Centro de Investigación Científica de Yucatán, A.C.

Posgrado en Ciencias Biológicas

Análisis de redes de coexpresión RNA-Seq para identificar patrones genéticos consenso en Arabidopsis en respuesta a la infección por hongos ascomycetes

Tesis que presenta

CYNTHIA GUADALUPE SOTO CARDINUALT

En opción al título de

DOCTOR EN CIENCIAS

(Ciencias Biológicas: Biotecnología)

Mérida, Yucatán, México

2023

CENTRO DE INVESTIGACIÓN CIENTÍFICA DE YUCATÁN, A. C.
POSGRADO EN CIENCIAS BIOLÓGICAS

RECONOCIMIENTO

Por medio de la presente, hago constar que el trabajo de tesis de **Cynthia Guadalupe Soto Cardinault** titulado **"Análisis de redes de coexpresión RNA-Seq para identificar patrones genéticos consenso en *Arabidopsis* en respuesta a la infección por hongos ascomycetes",** fue realizado en la Unidad de Biotecnología, en la línea de investigación de Agrobiotecnología, en el laboratorio Bioinformática del Centro de Investigación Científica de Yucatán, A.C. bajo la dirección de la **Dra. Elsa Beatriz Góngora Castillo**, dentro de la opción de Biotecnología, perteneciente al Programa de Posgrado en Ciencias Biológicas de este Centro.

Atentamente

Dra. Cecilia Hernández Zepeda
Directora de Docencia

Mérida, Yucatán, México, a 07 de junio de 2023

DECLARACIÓN DE PROPIEDAD

Declaro que la información contenida en la sección de Materiales y Métodos, los Resultados y Discusión de este documento proviene de las actividades de investigación realizadas durante el período que se me asignó para desarrollar mi trabajo de tesis, en las Unidades y Laboratorios del Centro de Investigación Científica de Yucatán, A.C., y que a razón de lo anterior y en contraprestación de los servicios educativos o de apoyo que me fueron brindados, dicha información, en términos de la Ley Federal del Derecho de Autor y la Ley de la Propiedad Industrial, le pertenece patrimonialmente a dicho Centro de Investigación. Por otra parte, en virtud de lo ya manifestado, reconozco que de igual manera los productos intelectuales o desarrollos tecnológicos que deriven o pudieran derivar de lo correspondiente a dicha información, le pertenecen patrimonialmente al Centro de Investigación Científica de Yucatán, A.C., y en el mismo tenor, reconozco que si derivaren de este trabajo productos intelectuales o desarrollos tecnológicos, en lo especial, estos se regirán en todo caso por lo dispuesto por la Ley Federal del Derecho de Autor y la Ley de la Propiedad Industrial, en el tenor de lo expuesto en la presente Declaración.

Firma: _____

Cynthia Guadalupe Soto Cardianault

Este trabajo se llevó a cabo en el laboratorio de Bioinformática de la Dra. Elsa Gongora del Centro de Investigación Científica de Yucatán, A.C., y forma parte del proyecto titulado “Análisis de redes de coexpresión RNA-Seq para identificar patrones genéticos consenso en *Arabidopsis* en respuesta a la infección por hongos *ascomycetes*” bajo la dirección de la Dra. Elsa Beatriz Góngora Castillo.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada No. 552999.

Al Centro de Investigación Científica de Yucatán AC y a la Unidad de Tecnologías de Información y Comunicación por las facilidades otorgadas para utilizar su equipo de cómputo de alto rendimiento.

A la Unidad de Biotecnología por las instalaciones prestadas para el desarrollo de la parte experimental de este trabajo.

A la Universidad Estatal de Michigan y al *Dr. Kevin Childs* por las facilidades otorgadas para utilizar su equipo de cómputo de alto rendimiento.

A mi directora de tesis *Dra. Elsa Beatriz Góngora Castillo* por el asesoramiento para el uso del equipo de alto rendimiento y por guiarme durante todo el posgrado, camino lleno de retos y satisfacciones, su mentoría fue imprescindible para lograr los objetivos alcanzados.

A los miembros de mi comité de avances *Dr. Jorge Ramirez, Dr. Jorge Santamaría, Dr. Santy Peraza, Dr. Mario Martínez, Dr. Ernesto Pérez y Dr. Victor Uc*, por ofrecerme orientación durante las diferentes etapas de mi tesis, sus consejos y recomendaciones fueron muy importantes para el desarrollo de mi proyecto.

A la *Mtra. Adriana Quiroz* técnica de nuestro laboratorio, por la asesoría otorgada.

A mis compañeros de laboratorio por sus recomendaciones durante los avances de tesis y los buenos momentos de convivencia.

A la *Dra. Elsa Góngora, Dr. Simon Prochnik, Dra. Sofia Robb, Dra. Clelia de la Peña Seaman, Dr. Sergio Peraza Sánchez* y *a Helmsley Fellowship* por el financiamiento otorgado para asistir al curso de *Programming for Biology* en CSHL, NY.

LISTA DE LOS PRODUCTOS GENERADOS (ORDEN CRONOLÓGICO DESC.)

Artículo. Unlocking the potential of public RNA-seq data in plants: A framework to integrating Next Generation Sequencing data for gene co-expression analysis. Cynthia Soto-Cardinault, Elsa-Gongora, Kevin L. Childs. Submission May 9th, 2023. Status: In review

Podcast. SPOTIFY CANAL OFICIAL DEL CICY. Más Conciencia | Podcast en Spotify. Oct 18, 2022. https://open.spotify.com/show/2QKUcSrkiHaMZlqieNLzTK

Taller. Red Mexicana de Bioinformática (RMB). Redes de Coexpresión Génica con WGCNA. http://redmexicanadebioinformatica.org/noticia-29-abril-22.html. Expositor principal. Duración: 2 hrs. Abril 29, 2022.

Taller. Cold Spring Harbor Labs (CSHL), NY. Programming for Biology Workshop. http://programmingforbiology.org/. Teacher assistant. Duration: 100 hrs. Octubre 9 to 25 th, 2022.

Conferencia. Tecnológico Nacional de México y el Instituto Superior de Valladolid. TECNOLOGÍAS DE LA INFORMACIÓN EN LA INVESTIGACIÓN. Nov 25, 2021. Mx.

ePoster. CICESE. RNA-Seq Co-Expression Network Analysis to Identify Consensus Patterns in Arabidopsis in response to ascomycetes fungal infection. Oct 13-15, 2021. Baja California, Mx.

ePoster. Sociedad Mexicana de Biotecnología y Bioingeniería (SMBB). RNA-Seq Co-Expression Network Analysis to Identify Consensus Patterns in Arabidopsis in response to ascomycetes fungal infection. XIX National Congress of Bioinformatics and Bioengineering. Sep 27 al 1 Oct, 2021. México, Mx.

Capítulo de libro. Springer Protocols ISBN 978-1-4939-9865-4 pp. 65-84. DNA Methylation and Transcriptomic Next-Generation Technologies in Cereal Genomics. Cynthia Soto-Cardinault, Fátima Duarte-Aké, Clelia De-la-Peña, Elsa Góngora-Castillo (2020)

Video. YOUTUBE OFICIAL CICY. Respondiendo dudas sobre el coronavirus. Jun 16, 2020. https://www.youtube.com/playlist?list=PLueJZUjjSML4EUojMlDkSPIax16fTn0FB

ePoster. 4th International Symposium on Functional Genomics and Systems Biology. The Cell Dynamics Research Center. Meta-transcriptomic analysis using data-mining strategies to identify similar responses in fungi infected plants. Yucatán, Mx. Nov. 20-22, 2019.

DEDICATORIAS

Agradezco a mi querida mamá *Alma Luisa* por su infinito apoyo durante estos años, su amor y apoyo incondicional me permitió alcanzar mis sueños.

Agradezco a mi esposo *Felipe Cornejo* por creer en mí, apoyar mis proyectos y ser incondicional en todo momento.

Agradezco a mis hijos *Milla* y *Damian* por su amor y comprensión, esto no hubiera sido posible sin ellos.

Agradezco a *Rosdhali* y *Mauricio* por todos los momentos de apoyo con mis hijos.

Esta tesis esta dedicada a todos ellos "Con amor".

Cynthia SC

ÍNDICE

LISTADO DE FIGURAS

LISTADO DE TABLAS

MATERIAL SUPLEMENTARIO

ABREVIATURAS

DNA: Ácido Desoxirribonucleico (Desoxirribonucleico Acid).
RNA: Ácido Ribonucleico (Ribonucleic Acid).
bRNA-Seq: Secuencias sin procesar RNA-Seq (Bulk RNA-Seq)
CDS: Genes codificantes de proteínas (Coding Sequence).
ETI: Inmunidad desencadenada por efector (Effector-Triggered Immunity).
FAO: Organización de las Naciones Unidas para la Alimentación y la Agricultura (Food and Agriculture Organisation).
FPKM: Fragmentos por kilobase asignada (Fragments Per Kilobase Mapped).
KDE: Estimación de la densidad del kernel (Kernel-Density-Estimation)
KR: Receptores kinases (Kinase Receptors).
LecRK: Quinasas Similares a Receptores de Lectina (Lectin Receptor-Like Kinases).
LPK: Proteínas Similares a quinasas (Like Protein Kinase).
LRR: Repetición Rica en Leucina (Leucine-Rich Repeat).
LysM: Motivo de Lisina (Lysine-Motif).
ML: Aprendizaje automático (Machine-Learning).
MOI: Integración multi-ómica (Multi Omic Integration).
NCBI: Centro Nacional de Información Biotecnológica (National Center for Biotechnology Information).
NCM: (Node-Connectivity-Mean)
NGS: Secuenciación de próxima generación (Next Generation Sequencing).
OMI: Integración múltiple de una sola capa (One-Omic Integration).
PAMP: Patrón molecular asociado a patógenos (Pathogen-Associated Molecular Pattern).
PCA: Análisis de componentes principales (Principal-Component-Analysis)
PRR: Receptores de reconocimiento (Pattern Recognition Receptors).
PTI: Inmunidad activada por patrón (Pattern-Triggered Immunity).
RLK: Receptor Tipo Quinasa (Receptor-Like Kinase)
RLP: Receptor Tipo Proteína (Receptor-Like Protein).
RPKM: Lecturas por kilobase asignada (Reads Per Kilobase Mapped).
SFT: modelo de topología libre de escala (Scale-Free-Topology)
SRA: Archivo de lectura de secuencia (Sequence Read Archive).
TOM: Matriz de superposición topológica (Topological Overlap Matrix).
TPM: Transcritos por millón asignados (Transcripts Per Million).
WAK: Quinasas Asociadas a la Pared (Wall associated Kinases).
WGCNA: (Weighted Gene Co-expression Network Analysis).

RESUMEN

La FAO estima pérdidas causadas por plagas y patógenos que rondan entre el 20% y 40% de la producción mundial de alimentos. En cinco de los principales cultivos de ingesta humana la pérdida global estimada es similar, alcanzando pérdidas de hasta el 95% en determinados cultivares. Dentro de las 10 primeras enfermedades reportadas, el 30% son atribuidas a patógenos fúngicos (Savary, S., 2019), alcanzando pérdidas de más del 50% en cultivos como soja y trigo. Los mecanismos de respuesta al estrés ensamblados por las plantas para hacer frente a las infecciones son muy complejos, y requieren de la percepción del estímulo para iniciar varias vías de señalización, y generar una respuesta. Los problemas actuales de globalización y el cambio climático extremo que enfrentamos, exacerban las probabilidades de enfrentar la aparición de enfermedades nuevas y emergentes, potenciando el ratio de diseminación que supone un riesgo para la seguridad alimentaria. Derivado de la situación emergente que enfrentamos, es necesario desplegar mecanismos alternos de análisis sobre la respuesta a la infección de las plantas desde una perspectiva sistémica. En este estudio utilizando un enfoque novedoso que incluye datos de transcriptomas RNA-Seq estudiados a partir de meta-análisis con redes de coexpresión, encontramos en la respuesta de *A thaliana* al estrés causado por *B cinerea* y *C higginsianum* a las 24 y 22 hpi, un clúster de respuesta consenso que involucra seis vías de respuesta en la etapa de infección asimilable a *in planta-appressorium*. Con soporte en la información disponible en 9 diferentes BDs ómicas, encontramos a las vías con repeticiones WD40/YVTN, metabolismo de esteroles, glicosiltransferasas, transporte de proteínas intracelulares, dedos de zinc y metiltransferasas, presentes en el momento de respuesta a la infección, con genes clave relacionados con la formación de calosa, la biogénesis de la pared celular, la señalización activada por Ca^{2+}, el transporte intracelular dependiente de clatrina (CME), la memoria activa al estrés por calor, y el control epigenético. Nuestros hallazgos también sugieren que el alto nivel de polimorfismos de secuencia encontrado en los genes anotados, podría ser clave para activar diversas vías de respuesta a la infección.

ABSTRACT

The FAO estimates losses caused by pests and pathogens of between 20% and 40% of world food production. In five of the main crops for human intake, the estimated global loss is similar, reaching losses of up to 95% in certain cultivars. Within the first 10 reported diseases, 30% are attributed to fungal pathogens (Savary, S., 2019), reaching losses of more than 50% in crops such as soybean and wheat. The stress response mechanisms assembled by plants to deal with infections are very complex, and require the perception of the stimulus to initiate various signaling pathways and generate a response. The current problems of globalization and the extreme climate change that we face exacerbate the chances of facing the appearance of new and emerging diseases, enhancing the rate of dissemination that poses a risk to food security. Derived from the emerging situation we are facing, it is necessary to deploy alternative analysis mechanisms on the response to infection of plants from a systemic perspective. In this study using a novel approach that includes RNA-Seq transcriptome data studied from meta-analyses with co-expression networks, we found in the response of *A thaliana* to stress caused by *B cinerea* and *C higginsianum* at 24 and 22 hpi, a consensus response cluster involving six response pathways at the stage of infection assimilable to in *plant-appressorium*. Supported by the information available in 9 different DBs, we find the *WD40/YVTN repeat-like-containing, Sterol metabolism, Glycosyltransferase, intracellular protein transport, Zinc finger,* and *Methyltransferases* pathways present at the time of response to infection, with key genes associate to callose formation, cell wall biogenesis, Ca^{2+} activated signaling, clathrin-dependent intracellular transport (CME), heat stress-activated memory, and epigenetic control. Our findings also suggest that the high level of sequence polymorphisms found in the annotated genes could be key to activating various response pathways to infection.

INTRODUCCIÓN

Uno de los mayores retos que enfrenta la agricultura es el combate a los hongos fitopatógenos, ya qué impactan los costos de producción para controlarlos, llegando incluso pueden comprometer la seguridad alimentaria. Aunque existen diversas medidas de control para combatir las enfermedades, como los fungicidas, los costos asociados merman la economía del agricultor, y eventualmente mal aplicados también pueden dañar la fertilidad del suelo, por ello han surgido alternativas de control como el mejoramiento genético. Sin embargo para aplicar estas técnicas desde las trincheras de la biología molecular, es necesario comprender mejor los mecanismo de defensa de las plantas desde una perspectiva sistémica, para lo cual los datos de secuenciación de próxima generación (NGS) (Kchouk et al., 2017) han resultado una gran promesa. Las tecnologías NGS producen datos a gran escala de genomas y transcriptomas para prácticamente cualquier organismo; hoy en día disponemos de enormes bases de datos (DB) de secuenciación parcialmente explorados que guardan un enorme potencial de investigación para ampliar las capacidades científicas, y alcanzar diagnósticos más precisos y robustos para fortalecer la agricultura. En la actualidad existen más de 17.9 *petabytes* (GenBank y WGS Statistics, 2022) de datos tan solo en las BDs del SRA-NCBI (Leinonen et al., 2011), qué junto con sus homólogos en Europa y Japón alojan la colección más grande de datos RNA-Seq del mundo, producto de miles de investigaciones conducidas por casi dos décadas; aquí, podemos encontrar datos NGS de diversas especies y grupos de trabajo, incluidos datos de secuenciación para el estudio de las interacciones planta-patógeno (Oh et al., 2022; Crandall et al., 2020).

Hoy en día muchas estrategias de investigación centradas en el análisis de la expresión génica con métodos de redes, basados en diversos modelos de aprendizaje automático (ML), han demostrado ser útiles para identificar patrones de expresión y genes diferenciados de forma consistente. Entre la gama de opciones están las redes de coexpresión génica como WGCNA (Weighted Gene Correlation Network Analysis) (Langfelder y Horvath, 2008) que se ha aplicado en múltiples estudios de transcriptómica que incluyen organismos de todo tipo (humanos, animales, microbios y plantas) (Iancu et al., 2012; Ghazalpour et al., 2008; Carlson et al., 2006; Horvath et al., 2006). En investigaciones con datos de plantas la encontramos ampliamente utilizada para anotación funcional, evolución molecular y rutas de regulación e inferencias de red (Lin et al., 2019; Liang et al., 2018; Childs et al., 2011). A través de casi dos décadas de su uso, las estrategia de análisis de redes génicas esta acreditada.

Hoy en día con el reconocimiento de la presión del cambio climático extremo sobre la agricultura, altamente dependiente del clima, esencialmente por un jugar un rol condicionante para el desarrollo exitoso de los patógenos de plantas, se teme la pérdida de continuidad de las estrategias de control previamente desarrolladas, debido a la alta dinámica de la presión del cambio extremo sobre las plantas, con lo que se espera que eventualmente aparezcan nichos de oportunidad para el desarrollo de patógenos oportunistas. Las enfermedades causadas por patógenos de plantas causan daños importantes en algunos cultivos que rondan entre el 20% y 40% (Savary et al., 2019) de las pérdidas agrícolas. De acuerdo con datos extraídos de la encuesta de Savary, los hongos patógenos de plantas se encuentran dentro de las primeras causas reportadas, con más del 50% de enfermedades atribuidas a estos patógenos. Se teme qué, el cambio climático extremo y la globalización propicien la aparición cada vez más frecuente de enfermedades de plantas tanto nuevas como emergentes. Ante esta disyuntiva, es imperante ampliar el estudio de los mecanismos de respuesta de las plantas ante la infección fúngica desde una perspectiva global (Organización de las Naciones Unidas para la Alimentación y la Agricultura, 2018), es decir, a nivel multisistemas, de forma que se pueda evaluar la respuesta de la planta a la infección ante múltiples estresores.

Debido a ello, en esta tesis investigamos la respuesta a la infección en arabidopsis causada por diversos hongos ascomycetes *d*entro de las primeras horas de infección. Se utiliza un enfoque novedoso para la integración de conteos de expresión génica RNA-Seq, y métodos de redes de coexpresión para su análisis, que en conjunto brindan la posibilidad de descubrir patrones de defensa activos en la planta que son comunes ante el ataque de diversos patógenos. Se utiliza una estrategia de análisis de multisistemas para identificar la respuesta de defensa consensuada de la planta en diversos pato-sistemas, se realizan pruebas de enriquecimiento

CAPÍTULO I

ANTECEDENTES

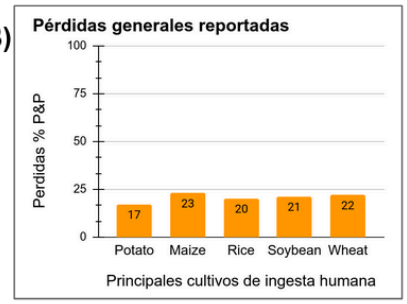
1.1. PÉRDIDAS AGRÍCOLAS CAUSADAS POR PATÓGENOS

La Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), estima pérdidas causadas por plagas y patógenos que rondan entre el 20% y 40% de la producción mundial de alimentos, constituyendo una amenaza para la seguridad alimentaria en algunas regiones dependiendo del tipo de cultivar. Se reporta a partir de una encuesta realizada a 219 expertos de 67 países que engloban las principales regiones productoras del mundo para los cinco principales cultivos de ingesta humana, pérdidas globales similares, 21.5% en trigo, 30% en arroz, 22.5% en maíz, 17.2% en papa y 21.4% en soya, con extremos de pérdidas entre el 69% y 95% dependiendo del cultivar (Savary et al., 2019) (**Figura 1.1A**). Datos extraídos de esta encuesta, también revela que más del 30% de las enfermedades reportadas corresponden a patógenos fúngicos (**Figura 1.1B**) alcanzado pérdidas de más del 50% en cultivos como la soja y el trigo (**Figura 1.1C**).

(A)

CULTIVO	ENFERMEDAD / PATÓGENO FÚNGICO
Papa *(Potato)*	Late blight (P infestans) Early blight (A solani) Early dying/Verticillium wilt (V albo-atrum, V. dahlia, Pratylenchus penetrans)
Maíz *(Maize)*	Northern corn leaf blight (Exserohilum turcicum; Setosphaeria turcica) Southern rust (Puccinia polysora) Diplodia ear and stem rot (Diplodia frumenti) Fusarium and Gibberella stalk/earl rots (F moniliforme, F graminearum)
Soja *(Soybean)*	White mold (S sclerotiorum) Soybean rust (Phakopsora pachyrhizi) Sudden death (F virguliforme, F tucumaniae, Heterodera) Phytophthora root and stem rot (P sojae) Rhizoctonia root rot, web blight (R solani) Pythium damping-off (Pythium spp)
Trigo *(Wheat)*	Septoria (Zymoseptoria) tritici blotch Stagonospora nodorum blotch Leaf (brown) rust (Puccinia triticina) Stem (black) rust (Puccinia graminis f. sp. tritici) Stripe (yellow) rust (Puccinia striiformis f. sp. tritici) Fusarium head blight – Scab (Fusarium spp., Microdochium spp) Tan spot (Pyrenophora tritici-repentis) Spot blotch (Cochliobolus sativus)
Arroz (Rice)	Leaf, neck, or panicle blast (Rhizoctonia solani) Sheath blight (Rhizoctonia solani) Brown spot (Cochliobolus miyabeanus) Sheath rot (Sarocladium oryzae)

(B)



(C)

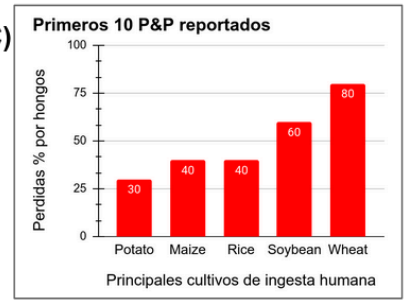


Figura 1.1. Carga de patógenos fúngicos en los principales cultivos alimentarios.

Aunque se conocen las cifras, los causales específicos de estas pérdidas no están detallados, debido a la amplia diversidad de la compleja red de interacciones entre plantas y patógenos que dificultan el control individual para cada uno de los cultivos existentes. Las

limitaciones en países con bajos ingresos *per cápita* también son una limitante. La comunidad internacional del cambio global para evaluar la vulnerabilidad de las plantas bajo el cambio climático (Sutherst et al., 2007) también señaló, que la base de las evaluaciones de riesgo de bioseguridad no es rentable para realizar análisis en profundidad para las importaciones en auge, del mismo modo que no es factible realizar análisis detallados de la mayoría de las plagas para determinar sus posibles impactos con el cambio climático. Se indica que las importaciones y el cambio climático son algunos de los causales de las enfermedades, por lo que para contrarrestar estos efectos se recurre a pesticidas como la forma más común de control para preservar los rendimientos. Sin embargo, su uso trás el aumento de la producción de alimentos desde la década de 1950 --que ha permitido satisfacer las necesidades de una población en crecimiento ha tenido un coste para el medio ambiente (Cooper y Dobson, 2007), ha impactado negativamente la biodiversidad, incluidas poblaciones de insectos, aves y peces, la calidad del suelo, el aire y el agua (Gill y Garg, 2014; Goulson, 2014; Sanchez-Bayo y Goka, 2014), así como, el riesgo que implica para la salud humana, con efectos tanto agudos como crónicos (K.-H. Kim et al., 2017; Bassil et al., 2007). Aunque conocidos los efectos negativos de estos mecanismos de control, la cantidad de plaguicidas sigue aumentando para cubrir la demanda de alimentos en todo el mundo, reportando incrementos de más del 78% de toneladas de ingredientes activos utilizados tan solo entre 1990 y 2016 (Organización de las Naciones Unidas para la Alimentación y la Agricultura, 2018).

Las límites en los recursos de las instituciones públicas y privadas para censar esta información, el amplio ámbito de los nichos de brote, y el estrecho margen de tiempo en la propagación de las enfermedades, sugieren una urgente necesidad de explorar nuevos enfoques de análisis y control de riesgos para contrarrestar los efectos de las pérdidas en los cultivos. Evaluar la salubridad de los cultivos requiere mucho tiempo, verificar la condición de cada planta varias veces en una temporada no es práctico, sencillo, ni económico, mucho menos en los grandes cultivares. La dificultad de acceder a algunos también puede complicar la prospección. FAO señala, *"las condiciones actuales exigen enfoques parsimoniosos que exploten conjuntos de datos mínimos y herramientas de modelado genéricas para responder preguntas relacionadas con numerosas especies de plagas a escala global"* (FAO, 2008), se alienta a explorar estrategias para la identificación automática de enfermedades (Boulent et al., 2019), el desarrollo de herramientas semi-automatizadas de diagnóstico, y el desarrollo de estrategias computacionales para identificar nuevos objetivos genéticos para producir plantas resistentes de más amplio espectro a múltiples enfermedades (Boutrot y Zipfel, 2017).

1.2. ENFERMEDADES EN LAS PLANTAS CAUSADAS POR HONGOS

Las enfermedades fúngicas son uno de los mayores retos que enfrenta la agricultura, ya que los hongos pueden atacar a las plantas antes, durante y después de la cosecha. Los hongos patógenos impactan los costos de producción generando pérdidas agrícolas y económicas, que incluso comprometen la seguridad alimentaria. Además un mismo hongo puede infectar a varios tipos de planta, aunque sean de diferentes especies (spc), y aunque por lo general el hongo pasa la mayor parte de su ciclo de vida como parásito en la planta (huésped) y el resto como saprófito en los residuos vegetales que quedan en el suelo, es común que el patógeno se reproduzca en la superficie de la planta o cerca de ella, y finalmente se dispersen fácilmente como espora, atacandola en forma local o general. La relación e intensidad entre un hongo y su huésped depende de muchos factores, entre los que están su estilo de vida, los factores genéticos y las condiciones ambientales.

1.2.1. Estilos de vida de los hongos

En 1991 se estimó que había 1.5 millones de especies de hongos (Hawksworth, 1991), sin embargo la aparición de NGS cambió estas estimaciones de 2.2 a 3.8 millones de spc(s) (Cheek et al., 2020), por lo que bajo esta óptica es muy complejo clasificarlos. Una forma práctica de hacerlo en 6 categorías es a partir de su estilo de vida, en saprófito, patógenos, endófitos, epífitos, simbióticos, comensales o liquenizados en una amplia gama de huéspedes que incluyen animales y plantas. En cada uno de estos estilos de vida sus interacciones con el huésped varían, por lo que las interacciones entre los hongos y sus huéspedes son dinámicas y complejas. Si bien la mayoría de los hongos viven en materia orgánica muerta, existe evidencia de que pueden ser endófitos que cambian a otros estilos de vida en diversas condiciones. Algunas especies viven dentro del huésped como endófitos y cuando las condiciones cambian, se vuelven patógenos. Los factores que desencadenan estas transformaciones no se conocen por completo, quedando además por confirmar si estas variaciones en los estilos de vida se deben a diferencias en los niveles de expresión de genes fúngicos o los genotipos del huésped, o descifrar si algunas especies de hongos cambiar su estilos de vida de un huésped a otro (Cheek et al., 2020).

Siendo más específicos podemos decir que los hongos patógenos pueden ser de tres tipos, biotrofos, necrótrofos y hemibiotrofos. Desde esa perspectiva, menos del 2% de ~100 mil son conocidos como especies patógenas de plantas. Los hongos biotrofos suelen mantener al

huésped vivo y usualmente exhiben especialización por spc. de planta, como *powdery mildew* que infecta a diferentes spc. Usualmente si un biotrofo patógeno coloniza a una planta, el metabolismo y desarrollo de la misma se verá severamente alterado, exhibiéndose problemas de crecimiento y desarrollo. En el caso de los hongos necrotrofos, estos producen enzimas que degradan la pared celular de la planta y tienden a atacar a un amplio número de especies, como *Botrytis cinerea* que ataca a más de 1000 spc. de plantas. Otras como *Sclerotinia sclerosis* que puede infectar a más de 400 especies, tiene un modo de acción específico que inhibe las señales de las plantas, segregando toxinas (*oxalyc acid)* que tiene la función de bloquear el sistema de defensa temprano. Los hongos hemibiotrofos por su parte, primero exhiben un estilo de vida biotrofo y después cambian a un estilo de nutrición necrotrofo; el cambio podría estar activado por la alta demanda nutricional derivado del crecimiento de la biomasa del patógeno y el desarrollo de su ciclo sexual; muchos fitopatógenos devastadores para la agricultura se encuentran en esta categoría.

1.2.2. Los hongos ascomycetes

El *phylum* de los *Ascomycota* contiene más del 60% de los hongos descritos en la actualidad, aquí se encuentra un gran número de los hongos patógenos de plantas. Según Cavalier Smith (1998) en este phylum hay 15 clases, 68 órdenes, 327 familias y 6355 géneros, con más de 64163 especies. Otras taxonomías basadas en datos NGS los han reclasificado, por ejemplo, *JGI Fungal Program* (Grigoriev et al., 2014) los agrupa en 10 clases (**Figura 1.2**).

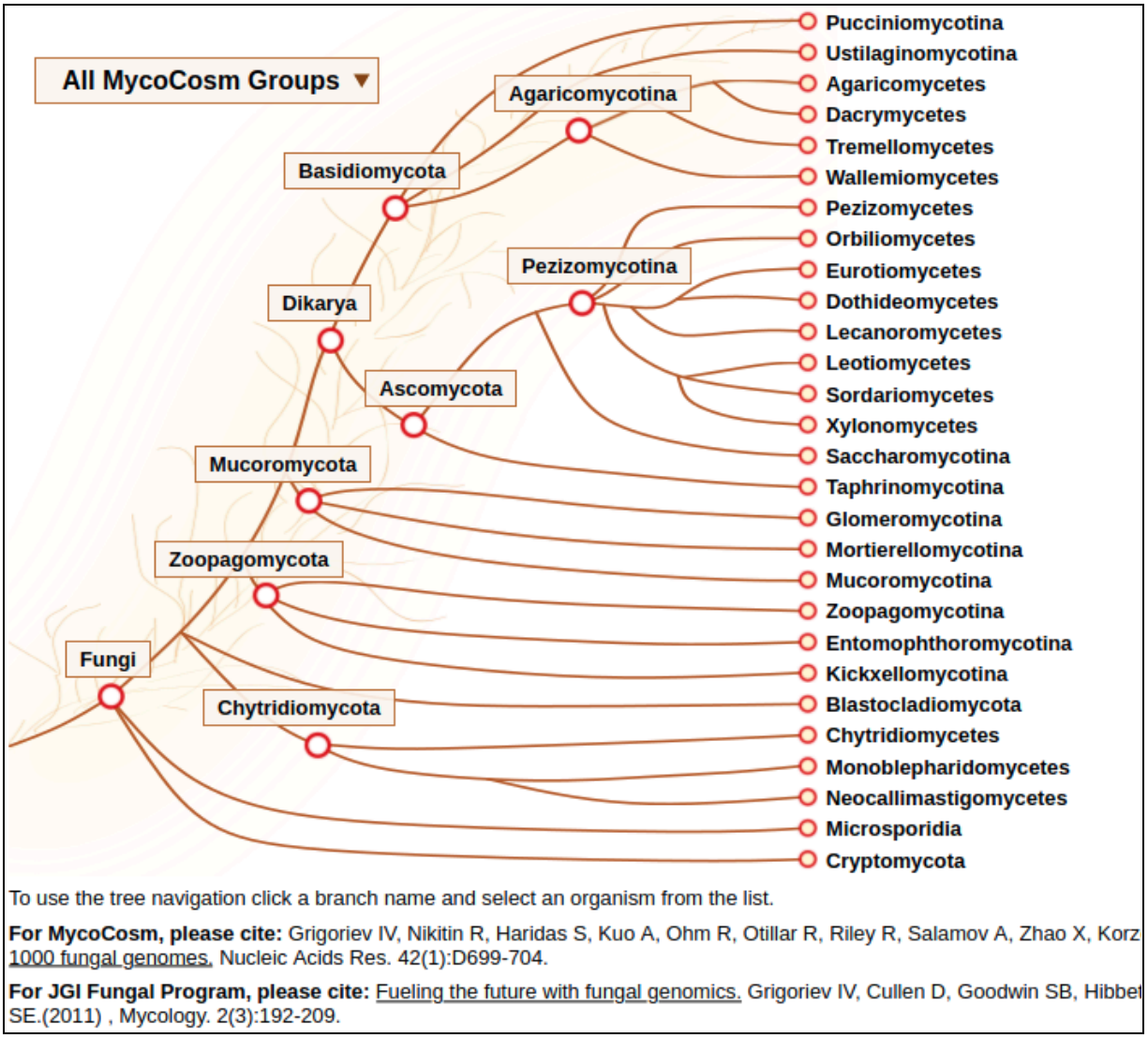


Figura 1.2. Taxonomía del reino de los hongos. Extraída de MycoCosm website (JGI Fungal Program).

Los *Ascomycetes* son organismos unicelulares (levadura) y multicelulares, marinos, de agua dulce y terrestres, con un amplio rango de patógenos no solo de plantas, sino también de animales y humanos. Su característica morfológica distintiva es el *ascus*, una microscópica estructura sexual en la cual esporas inmóviles llamadas *ascosporas* son formadas. Muchos patógenos *Ascomycetes* de plantas de las clases *Dothideomycetes, Sordariomycetes* y *Leotiomycetes,* de la subdivisión *Pezizomycotina* son muy dañinos para el sector agronómico (Doehlemann et al., 2017). Algunas enfermedades fúngicas dentro de estas clases bajo vigilancia epidemiológica son por ejemplo, el moho gris en la fresa causado por *Botrytis cinerea*, el rizo de hoja de melocotón causado por *Taphrina spp*., la marchitez vascular causado por *Verticillium, Fusarium*, la pudrición marrón cereza causado por *Monilinia fructicola* y *M. laxa*, el *tizón temprano* del tomate causado por *Alternaria solani*, el tizón tardío del tomate y la papa

causado por *Phytophthora infestans*, la antracnosis en el aguacate causado por *Colletotrichum spp*. o en papaya causado por *C. gloeosporioides*, y el cornezuelo de centeno causado por *Sclerotinia sclerotium*. Los signos y síntomas son variables, pero en general se presentan diversos trastornos foliares, podredumbres de frutas, verduras, raíces y tallos, cancros y antracnosis (Doehlemann et al., 2017) (**Figura 1.3**).

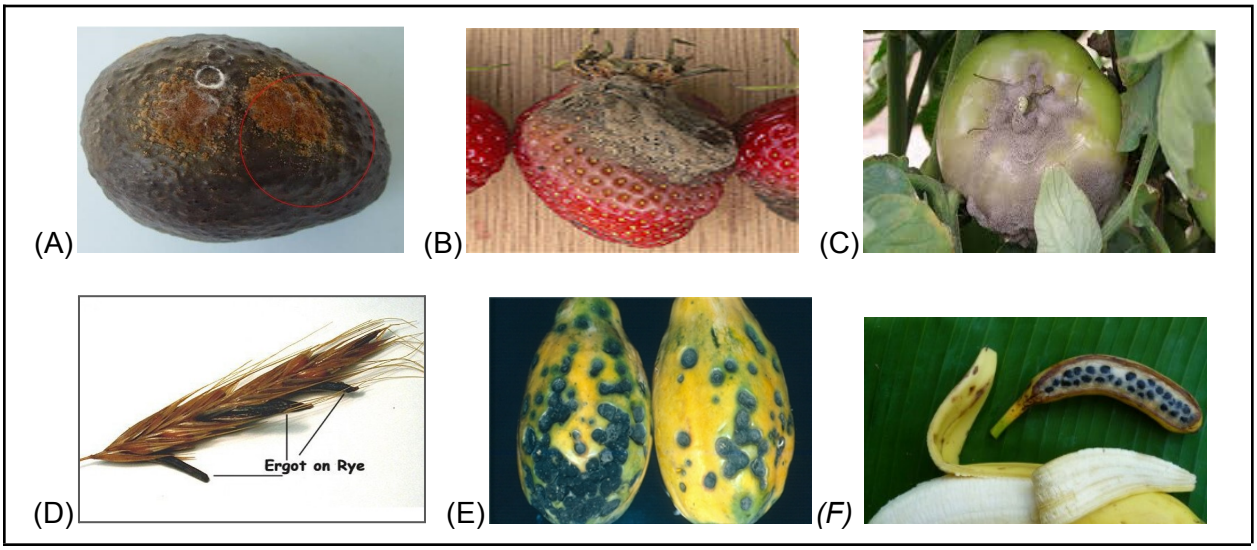


Figura 1.3. Enfermedades fúngicas causadas por *Ascomycetes*. Ejemplos de enfermedades: (A) Antracnosis en aguacate (*Colletotrichum* spp), (B) Moho gris en fresa (*B cinerea*), (C) Moho gris de tomate (*B cinerea*), (D) Cornezuelo de centeno (*S sclerotium*), (E) Antracnosis en papaya (*C gloeosporioides*) y (F) Marchitez del banano (*F oxysporum*).

1.3. ORGANISMO MODELO VEGETAL *ARABIDOPSIS THALIANA*

A thaliana, nombre científico dado por Heynh, fue introducida en el laboratorio hace más de medio siglo, siendo utilizada como organismo modelo vegetal. Sus características particulares de desarrollo y su genoma pequeño la han hecho idónea para investigaciones sobre el desarrollo, crecimiento y floración de la planta, sirviendo a múltiples investigaciones para estudiar la función génica en especies de cultivos y otras plantas de importancia para los humanos (Provart et al., 2016). Arabidopsis es una pequeña planta de la familia *Brassicaceae* (mostaza) que comparte familia con un amplio abanico de especies de interés agrícola, como la *Brassica oleracea* (col), la *Brassica napus* (nabo), el *Raphanus sativus* (rábano), las *Brassica campestris* (diversas mostazas), la *Sinapis alba* (mostaza blanca) y *B nigra* (mostaza negra). En el año 2000 el proyecto “Iniciativa para el Genoma de Arabidopsis” (AGI) obtuvo el primer mapa genético de la planta, reportando *25498* genes codificantes (CDS) de proteínas; en 2010 se liberó la más reciente actualización del genoma TAIR10 (Lamesch et al., 2012), y en 2017 la más reciente anotación Araport11 (Cheng et al., 2017) que reportó *27655* CDS, además de

nuevas características génicas relevantes. Al día de hoy, 50% de sus genes se encuentran en estado desconocido para la función molecular y proceso biológico, y 30% para los componentes celulares (GO en TAIR10, Diciembre 2020). Debido a la refinada y consistente descripción que se tiene ahora de su genoma, así como a la amplia descripción molecular de sus funciones y procesos moleculares, arabidopsis sigue vigente como modelo molecular vegetal para estudios realizados con datos NGS.

1.4. MECÁNISMOS DE RESPUESTA AL ESTRÉS

La respuesta inmune de la planta es tradicionalmente esquematizada en dos umbrales de respuesta, PTI (Pattern Triggered Immunity) y ETI (Effector Triggered Immunity). PTI se relaciona con el reconocimiento de patrones moleculares asociados a patógenos (Pathogen-Associated Molecular Pattern / PAMPs), el cual actúa sobre la superficie de la célula huésped a partir de receptores PAMPs, y ETI implica receptores intracelulares de virulencia (efectores) de respuesta amplia y variable, con una extremada diversificación tanto dentro como entre especies (Jones y Dangl, 2006) (**Figura 1.4**). La PTI está asociada con la resistencia sistémica adquirida (SAR) la cual se establece contra patógenos biotróficos, y está regularmente controlada por ácido salicílico (SA), mientras a la ETI se asocia con la resistencia sistémica inducida (ISR) que requiere de la vía de señalización del ácido jasmónico (JA) y el etileno (ET), se le encuentra como una respuesta a patógenos necrotróficos (Bürger y Chory, 2019).

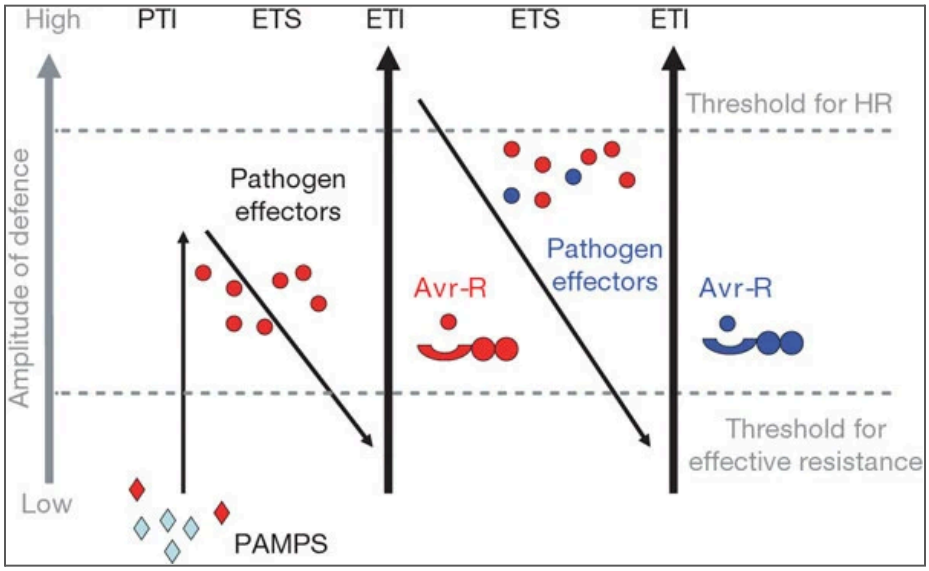


Figura 1.4. Modelo Zig-Zag de inmunidad cuantitativa. Esquema que ilustra la lucha molecular ante ataque de patógenos propuesta por Dangl y Jones en 2006.

En el PTI, la respuesta inmune de la planta es censada por receptores de reconocimiento (PRR) que constituyen la primera capa de inmunidad que restringe la proliferación de patógenos (Ranf, 2018; J. Zhang et al., 2010), considerada ampliamente conservada entre las familias, no existiendo una diversificación extrema de los receptores y efectores tanto dentro como entre especies (Macho y Zipfel, 2014; Dodds y Rathjen, 2010; Jones y Dangl, 2006). Los PRR de plantas son en su mayoría proteínas quinasas/cinasas (PK) localizadas en la superficie de la planta o proteínas similares a quinasas (RLP) que contienen varios ectodominios de unión a ligandos que perciben los PAMPs. Los pares conocidos de PAMP-PRR de plantas ilustran distintos mecanismos de reconocimiento molecular (Zipfel C, 2014). Los PRR son parte de complejos de proteínas multiméricas en la membrana plasmática, que reclutan quinasas citoplasmáticas de manera diferencial que conectan los complejos PRR con componentes de señalización descendentes. La unión del ligando inicia una serie de eventos de fosforilación dentro de los complejos PRR que activa la señalización inmune celular, que incluye el estallido intracelular de especies reactivas al oxígeno (ROS), los canales de calcio, la activación de cascadas de quinasas citoplasmáticas, y la reprogramación transcripcional para activar los genes de defensa. Al igual que en los mamíferos, la activación excesiva de las respuestas inmunitarias de las plantas puede tener consecuencias perjudiciales. Por tanto, un sistema regulador negativo complejo controla diferentes componentes inmunitarios para mantener la homeostasis celular. Aún con este sistema tan sofisticado de respuesta, muchos patógenos son capaces de subvertir el sistema inmunológico de la planta al secretar moléculas especializadas, llamadas efectores, que a menudo imitan el modo de acción de los reguladores negativos de la señalización inmunitaria del huésped (Macho y Zipfel, 2014). Un ejemplo de esta activación de señalización en arabidopsis la encontramos en los receptores LysM-RLK CERK1 / RLK1, los cuales son necesarios para la percepción de quitina en la pared celular. Muchas investigaciones se enfocan en el estudio de las relaciones PAMP-PRR debido a su potencial para el diseño de organismos vegetales resistentes a enfermedades de amplio espectro.

Existe una variada clasificación de efectores de diverso contenido y estructura química, muchos caracterizados en *A thaliana, C annuum, S lycopersicum, N tabacum, B napus, G max, N benthamiana, O sativa, T aestivum y Z mays*. Entre los efectores fúngicos de proteína tenemos por ejemplo *Cellulase, Avr9, Ethylene-inducing xylanase (EIX), Avr4, Necrosis-inducing protein1 (NIP1), Ecp2, Cerato-platanin, Ecp1, Ecp4, Ecp5, Avr2, Endopolygalacturonase, Avr3/Six1, Avr4E, PemG1, Nascent polypeptide-associated, complex*

(NAC) α-polypeptide, Ave1, PevD1, Hypersensitive response-inducing protein (HRIP), Serine protease (AsES), Avr5, Cutinase, Hydrophobin Cyclodipeptides, CS20EP, Rapid alkalinization factor (RALF), Avr1/Six4, SnTox1 y AvrStb6. A base de carbohidratos están la *Chitin, Oligochitosan y β-1,3-glucan.* De glicopéptido esta *gp8.* De lípidos el *Ergosterol.* De metabolitos secundarios el *Chrisophanol y la Cerebroside.* También se han identificado variedades de PRRs con ligandos / agonistas, entre los que se encuentran los KR de las familias LRR XI (PEPR1, PEPR2, RLK7), XII (EFR, XPS1, CORE), WAK (WAK1, Snn1), LysM (AtCERK1, AtLYK5, APR3, AtLYM1, AtLYM2 y AtLYM3), L-Lec (LecRK-1.9), G-Lec (SD1-29, I-3), algunos receptores tipo quinasas LRR (LePR3, RLP1, RLP23, RLP30, RLP42) (**Figura 1.4**) (Saijo et al., 2018; Boutrot y Zipfel, 2017).

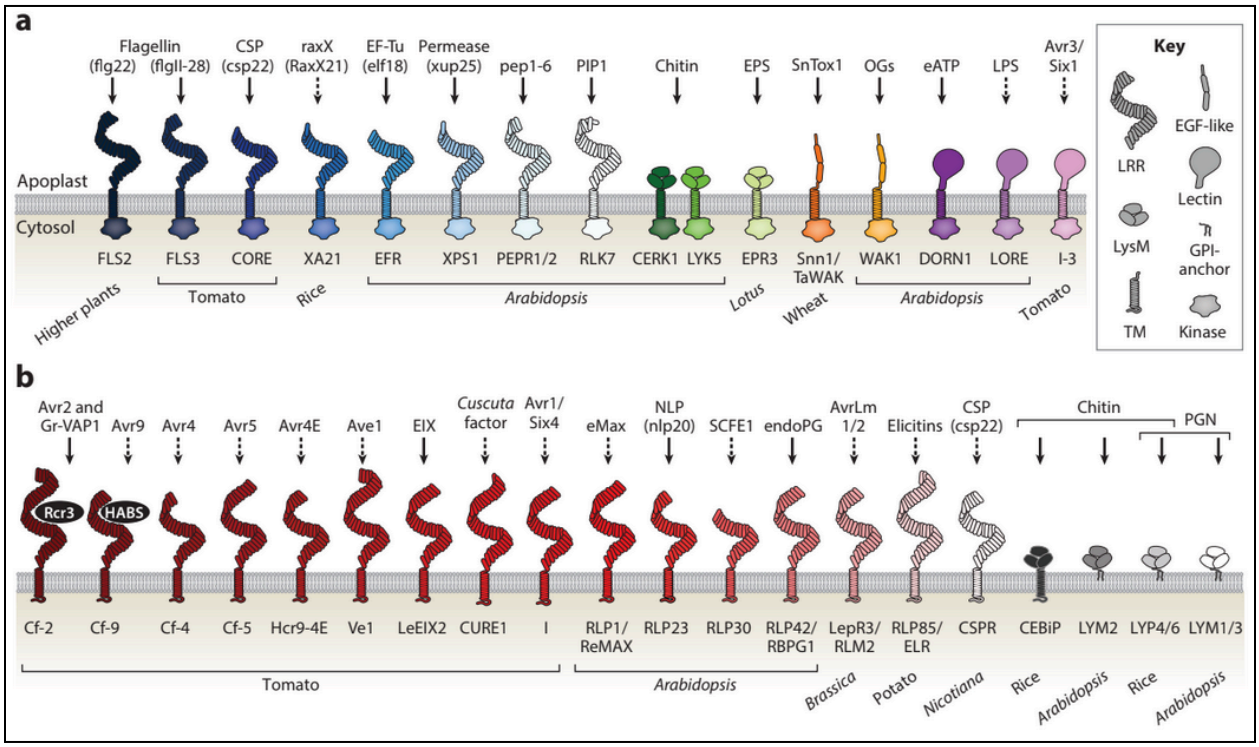


Figura 1.5. Estructura de los PRRs de plantas. (a) Quinasas receptoras (RK). (b) Proteínas de tipo receptor (RLP). Las flechas continuas indican unión directa demostrada; Las flechas discontinuas indican una falta actual de evidencia de unión directa. Abreviaturas: EGF, factor de crecimiento epidérmico; EIX, xilanasa inductora de etileno; EPS, polisacáridos extracelulares; GPI, glicofosfatidilinositol; LPS, lipopolisacárido; LRR, repetición rica en leucina; OG, oligogalacturónidos; PGN, peptidoglicano; TM, transmembrana.

1.4.1 Crosstalk de la respuesta de las plantas al estrés

Investigaciones que integran múltiples sistemas planta-patógeno con base en datos NGS, con el objetivo de identificar patrones de respuesta consenso diferenciados, son escasos

en la literatura, tanto para microarreglos como para RNA-Seq, probablemente esta escasez se debe a las dificultades que se tiene para integrar, normalizar y analizar datos que proceden de diversos orígenes, ya que invariablemente se conlleva problemas de alta variabilidad técnica (efecto batch) durante el análisis, que puede confundir la variabilidad biológica, perdiendo la huella biológica de la pregunta de interés.

Algunos estudios que han abordado este enfoque de análisis a nivel multi-sistemas con datos de plantas, aunque no precisamente en relación con sus mecanismos de defensa, son el de (Shaik y Ramakrishna, 2013) donde integra ~300 muestras de microarreglos de 20 estudios en arroz y arabidopsis para identificar qué ~38.5% y ~28.7% de genes diferencialmente expresados están presentes en respuesta a la sequía, y también responden a estrés bacteriano; los estudios de (Das et al., 2017) y (Childs et al., 2011), donde identifican la respuesta común al estrés biótico causado por aluminio en la soja, y para realizar anotación funcional para genes de arroz, respectivamente; y el de (Shahan et al., 2018) donde utiliza 46 bibliotecas RNA-Seq de fresa en diferentes condiciones para evidenciar el fenómeno del aumento del transporte de hierro para el desarrollo de frutos.

El único en que encontramos específicamente realizado para analizar la respuesta de la planta ante diferentes tipos de infección con patógenos es el de (Rodrigo et al., 2012), donde específicamente se integran datos de microarreglos para analizar la interacción de arabidopsis tras la infección con 7 diferentes virus, encontrando 7 VRG regulados al alza en común por seis virus, de los cuales, sorprendentemente, seis juegan un papel en la migración celular (At3g57260, At5g10380, At3g14990, At3g28510, At5g52640 y At4g24690) y uno (At1g75040) codifica una taumatina PR-5 LK, factores conocidos por su participación en las respuestas de los patógenos. No se identificó ningún VRG en común entre las ocho infecciones. 1 VRG fue sistemáticamente regulado por siete virus (es decir, todos excepto PPV) y se descubrió que codifica una aspartil proteasa involucrada, nuevamente, en la migración celular. De acuerdo a sus hallazgos, la migración celular juega un rol importante en esta respuesta a la infección. La escasa cantidad de estudios disponibles realizados con datos NGS para el análisis de la respuesta consenso de la planta ante múltiples escenarios de infección, dificulta obtener referencias consistentes sobre el crosstalk de la respuesta, por lo que es necesario realizar más estudios bajo esta perspectiva.

1.5. LA EXPRESIÓN GÉNICA CON TECNOLOGÍAS NGS (RNA-Seq)

La expresión genética es un proceso bioquímico estrictamente regulado que permite a una célula responder dinámicamente a los estímulos ambientales y a sus propias necesidades cambiantes, exceptuando a los genes constitutivos, todos los demás genes se expresan o no dependiendo de la función de la célula en un tejido particular; por ejemplo, genes codificantes de proteínas responsables del transporte axonal se expresan en las neuronas y no en linfocitos donde se expresan los genes de la respuesta inmune. También existe especificidad temporal, donde diferentes genes se expresan en diferentes momentos de la vida de un organismo; además la regulación también varía según las propias funciones del gen.

A nivel molecular, la expresión genética inicia con la transcripción del DNA, cuyo resultado es la síntesis de tres tipos de moléculas RNA, RNA mensajero (mRNA), RNA transferente (tRNA) y RNA ribosómico (rRNA). En eucariotas, la información codificada de un gen es copiada del DNA al mRNA en el núcleo celular, transportando información que será utilizada durante la traducción en el citoplasma para ensamblar una proteína. El mRNA actúa como molde llevando información a los ribosomas que contienen rRNA, y en los que participa el tRNA como adaptador para convertir la información en aminoácidos que formarán proteínas, las cuales tendrán un rol específico, ya sea para formar estructuras o realizar una función biológica específica, proceso conocido como “el dogma central de la genética molecular”. El porcentaje de mRNA en la célula varía, incluso entre momentos diferentes de la vida de una misma célula, pero en general se encuentra en bajo contenido, por ejemplo, en *E.coli* el ~5% del RNA es mRNA, ~15% es tRNA y ~80% es rRNA (Klug, W. S. et.al., 2006).

A nivel estructural el RNA de estructura similar a la del DNA, esta formado por nucleótidos unidos en cadenas de polinucleótidos, el azúcar ribosa reemplaza a la desoxirribosa y el uracilo reemplaza a la timina, por lo que el uracilo es el complementario de la adenina durante la transcripción y durante el emparejamiento de las bases. Los genes de organismos eucariontes cuyas secuencias están interrumpidas por segmentos de nucleótidos que no se expresan en la secuencia de la proteína que codifican, son secuencias de DNA que se encuentran en el transcrito inicial de RNA, pero se eliminan antes de que el RNA maduro (mRNA) se traduzca. Las secuencias de DNA que sí están en el mRNA se llaman *exones* y las que no están *intrones*. En general, el corte y empalme dirige la eliminación de los *intrones* como resultado de un proceso de escisión, seguido de la unión de los exones (**Figura 1.6**) (Klug, W. S. et.al., 2006). La comprensión de estos conceptos moleculares y estructurales del

DNA/RNA es fundamental, ya que en función de estos, se define la característica molecular a medir, así como opciones de mapeo, alineación y cuantificación de las lecturas de datos NGS. De tal forma que se puede medir y contar el total de moléculas RNA, es decir el transcriptoma, o solo contar moléculas específicas, por ejemplo, moléculas de RNA no codificante (ncRNA).

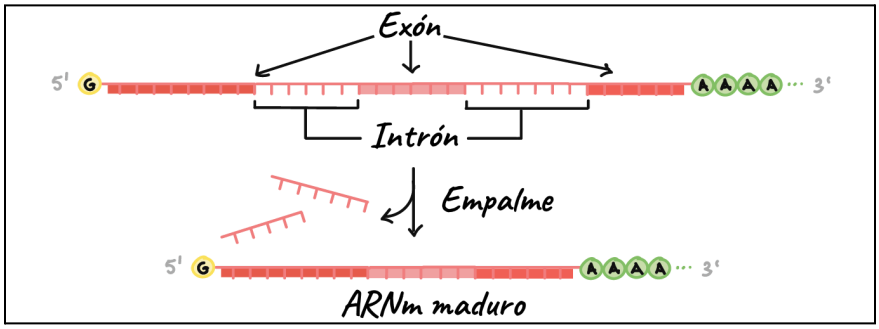


Figura 1.6. Secuencias intercaladas en el RNA inicial en un gen fragmentado. Los exones están intercalados por fragmentos de Intrones que son eliminados en el RNA maduro.

Las tecnologías de secuenciación de próxima generación (NGS), denominadas RNA-Seq, son cualquier tecnología que utilice procesamiento masivo paralelo y cuyas secuencias producidas (reads/spots) son cortas, generalmente de entre 100 a 300 pares de bases (pb). NGS tiene un amplio rango de aplicaciones debido a la capacidad que tienen de captar no sólo mRNA, sino otras moléculas de RNA de gran relevancia, como las ncRNA, los RNA largos no codificantes (lncRNA), el RNA nucleolar pequeño (snoRNA) y los microRNAs (Babarinde et al., 2019). La disminución en los costos de secuenciación aumentó la accesibilidad a la tecnología, ofreciendo muchas ventajas sobre sus predecesores, los microarreglos, sujetos a hibridación cruzada y con un mayor margen de error; RNA-Seq proporciona precisión y un rango dinámico más amplio (Z. Wang et al., 2009), es rápido y flexible; además de que permite sensar y medir el estado transcripcional de prácticamente cualquier especie, incluidos organismos que carecen de una secuencia genómica (Góngora-Castillo y Buell, 2013). A partir de su surgimiento paralelamente se ha impulsado el desarrollo de muchas aplicaciones que han permitido identificar y caracterizar varias clases de RNA, interacción RNA-proteína y su localización genómica (Reuter et al., 2015).

La estrategia RNA-Seq adoptada por un largo tiempo por muchos grupos de investigación a nivel global (Han et al., 2015), sigue en auge, y sus productos de datos han impactado el crecimiento exponencial de las BDs (Sayers et al., 2020). En tan solo una década el *archivo de lecturas de secuencias sin procesar* (SRA) del Centro Nacional de Información Biotecnológica (NCBI) de la Biblioteca Nacional de Medicina de Estados (Leinonen et al., 2011),

alcanzó 8.8 *petabytes* de datos de libre acceso, hasta agosto del 2020. En el último reporté de agosto del 2022, su contenido ya se había duplicado, alcanzando 17.9 *petabytes* (GenBank and WGS Statistics, 2022) en tan sólo dos años, sin duda, una mina de oro en datos para la investigación bioalimentaria y biomédica del presente. Aunque a NGS le preceden otras tecnologías NGS, como Oxford-Nanopore (Deamer et al., 2016) y PacBio (Eid et al., 2009), con varias ventajas y desventajas, en lo que se refiere al estudio de la expresión genética, la forma más común en la actualidad sigue siendo NGS RNA-Seq.

1.6 ANÁLISIS MULTISISTEMA CON TRANSCRIPTOMAS RNA-Seq

Con la llegada de NGS se desarrollaron una gran cantidad de estrategias y flujos de trabajo para la integración y análisis de datos RNA-Seq, acompañadas con cientos de recomendaciones generales y particulares para su manejo (Corchete et al., 2020; Conesa et al., 2016). En la actualidad las dos estrategias fundamentales que existen para combinar datos experimentalmente dependientes o independientes son la *integración en una sola capa* (OMI) o la *integración de múltiples capas* (MOI), entendiéndose por capa, al nivel molecular que se desea integrar, ya sea por ejemplo, datos de transcriptoma con datos de proteomas, exomas o epigenomas. Por ejemplo en el caso de la OMI, se puede optar por combinar transcriptomas de diversas especies o interacciones moleculares. Las combinaciones pueden ser muy diversas, con diversos retos y niveles de complejidad de acuerdo a los datos que se desea integrar. La MOI por su parte, es muy diversa, llegando a abarcar dos o tres tipos de capas, especies de varios filos y varios fenotipos (Kolenc et al., 2021; Crandall et al., 2020). La OMI regularmente es aplicada a estudios de expresión génica diferencial (DGEs)(Thomas et al., 2019), sin embargo sus aplicaciones en plantas es muy limitada.

Para ejemplificar su aplicación con datos de plantas, para OMI a nivel transcriptómico en 2013 se integraron ~300 muestras de microarreglos de 20 estudios en arroz y arabidopsis, identificándose genes DGEs en respuesta a la sequía y estrés bacteriano (Shaik y Ramakrishna, 2013); en 2017 se utilizó para identificar la respuesta común al estrés biótico causado por aluminio en la soja, y para realizar anotación funcional para genes de arroz, respectivamente (Das et al., 2017; Childs et al., 2011); en 2018 se integraron 46 bibliotecas RNA-Seq de fresa en diferentes condiciones para evidenciar el fenómeno del aumento del transporte de hierro para el desarrollo de frutos y semillas (Shahan et al., 2018). La OMI a nivel transcriptómico en plantas ha recorrido con éxito muy variables aplicaciones, sin embargo el estudio de los mecanismos de respuesta a la infección de las plantas bajo múltiples estresores,

es escaso, encontrando un estudio realizado con microarreglos para estudiar la interacción de arabidopsis tras la infección con 7 diferentes virus vegetales (Rodrigo et al., 2012), a partir de lo cual se identificó 1 módulo genético regulado en común, donde se identificaron 7 genes sensibles a virus regulados (+) en común por 6 de los 7 virus estudiados, de los cuales 6 genes juegan un rol específico en la célula y 1 codifica una proteína similar a la *taumatina PR-5*, conocida por su participación en las respuestas de la planta a patógenos.

El crecimiento exponencial en datos ofrece un gran oportunidad para continuar con estudios de multi-sistemas desde una perspectiva más integral, se desea alcanzar una base general sobre los mecanismos de defensa de la planta ante diversos estresores fúngicos, las condiciones actuales de la globalización y el cambio climático extremo lo ameritan.

1.6.1. Métodos de normalización de datos RNA-Seq

Durante la integración de datos públicos RNA-Seq se enfrentan muchos desafíos, debido a que las bibliotecas RNA-Seq tienen diferente diseño, tamaño y composición, es decir, se deben realizar correcciones en los datos para que los valores representados sean proporcionales entre sí. Quizá el reto más importante es la profundidad de la secuenciación, ya que estas diferencias entre las bibliotecas impactan las cuantificaciones de la característica génica en estudio, que a menudo dependerá de estas covariables (Finotello y Di Camillo, 2015). Por lo tanto, es necesario normalizar los datos para reducir el sesgo técnico inducido por la secuenciación, y conservar únicamente la variabilidad biológica. Existen numerosos métodos de normalización para RNA-Seq (Zouine et al., 2017; Zyprych-Walczak et al., 2015; Dillies et al., 2013); entre los más utilizados están RPKM, FPKM y TPM (Geistlinger et al., 2021; S. Liang et al., 2018; Badet et al., 2017; C. Zhang et al., 2017; Coolen et al., 2016). Extensas revisiones con las que se cuenta hoy en día sugieren recomendaciones para diversos tipos de casos, nosotros optamos por el método de “Transcritos por millón” (TPM) por estar ampliamente probado en transcriptómica (Y. Zhao et al., 2021; S. Zhao et al., 2020; Zyprych-Walczak et al., 2015; Dillies et al., 2013; Maza et al., 2013) y ser adecuado para protocolos donde la secuenciación de las lecturas dependen de la longitud del gen. Además dicho método, se encuentra implementado de diversas herramientas de cuantificación (Patro et al., 2017) y KDB (Oh et al., 2022).

1.6.2. Preprocesamiento y estimación de los niveles de expresión

Previo a la estimación de los niveles de expresión (cuantificación), los datos son

verificados mediante un puntaje de calidad llamado *Phred*, el cual se ejecuta para identificar y corregir lecturas de baja calidad, presencia de adaptadores y secuencias repetidas, entre otros factores, esto con el objetivo de eliminar dichas lecturas de baja calidad o corregirlas de ser posible. El archivo resultante de este preprocesamiento es utilizado para la cuantificación, por lo que en lo sucesivo la secuencia de pasos a seguir estará estrechamente ligada a la estrategia y característica genética definida durante el diseño experimental. A modo de dar contexto, se pueden seguir dos estrategias, la primera cuando se tiene una referencia, o la segunda cuando se realiza una cuantificación de *novo* para identificar nuevos objetivos génicos. Para el primer caso, el proceso básico consiste en alinear, mapear y cuantificar las lecturas preprocesadas a partir de la definición de una característica génica definida en un archivo de referencia, que puede ser un genoma o un transcriptoma; la existencia y completitud de una anotación genómica es determinante para decidir la referencia del alineamiento, ya que organismos que no tengan un genoma secuenciado o lo tengan parcialmente anotado, tendrán que seguir la estrategia de *novo*. En organismos que tienen un genoma "completo", la cuantificación a nivel gen puede ser todo lo que se necesita para muchos propósitos, debido en parte a que la eliminación de variantes duplicadas y ruido de los datos puede facilitar las labores de análisis (Babarinde et al., 2019). Sin embargo, todo dependerá del objetivo.

1.6.3. Análisis con redes de coexpresión génica

Las redes de coexpressión se remontan a los primeros estudios en levadura (Wu et al., 2002; Eisen et al., 1998), seguido por organismos superiores como humanos, moscas y gusanos (Stuart et al., 2003). Sin embargo es realmente a partir de NGS que se desarrollan herramientas con métodos más sofisticados. Las redes de coexpresión en general pueden describirse como herramientas análiticas que incorporan métodos de ML para asociar genes de función desconocida con procesos biológicos, priorizar genes candidatos para enfermedades o discernir genes reguladores de la transcripción (van Dam et al., 2018). Los métodos disponibles son muy amplios y variados, y en un contexto general se representa en dos categorías, las herramientas para hacer predicciones y las utilizadas para identificar patrones (Al-Mhairat et al., 2019; Xu y Tian, 2015). Algunos ejemplos de las más utilizadas en bioinformática son WGCNA (B. Zhang y Horvath, 2005), utilizada para estudios con datos NGS de plantas, animales y humanos, CEMiTool (Co-Expression Modules identification) (Russo et al., 2018) y Coseq (Co-Expression Analysis of Sequencing Data) (Godichon-Baggioni, 2019), ambas muy utilizadas en investigaciones biomédicas (Leishmania, miocardiopatía). El análisis de coexpresión es muy poderoso, sin embargo hay que tener en cuenta que no proporcionan

información sobre causalidad, ya que no se pueden hacer inferencias directas acerca de qué genes son regulados o reguladores, por lo que diferentes estrategias deben incorporarse en el análisis de la red para permitir identificar estas asociaciones (van Dam et al., 2018).

En una red de coexpresión cada nodo representa un gen, y cada borde la correlación. Redes *no-ponderadas* (unweighted) no pueden proporcionar información sobre la fuerza de la correlación, mientras redes ponderadas (weighted) sí permiten medir la correlación entre dos nodos. Los agrupamientos identificados en la red son módulos génicos que contienen el patrón de la expresión más similar para múltiples muestras, por lo que hay que considerar en su diseño, que la relación entre la cantidad y heterogeneidad de las muestras impactará el resultado. Por ejemplo, módulos coexpresados digamos, "específicos de tejido", pueden no ser detectables en una red construida a partir de múltiples tejidos, debido a que la señal de correlación de los agrupamientos se diluirá por una falta de correlación entre los otros tejidos. Por otro lado, limitar el estudio a un solo tejido, también reduce el tamaño de la muestra, lo que disminuye el poder estadístico para detectar módulos compartidos (van Dam et al., 2018). Los métodos que no distinguen entre tejidos o condiciones deben usarse para la identificación de módulos de coexpresión comunes, mientras que la coexpresión diferencial comparando diferentes tejidos (o condiciones) será mejor para identificar módulos únicos para un tejido específico. Dentro de las técnicas disponibles para construir este tipo de redes, el "Agrupamiento Jerárquico" (Clustering/hClust) es una de las técnicas más comunes. El hClust realiza un proceso iterativo de la evaluación de las agrupaciones y se retroalimenta con los cambios en la configuración del algoritmo hasta lograr el resultado apropiado. Por lo tanto, el proceso es dinámico y cambiante. Existen diferentes métricas de correlación (*Pearson, Spearman,* etc*)* (Van Someren, 2006), y frecuentemente surgen nuevas. La selección de la herramienta por lo regular se basa en las características de los datos, la distribución, número de muestras, atípicos y la densidad esperada, ya qué por ejemplo las cuantificaciones RNA-Seq contienen valores exponenciales que difícilmente se ajustan a una distribución normal. Algoritmos para hClust tienen la ventaja de ser deterministas y trabajar bien los detalles finos sobre las relaciones entre los objetos, sin embargo, son sensibles al ruido y a los valores atípicos, además de ser computacionalmente exhaustivos. La **Tabla 1.1** resume las características de tres de las principales técnicas de agrupamiento para datos RNA-Seq.

Tabla 1.1 Técnicas de agrupamiento más utilizadas en el estudio de patrones de coexpresión genética.

Descripción	Requisitos	Fortalezas	Debilidades
Agrupación particional (Partitional clustering)			
Divide los objetos de datos en grupos que no se superponen. Ningún objeto puede ser miembro de más de un clúster y cada clúster debe tener al menos un objeto.	Requiere que se especifique el número de clusters (k). Muchos algoritmos de este tipo funcionan a través de un proceso iterativo para asignar subconjuntos de puntos de datos en 'k' grupos. Ejemplos de estos algoritmos incluyen a las **K-means, K-medoids, PAM, CLARA y CLARANS.**	Funcionan bien cuando los clusters tienen **forma esférica**. Son **escalables** con respecto a la complejidad del algoritmo.	Ambos algoritmos son **no deterministas*.** No son adecuados para clústeres con formas complejas y diferentes tamaños. Se descomponen cuando se usan con grupos de diferentes densidades.
Agrupación jerárquica (Hierarchical clustering)			
Determina las asignaciones de grupos mediante la construcción de una jerarquía. Tiene dos enfoques: la **agrupación aglomerativa** (Agglomerative clus - ascendente), la cual fusiona los dos puntos que son más similares hasta que todos los puntos se hayan fusionado en un solo grupo; la **agrupación dividida** (Divisive clus – descendente), la cual comienza con todos los puntos como un grupo y divide los grupos menos similares en cada paso hasta que solo quedan puntos de datos únicos. Estos métodos producen una jerarquía de puntos basada en árboles llamada **dendrograma.**	A menudo requieren que se especifique el número de clusters (k). Los grupos se asignan cortando el dendrograma a una profundidad específica que da como resultado 'k' grupos de dendrogramas más pequeños. Ejemplos de estos algoritmos incluyen a **BIRCH, CURE, ROCK, Chameleon y WGCNA.**	La agrupación jerárquica es un proceso **determinista*.** A menudo revelan los detalles más finos sobre las relaciones entre los objetos de datos. Proporcionan un dendrograma interpretable.	Son computacionalmente costosos con respecto a la complejidad del algoritmo. Son **sensibles al ruido y a los valores atípicos.**
Agrupación basada en densidad (Density-based clustering)			
Determina las asignaciones de clústeres en función de la densidad de puntos de datos en una región. Los clústeres se asignan cuando hay altas densidades de puntos de datos separados por regiones de baja densidad.	No requiere que el usuario especifique la cantidad de agrupaciones (k). Existe un parámetro basado en la distancia que actúa como un umbral suave que determina qué tan cerca deben estar los puntos para ser considerados miembros del clúster. Ejemplos de estos algoritmos son: **DBSCAN, OPTICS y Mean-Shift**	Se sobresalen en la identificación de grupos de formas no esféricas. Son resistentes a los valores atípicos.	No son adecuados para agruparse en espacios de gran dimensión. Tienen problemas para identificar grupos de diferentes densidades.

*No determinista: significa que podrían producir resultados diferentes a partir de dos ejecuciones separadas incluso si las ejecuciones se basaran en la misma entrada.

**Determinista: significa que las asignaciones de clústeres no cambiarán cuando se ejecute un algoritmo dos veces en los mismos datos de entrada. Extraído de (Xu y Tian, 2015).

1.6.3.1 Red de coexpresión ponderada wgcna

Las herramienta WGCNA (Weighted Gene Correlation Network Analysis) (Langfelder y Horvath, 2008) contiene una colección de funciones para diseñar, construir y analizar redes de coexpresión génica, se basa en hClust aglomerativo y se ha destacado por su potencial de análisis para generar inferencias para un amplio número de organismos (Shahan et al., 2018; Childs et al., 2011; ; Ghazalpour et al., 2008; Carlson et al., 2006). En WGCNA se pueden construir redes ponderadas y no ponderadas, además se puede definir la dirección de los perfiles de los nodos (genes) a partir de la configuración de la red (signed-ntw o unsigned-ntw) para conocer sí los agrupamientos están regulados positiva o negativamente.

Cualquier configuración de red se construye a partir de una matriz de adyacencias que nos dicta la relación existente entre las instancias de entrada (los valores de expresión). En la red ponderada (weighted) se usa una función de adyacencia de potencia β que define las fuerzas de conexión entre pares de genes. Por lo tanto, la fuerza de la conexión es igual a la correlación elevada a una determinada potencia β, llamada umbral suave. En la red no ponderada (unweighted) se establece un umbral estricto del valor absoluto de la matriz de correlación, por lo que se utiliza un punto de corte estático. WGCNA agrupa los genes en módulos utilizando una aproximación a una matriz de superposición topológica (TOM). En la red ponderada, la TOM generaliza la matriz utilizando los números reales de la fórmula para el cálculo de la matriz de adyacencias, qué en una red no ponderada son valores binarios (B. Zhang y Horvath, 2005). Finalmente dependiendo de la configuración, se obtendrán agrupamientos definidos naturalmente con valores de expresión lo más similares posibles entre sus miembros, y lo más diferentes posibles entre los distintos agrupamientos.

1.7. BASES DE DATOS, REUTILIZACIÓN E INTEGRACIÓN DE DATOS

La enorme montaña de datos en las BDs ómicas y el despliegue de nuevas herramientas de análisis ha alentado el interés en la reutilización de los datos en todos los niveles (Wu et al., 2021; Crandall et al., 2020; Subramanian et al., 2020). Sin embargo puede notarse que el crecimiento de las BDs no ha contribuido directamente a la reutilización de los mismos, por lo tanto ahora tenemos enormes colecciones de datos parcialmente exploradas, debido a que por décadas han sido generados datos para un-solo-uso. La reutilización se estancó debido a la falta de estrategias para su integración, ya que no es trivial y contiene datos “desorganizados” con escalas muy variables (Jamil et al., 2020), en diferentes

magnitudes y con diversos tipos de sesgos (Y. Zhao et al., 2021; Corchete et al., 2020; C. Zhang et al., 2017; Conesa et al., 2016). Por tal motivo, tras más de una década se sigue priorizando la generación de datos con respecto a la reutilización. La reutilización en las ómicas es muy relevante porque cumple con diversos objetivos, favorece y acorta los tiempos de investigación, mejora la potencia de las pruebas y ofrece distintas perspectivas, sin omitir, que generar datos biológico conlleva gestiones presupuestarias y legales, procesos largos para la generación de las muestras, y muchas veces restricciones geográficas de acceso para su adquisición.

A pesar de las complejidades existentes en la integración de datos ómicos, para datos de plantas se ha logrado desplegar una cantidad decente de BDs para entornos web con distintos niveles de conocimiento (knowledge databases / KBDs) y flexibilidad para el usuario (user-friendly). Algunos ejemplos son ATTEND-II (http://atted.jp/), Expression Angler (http://bar.utoronto.ca/ExpressionAngler/), AtCAST (http://atpbsmd.yokohama-cu.ac.jp/cgi/atcast/home.cgi), Aranet (https://www.inetbio.org/aranet) y PlaD (http://systbio.cau.edu.cn/plad/index.php). Sin embargo, no se debe ignorar el hecho de que pesé a su enorme utilidad para la comunidad científica, todas las BDs ofrecen flexibilidad limitada para comparar y extraer información, además la mayoría están estrictamente ligadas a los datos pre-integrados y bajo un solo método de análisis. En un sentido más amplio, lo deseable sería acceder a configuraciones más personalizadas o ampliar los recursos de enseñanza para fortalecer la capacidad de la comunidad científica para explotar estos datos, ya qué comunemente encontraremos estrategias muy *específicas* (Ibrahim et al., 2021; Thomas et al., 2019; X. Liang et al., 2018; Shaik y Ramakrishna, 2013), que resultan finamente ligadas al caso de estudio en cuestión, y por lo tanto puede resultar difícil adecuarlas a nuestro estudio, o muy *generalizadas* (Corchete et al., 2020, Conesa et al., 2016), en formatos de revisiones que solo permiten comparar las diversas estrategias. Por ello, se resalta la importancia de seguir trabajando en el diseño de más flujo de análisis para integrar datos RNA-Seq a nivel multi-sistema, respondiendo a la demanda de alternativas que permitan extraer, combinar y asociar críticamente diferentes conjuntos de datos (Fondi y Liò, 2015; Hughes, 2015), imperantes para conducir estudios centrados en multisistemas (Oh et al., 2022; Crandall et al., 2020).

JUSTIFICACIÓN

Es de vital importancia comprender los mecanismos de defensa de las plantas a múltiples estresores fúngicos desde una perspectiva multisistémica, debido a que las enfermedades fúngicas ocupan los primeros lugares de pérdidas agrícolas en los principales cultivos de ingesta humana; además se prevé, que el impacto del ambiente en que se desarrollan estas enfermedades detonará la aparición de más enfermedades derivado del cambio climático extremo y los efectos de la globalización. Por tanto en esta tesis, estudiamos los mecanismos de defensa de *A thaliana* a nivel multi-sistema en relación con diversos patógenos fúngicos de relevancia agrícola.

Así mismo para responder a la demanda de alternativas metodológicas para combinar y asociar críticamente diferentes conjuntos de datos, imperante para conducir estudios centrados en multisistemas, se diseñó e implementó un marco metodológico novedoso para integrar y analizar datos públicos RNA-Seq experimentalmente independientes. Se utilizaron datos de la planta *A thaliana*, sana e infectada con algún hongo *Ascomycetes,* debido a la completitud y consistencia de su genoma y anotación, que permite tener una base sólida de estudio; la interacción con el hongo se decidió en función de los reportes globales de patogenicidad y la disposición de transcriptomas en las BDs del SRA-NCBI.

HIPÓTESIS

Existe un agrupamiento de genes core asociados al mecanismo de respuesta de plantas de *A. thaliana* que responde en forma ubicua a infecciones por diferentes hongos patógenos Ascomycetes.

OBJETIVO GENERAL

Identificar genes de la respuesta consenso en plantas de *A thaliana* infectadas con distintos hongos *Ascomycetes* a partir de un meta-análisis con datos de transcriptomas RNA-Seq disponibles en forma pública.

OBJETIVOS ESPECÍFICOS

1. Preparar un diseño experimental con datos de transcriptomas RNA-Seq disponibles en forma pública para plantas de *A thaliana* infectadas con distintos hongos Ascomycetes durante el umbral de respuesta PAMP-triggered immunity (PTI).

2. Definir una metodología para integrar los conjuntos de datos RNA-Seq independientes en matrices de expresión para su análisis con métodos de redes de coexpresión ponderadas WGCNA.

3. Identificar agrupamientos de genes de consenso en las redes ponderadas relacionados con la respuesta de defensa de la planta ante la infección causada por estos hongos.

ESTRATEGIA EXPERIMENTAL

A partir de la identificación y extracción de transcriptomas RNA-Seq en las BDs públicas de la planta arabidopsis en condición sana o infectada por hongo ascomycete, se crearon perfiles de expresión génica que se integraron en redes de coexpresión, una para plantas sanas y otra para infectadas. En ambas redes se identificaron agrupamientos (módulos) con *coeficiente de correlación* de moderado a alto (>0.50 / *Pearson*) para el control y para estudiar la respuesta "*consenso"* de la planta ante la infección con diversos patógenos *ascomycetes*; los módulos *consenso* considerados en el estudio están altamente diferenciados de los módulos de control (>75%) en relación con los genes asignados; con estos módulos se realizó análisis de sobre-representación génica y de enriquecimiento (**Figura 1.7**) para estudiar su relación con los mecanismos de defensa de la planta; adicionalmente se realizó análisis de polimorfismos de secuencia, localización cromosomal y análisis de funciones compartidas.

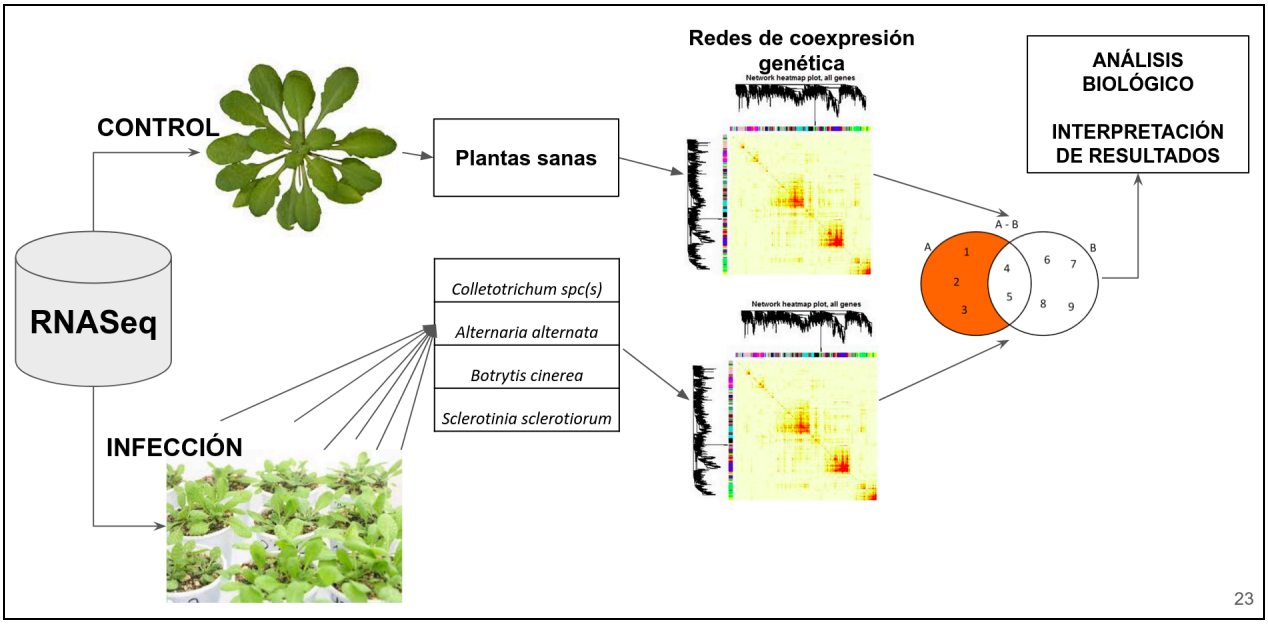


Figura 1.7. Estrategia experimental.

METODOLOGÍA

A partir de la identificación y extracción de transcriptomas RNA-Seq del SRA-NCBI de la planta *A thaliana* en condición sana o infectada por hongo *ascomycete* dentro las primeras 48 horas (hpi), se crearon perfiles de expresión génica que se integraron en dos redes de coexpresión ponderadas, utilizando la herramienta WGCNA (Weighted Gene Correlation Network Analysis) (Langfelder y Horvath, 2008), se creó una para plantas sanas y otra para infectadas. En la red de plantas sanas se extrajeron los agrupamientos génicos (módulos) con *coeficiente de correlación* (CC)>0.75 (*Pearson*) para *control*, y en la red de plantas infectadas se extrajeron módulos *consenso* con CC>0.50 para su análisis, los cuales se filtraron con lógica de conjuntos contra los módulos de *control* para conservar únicamente los módulos altamente diferenciados (>75% de sus genes únicos). Posteriormente se realizó anotación GO y anotación de clúster enriquecido a diversos niveles de estridencia estadística contra 9 BDs de diferente orden, se extrajeron únicamente los clústeres enriquecidos con valor-p<0.05, se fusionaron los genes repetidos en múltiples clústers (**Figura 1.8**). Finalmente, con los genes sobrevivientes del módulo de *consenso*, se estudiaron sus vías biológicas, procesos y componentes moleculares identificados en relación con los mecanismos de defensa conocidos, se analizaron los polimorfismos de secuencias de los genes, su localización cromosomal y funciones compartidas. Se sintetizaron los hallazgos más relevantes.

Se eligió a la planta arabidopsis debido a que posee un genoma completo sin cambios

desde el 2010 con una reanotación mejorada (Cheng et al., 2017) liberada en 2016. El *phylum* de hongo a*scomycete* se determinó en función de que muchos hongos patógenos vegetales causantes de enfermedades muy devastadoras para muchos cultivares se encuentran en este grupo. Además, los miembros en este *phylum* existen prácticamente en cualquier entorno sobre la tierra, llegando a comprender más del 60% de los patógenos descritos en la actualidad (Cavalier-Smith,T., 1998). Se estableció un umbral de infección planta-patógeno entre las 0 y las 48 horas esperando capturar el patrón de la respuesta consenso activo de la planta por el reconocimiento de patrones moleculares asociados a patógenos (PAMP-triggered immunity / PTI), la cual constituye la primera capa de inmunidad vegetal que restringe la proliferación de patógenos, y es descrito como un umbral de respuesta conservado entre las familias, no existiendo una diversificación extrema de los receptores y efectores tanto dentro como entre especies (Macho y Zipfel, 2014; Dodds y Rathjen, 2010; Jones y Dangl, 2006). Otros criterios técnicos para la selección de las muestras a nivel técnico se establecieron con la finalidad de reducir el sesgo técnico derivado del método de secuenciación utilizado (**Tabla 1.2**).

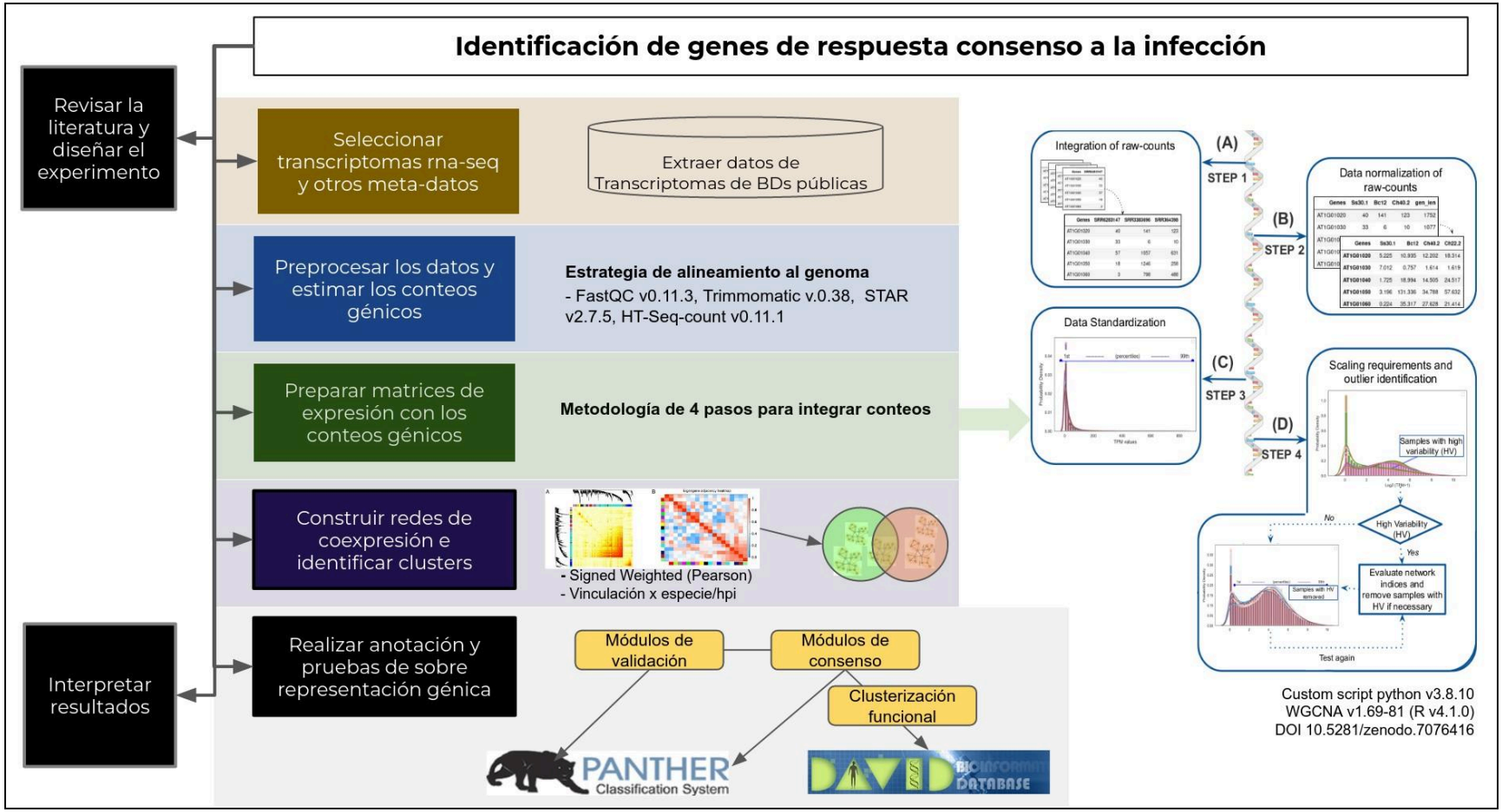


Figura 1.8. Metodología. Primero se identificaron y extrajeron transcriptomas RNA-Seq del SRA-NCBI de la planta *A thaliana* en condición sana o infectada por hongo *ascomycete* dentro las primeras 48 horas (hpi), después se preprocesaron los datos y se crearon perfiles de expresión, los cuales se integraron con el marco metodológico de 4 pasos. Con estas matrices se crearon las redes de coexpresión ponderadas con la herramienta WGCNA (Langfelder y Horvath, 2008). En ambas redes se identificaron agrupamientos génicos con *coeficiente de correlación* medio a moderados para *control*, y análisis de la respuesta consenso a la infección. Los agrupamientos se filtraron con lógica de conjuntos para conservar únicamente los módulos altamente diferenciados (>75% de sus genes únicos). Posteriormente se realizaron validaciones basadas en anotación GO y anotación de clúster enriquecido para el módulo de *consenso*. Se estudiaron las vías biológicas, procesos y componentes moleculares identificados en relación con los mecanismos de defensa conocidos, polimorfismos de secuencias y localización cromosomal.

Tabla 1.2. Criterios de selección de las muestras.

Biológicos		**Técnicos**	
Huésped:	*A thaliana*	Origen:	Transcriptoma
Especie:	Col-01	Estrategia:	RNA-Seq
Patógeno:	Hongo *ascomycete*	Plataforma de sec:	Illumina
RNA:	Extraído de hoja	Tamaño de la librería:	>5 Gpb de lecturas
h.p.i:	0-48	Long. secuencia:	>100 pb

NOTA. Se diseñó e implementó un marco metodológico novedoso para la fase de integración y preparación de las matrices de expresión, el cual se presenta en el **Capítulo II.** En el **Capítulo III** se presentan los resultados de la respuesta de defensa consenso de arabidopsis ante los patógenos ascomycetes incluidos en este estudio.

CAPÍTULO II

ANÁLISIS MULTISISTEMA EN ARABIDOPSIS PARA LA IDENTIFICACIÓN DE PATRONES GÉNICOS DE LA RESPUESTA CONSENSO AL ESTRÉS FÚNGICO

2.1. INTRODUCCIÓN

En la última década el método RNA-Seq se posicionó como una estrategia básica para muchos grupos de investigación a nivel global debido en parte a la precisión del método y la caída de los precios por base secuenciada, llenando a nivel exponencial las BDs ómicas públicas (Sayers et al., 2020), detonando una avalancha de métodos y herramientas para dar manejo, decenas de recomendaciones tanto generales como específicas para su manejo (Corchete et al., 2020; Conesa et al., 2016). La transcriptómica, seguida por la proteómica, son dos de las ómicas más utilizadas en la investigación bioinformática, por lo que perfeccionar y proponer métodos alternativos para su uso sigue vigente, incluso trás casi dos décadas de su lanzamiento. La transcriptómica no esta vigente, sino que además hoy es incorporada a estudios multi-omicos o de multi-sistemas que persiguen mejorar el poder estadístico de las pruebas y la capacidad del quehacer científico (Thomas et al., 2019).

Pese a los muchos casos que existen de la aplicación multisistémica con transcriptomas en muchas áreas científicas, como la biomédica, su implementación con datos de plantas es escasa. Algunos casos de estudio con datos de plantas donde se ha empleado con éxito la integración de multisistemas (Tsushima et al., 2019; Shahan et al., 2018; Badet et al., 2017; Iancu et al., 2012; Childs et al., 2011; Ghazalpour et al., 2008; Carlson et al., 2006) han impulsado un renovado interés en los predictores con datos de expresión, sin embargo, el reúso de datos no es trivial y ha estado mermado por la carencia de estrategias flexibles para la apropiada integración y manejo de los datos. La integración de perfiles de expresión procedentes de datos independientes (RNA-Seq) presenta desafíos importantes (Jamil et al., 2020); (Y. Zhao et al., 2021; Corchete et al., 2020; C. Zhang et al., 2017, 2016; Conesa et al., 2016) que requieren la aplicación de diversas estrategias de normalización, estandarización y filtrado de los datos para lograr una adecuada representación de los mismos.

Por lo tanto, con la finalidad de probar la hipótesis que antecede a este trabajo de investigación, y promover la reutilización de datos RNA-Seq experimentalmente independientes alojados en las BDs públicas, diseñamos e implementamos un marco metodológico de cuatro

pasos para integrar conteos de expresión para su análisis con redes de coexpresión génica. Con la finalidad de facilitar la apropiación del marco metodológico para su uso habitual en el análisis exploratorio, adoptamos conceptos de estadística comunes (medidas de tendencia central y de distribución) para definir las métricas de evaluación (Larson, 2006); 2) así como gráficos de distribución con Kernel-Density-Estimation (KDE) para facilitar las visualizaciones e interpretación, y seguimiento a los cambios en la variabilidad técnica e identificación de atípicos (Waskom, 2021); y 3). También utilizamos los índices de red del modelo de topología libre de escala (SFT) disponibles en la herramienta WGCNA (Langfelder y Horvath, 2008) para evaluar los coeficientes de correlación alcanzados por la matriz de expresion procesada (normalizada y estandarizada) (**Figura 2.1**). La estrategia de integración de los datos conducida, demostró ser eficaz para utilizarse con clustering jerárquico. Con más del 80% de los valores de expresión que sobrevivieron al análisis, logramos identificar agrupamientos de genes coexpresados consistentes con la respuesta esperada a nivel fenotipo.

2.2. MATERIALES Y MÉTODOS

2.2.1 Datos, preprocesamiento y estimación de la expresión génica

Datos RNA-Seq de *A. thaliana,* plantas sanas y bajo estrés biótico fueron identificadas en el *archivo de lecturas sin procesar* del (SRA) (Leinonen et al., 2011) del Centro Nacional de Información Biotecnológica (NCBI). Los datos identificados se descargaron con la herramienta SRA-Toolkit-Box (SRA-Toolkit, 2021); se verificó y corrigió la calidad de las lecturas con FastQC v0.11.5 (Babraham Bioinformatics, 2010) y Trimmomatic v.0.38 (Bolger et al., 2014), estableciendo un *Phred Score* > 20 y una longitud mínima de secuencia > 40 pb.

Las lecturas de los archivos preprocesados se cuantificaron siguiendo la estrategia de alineamiento al genoma, utilizando el genoma de arabidopsis TAIR10 (Lamesch et al., 2012) (GenBank accessions CP002684 – CP002688) y la anotación Araport11 (Cheng et al., 2017) para genes codificantes de proteínas. El alineamiento se realizó con STAR v2.7.5 (Dobin et al., 2013) ajustando los parámetros *sjdbOverhang=92* y *genomeSAindexNbases=7* para adecuarlos al tamaño del genoma, y los parámetros *alignIntronMin=8* y *alignIntronMax=1999* para ajustar la longitud mínima y máxima de intrones (Chang et al., 2017). La calidad de los alineamientos (SAM/BAM) (Li et al., 2009) se evaluó con HTSeq-qa v0.11.1 (Anders et al., 2015); se verificó el porcentaje de cobertura, y se estimó la abundancia de expresión (perfiles) no normalizada (conteos crudos) con HTSeq-count v0.11.1 (Anders et al., 2015).

2.2.2 Marco metodológico para la Integración de conteos en matrices de expresión

Se integraron los conteos de expresión obtenidos mediante el marco metodológico de cuatro pasos a continuación descrito (**Figura 2.1**), el cual se complementa con la herramienta WGCNA (Langfelder y Horvath, 2008). Las tres métricas que se utilizan para dar seguimiento a los cambios son la desviación estándar (σ), la media (x) y el rango (R) (Larson, 2006); se utilizan histogramas con ajuste de densidad del kernel (KDE) (Waskom, M.L., 2022; VanderPlas Jake, 2016) para obtener las visualizaciones y hacer los ajustes pertinentes. La metodología se implementó en Python v3.8.10 (Welcome to Python.org, 2019) y está disponible en Github https://github.com/cyntsc/RNA-Seq-raw-integration (DOI 10.5281/zenodo.7076416) (**Table 2.1**).

2.2.2.1 Paso 1: integración de conteos de expresión RNA-Seq

Se integraron los conteos sin normalizar en una matriz de expresión por condición (planta sana e infectada) utilizando el identificador único del gen. Se eliminaron los genes con valor de expresión igual a cero entre las muestras. Se calculó la desviación estándar, la media y el rango (σ, x, *R),* y se generaron los histogramas con ajuste KDE (**Figura 2.1A**). Se utilizó el script 1_Step1_integrating_raw_counts.ipynb (**Table 2.1**).

2.2.2.2 Paso 2,3: control de la variabilidad (normalización y estandarización)

Se normalizaron los valores de las matrices de expresión a transcritos por millón (Transcripts-Per-Million / TPMs) (Zhao et al., 2021) utilizando la biblioteca para python TPM normalization v0.9.1 (c, 2020). Se recalcularon las métricas de evaluación (α,x,*R*) y los gráficos con KDE (**Figura 2.1B**). Se utilizó el script 2_Step2_TPM_normalization.ipynb (**Table 2.1**). Posteriormente, se calcularon los percentiles y las distribuciones de cada muestra en las matrices de expresión, para identificar los valores extremos de las colas, y eliminar los genes que tiene baja variabilidad, están subrepresentados o contienen valores exponenciales extremos. Se definió punto de corte para los genes con valores TPM<0.1 en al menos el 70% de las muestras, y los genes con valores TPM sobre el 99th percentil. Se recalcularon las métricas de evaluación y los histogramas con KDE (**Figura 2.1C**). Se utilizó el script 3_Step3_TPM_standardization.ipynb (**Table 2.1**).

2.2.2.3 Paso 4: identificación de muestras atípicas

Las matrices de datos resultantes se transformaron a $\log_2$(TPM+1). Con el objetivo de resaltar el efecto negativo de la alta variabilidad en las redes, se insertaron 3 muestras con distribuciones atípicas (control negativo). Se recalcularon las métricas y se generaron los gráficos con KDE. Se implementó con el script 4_Step4_Log2_scale.ipynb (**Figura 2.1D**).

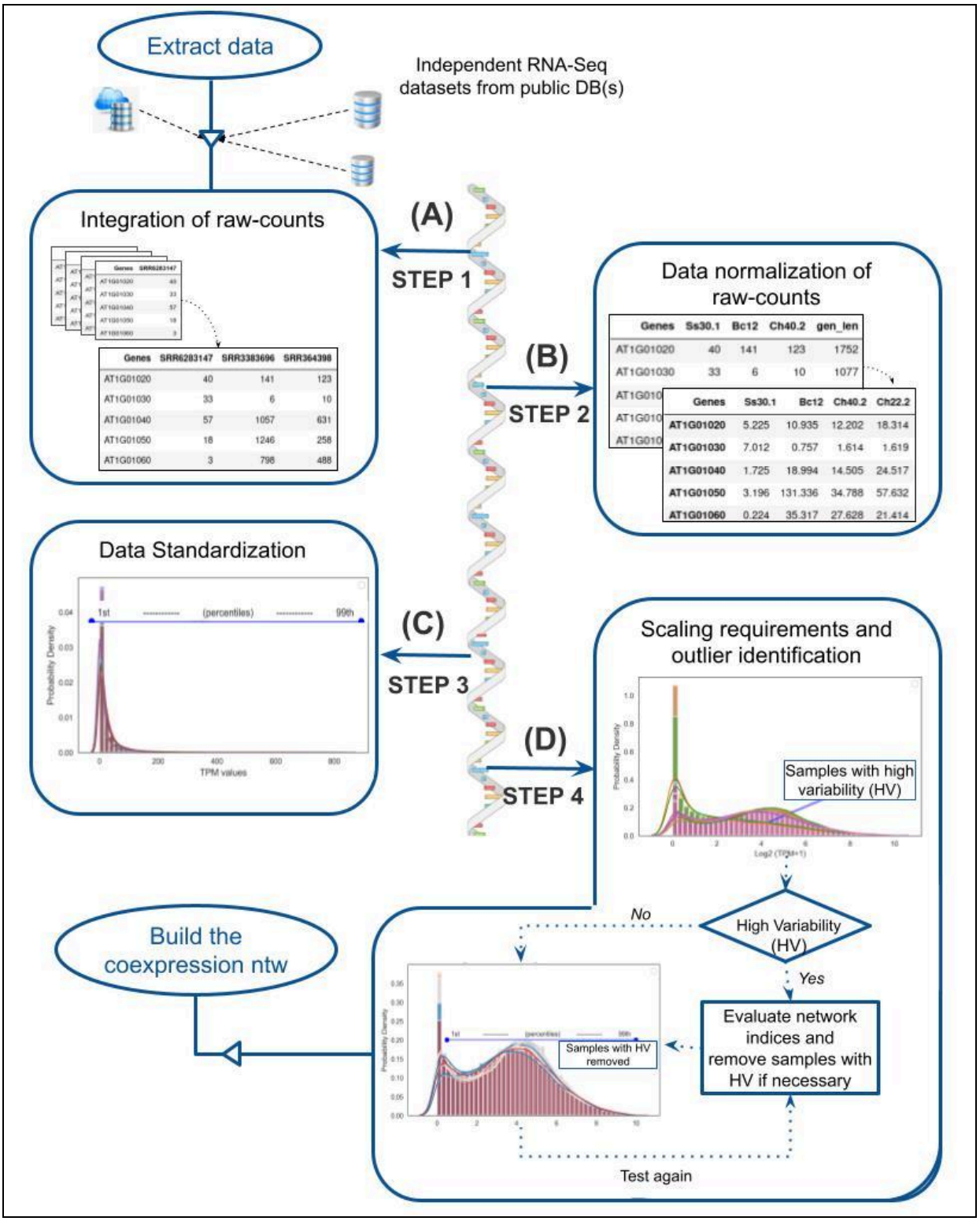


Figura 2.1. Marco metodológico para integrar y estandarizar conteos de datos RNA-Seq. (A) Los conteos sin normalizar se integran en matrices de datos y se calculan las métricas de evaluación de referencia (desviación estándar (σ), media (μ) y rango (R)). (B) Se normalizan los conteos en TPMs y se recalculan las métricas. (C) Se calculan percentiles para identificar puntos de corte (umbrales). Genes con valores

por debajo del 1er percentil 1 y por encima del 99vo son eliminados y se recalculan las métricas. (D) Se transforman los valores de expresión a $\log_2$ (TMP+1) para reducir la escala de los datos y resaltar distribuciones atípicas. Se introduce un control negativo y se recalculan las métricas.

2.2.3 Construcción de las redes de coexpresión ponderadas con WGCNA

Con las matrices resultantes se construyeron dos redes de coexpresión ponderadas con signo (signed-ntw) con el método de *Pearson*, una para plantas sanas y otra para las infectadas*;* se utilizó WGCNA v1.69-81 (Langfelder y Horvath, 2008) para R (R: The R Project for Statistical Computing, s. f.). Con la función *pickSoftThreshold* se calcularon los índice de red con ajuste al modelo de red SFT. Se analizó el coeficiente de correlación (r^2) y la media de conectividad de nodo (NCM), fijando corte en el umbral de potencia β~0.80. Se utilizó disimilitud basada en TOM para asignar los genes a los agrupamientos, fijando un mínimo de cluster=20; posteriormente, cada red se fusionó (merged) a 0.1 distancias para reagrupar los agrupamientos con valores más cercanos. Los agrupamientos (módulos génicos) obtenidos se vincularon en mapas de calor (heatmaps) por hora del tratamiento simulado para la red de control, y por especie de hongo y hora de infección para la red de las plantas infectadas. Se calculó el PCA y valor-p de los módulos utilizando las funciones *moduleEigengenes*, *cor* y *corPvalueStudent*. Se identificó en ambas redes módulos con coeficientes de correlación (CC) $r^2>0.75$ y CC $r^2<0.75$, y valor-p<0.05. En la red de plantas infectadas se aplicó un criterio flexible de búsqueda no menor a $r^2<0.50$ (medio-moderado). Con comparaciones lógicas se identificaron los módulos de la red de plantas infectadas mejor diferenciados de los controles (>75% de sus genes únicos). Los scripts de implementación se encuentran en la **Tabla 2.1**.

2.2.4 Validación y anotación funcional de los agrupamientos

Para validar sí los módulos identificados en las redes se corresponden con los fenotipos incluidos, se realizó prueba de sobrerrepresentación de ontología génica (GO) en la plataforma PANTHER v17.0 (Thomas et al., 2022; Mi et al., 2019) --released 2022-02-22 (**Tabla 2.1**). Se utilizó la prueba *binomial* con corrección *Bonferroni* para identificar los términos asignados a las distintas categorías GO (The Gene Ontology Consortium, 2021). Se utilizó la anotación de arabidopsis araport11 (Cheng et al., 2017) como referencia para la prueba en las categorías de función molecular (MF), proceso biológico (BP) y componente celular (CC) de GO Slim (versiones reducidas de GO que simplifican las operaciones de anotación y proporcionar una descripción general de las funciones y procesos), salvo los casos donde no se obtuvo hit en las

versiones slim, se realizó la prueba en la versión GO completa. Se presentan únicamente los resultados de Bonferroni-corregido para las clases sobrerrepresentadas con valores-p<0.05.

Tabla 2.1. Bases de datos, herramientas bioinformáticas y código fuente.

Bases de datos para extraer datos masivos de RNA-Seq y archivos complementarios		
BDs o archivo	**Descripción**	**Sitio web**
SRA-NCBI DB	Bulk RNA-Seq data	https://www.ncbi.nlm.nih.gov/sra
TAIR10 Genome Fasta File	TAIR10 genome release 2010	https://www.arabidopsis.org/download/index-auto.jsp?dir=/download_files/Sequences
Araport11 Genome Annotation (GFF File)	Genome Annotation Release 2016	https://www.arabidopsis.org/download/index-auto.jsp?dir=/download_files/Sequences
Herramientas bioinformáticas para estimar los conteos sin procesar RNA-Seq		
Herramienta	**Objetivo**	**website**
SRA-Toolkit	SRA accessions	https://hpc.nih.gov/apps/sratoolkit.html
FastQC	Quality control analysis	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trimmomatic	Preprocessing tasks	http://www.usadellab.org/cms/?page=trimmomatic
STAR	Read alignment	https://github.com/alexdobin/STAR
HT-Seq-qa HT-Seq-count	SAM quality test Gene estimation	https://htseq.readthedocs.io/en/master/
Metodología de 4 pasos para la integración de los conteos sin procesar RNA-Seq Python scripts (v3.8.10) https://github.com/cyntsc/RNA-Seq-raw-integration DOI 10.5281/zenodo.7076416		
PASOS	**Nombre del script y descripción**	
PASO 1: Raw-count integration PASO 2: TPM normalization PASO 3: Data standardization PASO 4: Data log transformation and atypical sample identification Script to extract gene lengths for step 2. Script to extract gene modules for annotation tasks	All the scripts are available on https://github.com/cyntsc/RNA-Seq-raw-integration 1_Step1_integrating_raw_counts 2_Step2_TPM_normalization.ipynb 3_Step3_TPM_standardization.ipynb 4_Step4_Log2_scale.ipynb Gene_length_extraction_from_GTF.ipynb Venn_diagram_genes_in_ceros.ipynb 6_modules_percentual_differentiation.ipynb	
Implementación de redes e identificación de módulos génicos consenso R scripts (v4.1.0) https://github.com/cyntsc/RNA-Seq-raw-integration DOI 10.5281/zenodo.7076416		
Nombre Script	**Descripción**	

01_Healthy_SignedNtw_D.R	Network implementation for a signed-ntw for the healthy plants
01_Infected_SignedNtw_E.R	Network implementation for a signed-ntw for the infected plants
02_Healthy_GS_MM_24hpi.R	Gene-Significance and Module-Membership for the healthy network
02_Infected_GS_MM_24hpi.R	Gene-Significance and Module-Membership for the infected network
Recursos de anotación de ontología de genes	
Panther v17.0	http://pantherdb.org/webservices/go/overrep.jsp (Validaciones)
DAVID v6.8	https://david.ncifcrf.gov/ (clusterización funcional de enriquecimiento)

2.3. RESULTADOS

Con la finalidad de organizar y facilitar la presentación de resultados, en este capítulo (II) se presenta, el proceso de obtención e integración de los datos, incluidos los resultados del marco metodológico (**Figura 2.1**) ideado para integrar los datos RNA-Seq para su análisis con redes de coexpresión génica; el cual se propone como una alternativa novedosa flexible para la identificación de patrones génicos (módulos) altamente diferenciados en datos públicos NGS experimentalmente independientes. Para consultar directamente los resultados de los patrones génicos activos encontrados en la respuesta de defensa de *A. thaliana* que son comunes ante el ataque de diversos patógenos, favor de consultar el Capítulo III.

2.3.1 Selección de datos y perfiles de expresión

Se procesaron un total de 25 bibliotecas de datos RNA-Seq de tejido de hoja de la planta *A. thaliana* extraídos de las BDs del SRA (Leinonen et al., 2011) que corresponden a cuatro proyectos independientes (PRJNA148307, PRJNA315516, PRJNA593073 y PRJNA418121). Ocho transcriptomas son de plantas sanas (control) y diecisiete son de plantas infectadas por hongos (tratamientos), de las cuales ocho están infectadas con *Colletotrichum higginsianum* (Ch), seis con *Botrytis cinerea* (Bc) y tres con *Sclerotinia sclerotorium* (Ss). El rango de recolección de muestras va de las 12 a 30 horas para el control, y de las 12 a 40 horas para los tratamientos. Con la finalidad de reducir parte del sesgo técnico, todos los transcriptomas se secuenciaron en plataformas Illumina, el número de lecturas oscila entre los 10 y 30 millones de pb, y las longitudes de lectura van de las 93 pb a las 150 pb. Las muestras de control se etiquetaron como healthy12, healthy12.1, healthy18, healthy18.1, healthy24, healthy24.1, healthy30 y healthy30.1, y las muestras con los tratamientos como Ch22, Ch22.1, Ch22. 2, Ch22.3, Ch40, Ch40.1, Ch40.2 y Ch40.3 (*C higginsianum*); Bc12, Bc12.1, Bc18, Bc18.1, Bc24 y Bc24.1 (*B cinerea*); y Ss30, Ss30.1 y Ss30.2 (*S sclerotorium*) (**Tabla 2.2**) (**Suplementario 2.1**).

Filtrados los datos por calidad de secuencia Phred>30, se alinearon las lecturas a la referencia genómica para arabidopsis TAIR10 (Lamesch et al., 2012), lográndose coberturas de alineamientos único entre el 97.5% y 98.3% en las muestras de control, y entre el 77.4% y 98.2% en los tratamientos, exceptuando las muestras Ss30, Ss30.1 y Ss30.2, incluidas a propósito como control negativo para contrastar las curvas de distribución de los conteos; dichas muestras tiene porcentajes de cobertura<37% (**Figura 2.2**), ya que muchas de sus lecturas se encuentran asociadas con la respuesta de los hongos, lo cual es conveniente para contrastar el efecto de los ajustes realizados en la integración de datos para eliminar ruido.

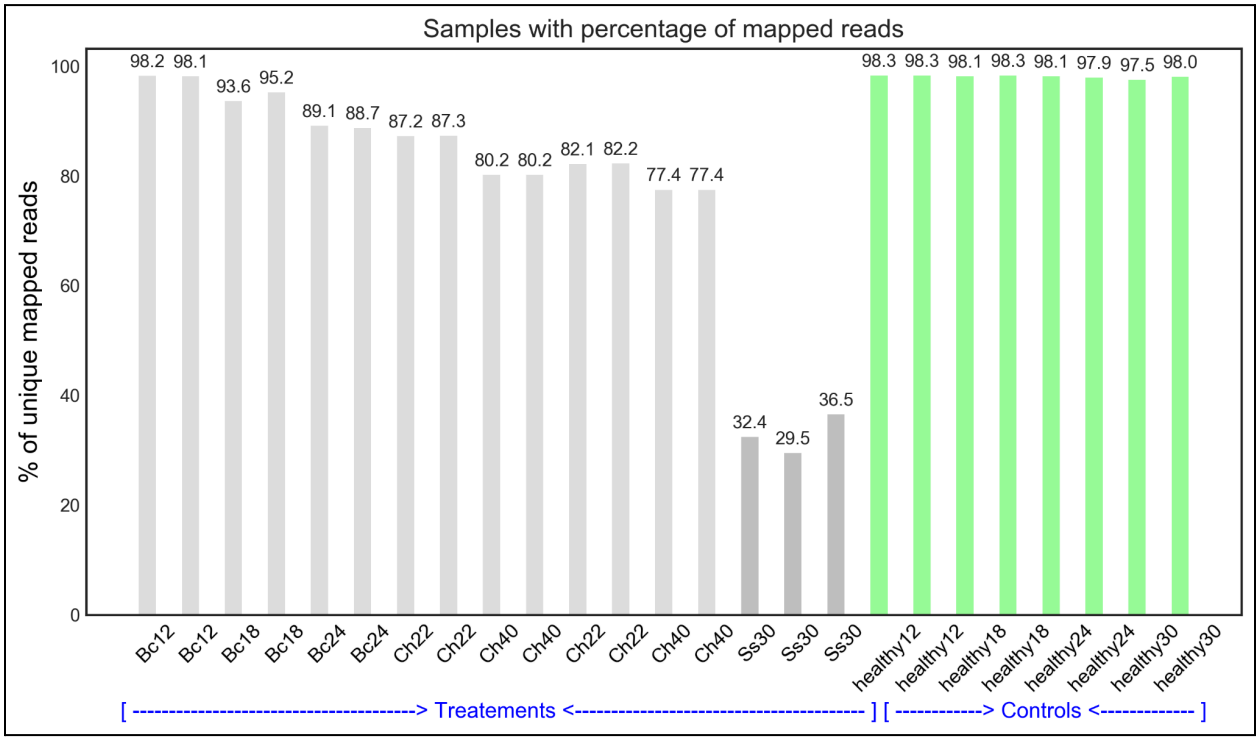


Figura 2.2. Porcentajes de cobertura de alineación. Porcentaje de lecturas asignadas únicas al genoma de arabidopsis (TAIR10) de muestras para el control y los tratamientos.

Tabla 2.2. Transcriptomas RNA-Seq descargados del SRA-NCBI.

Bioproject	Muestra ID	hpi (*)	Accession	Diseño	Lecturas (M)	%Lecturas-limpias	%Lecturas-Alineadas
TT: *A. thaliana* with *C. higginsianum* PRJNA148307 (Tsushima et al., 2019; O'Connell et al., 2012)	Ch22	22	SRR364389	SE	12.6	92.12	87.17
	Ch22.1		SRR364390	SE	12.4	92.03	77.45
	Ch22.2		SRR364391	SE	12.4	92.18	77.4
	Ch22.3		SRR364392	SE	12.2	92.08	87.29
	Ch40	40	SRR364400	SE	11.9	91.40	82.08
	Ch40.1		SRR364401	SE	11.9	91.40	82.25
	Ch40.2		SRR364398	SE	13.2	92.55	80.25
	Ch40.3		SRR364399	SE	13.2	92.73	80.25
TT: *A. thaliana* with *B. cinerea* PRJNA315516 Submitted by the Utrecht University (2016) PRJNA593073 Submitted by the Beijing University (2019)	Bc12	12	SRR3383696	SE	12.1	100.00	98.2
	Bc12.1		SRR3383697	SE	15	100.00	98.14
	Bc18	18	SRR3383779	SE	10.3	97.37	93.55
	Bc18.1		SRR3383780	SE	13.6	97.26	95.2
	Bc24	24	SRR10586397	PE	22.2	95.53	89.07
	Bc24.1		SRR10586399	PE	22	95.70	88.69
TT: *A. thaliana* with *S. sclerotiorum* PRJNA418121 (Badet et al., 2017)	Ss30	30	SRR6283146	SE	20.8	95.01	36.48
	Ss30.1		SRR6283147	SE	20.9	96.09	32.35
	Ss30.2		SRR6283148	SE	21.2	92.90	29.5
CT: *A thaliana* Healthy PRJNA315516 y PRJNA418121	healthy12	12	SRR3383640	SE	10.9	97.12	98.31
	healthy12.1		SRR3383641	SE	22.6	97.50	98.34
	healthy18	18	SRR3383782	SE	29.8	97.62	98.13
	healthy18.1		SRR3383783	SE	14.2	97.69	98.34
	healthy24	24	SRR3383821	SE	15	97.18	98.09
	healthy24.1		SRR3383822	SE	10.2	95.98	97.87
	healthy30	30	SRR6283144	SE	22.1	95.31	97.51
	healthy30.1		SRR6283145	SE	19.7	95.72	97.99

(*)hpi: hours post infection; (M): Millions of reads. (**Suplementario.2.1**)

2.3.2 Integración y ajuste de los conteos en matrices de expresión

2.3.2.1 Paso 1: integración de conteos de expresión RNA-Seq

Los 25 archivos con los conteos sin normalizar se integraron en matices de expresión mediante el identificador único del gen. La matriz de control se compone de 221 mil valores de expresión (27,655 CDS x 8 muestras), y la de los tratamientos de 470 mil (27,655 CDS x 17 muestras). Se eliminaron un total de 2056 (~7.43 %) genes con ceros en la matriz de control, y 243 (~0.87 %) en la matriz con los tratamientos. Se identificaron 3172 (~ 11.46 %) genes sin conteos en común entre ambas matrices (**Figura 2.3**). Se retuvieron 22,426 genes en la matriz de control con una μ=785.5±391.5, σ=4983±2925 y R=300,000, y 24,239 genes en la matriz de tratamientos con una μ=429.5±232.5, σ=2041 ±829 y R=250,000 (**Suplementario 2.2)**.

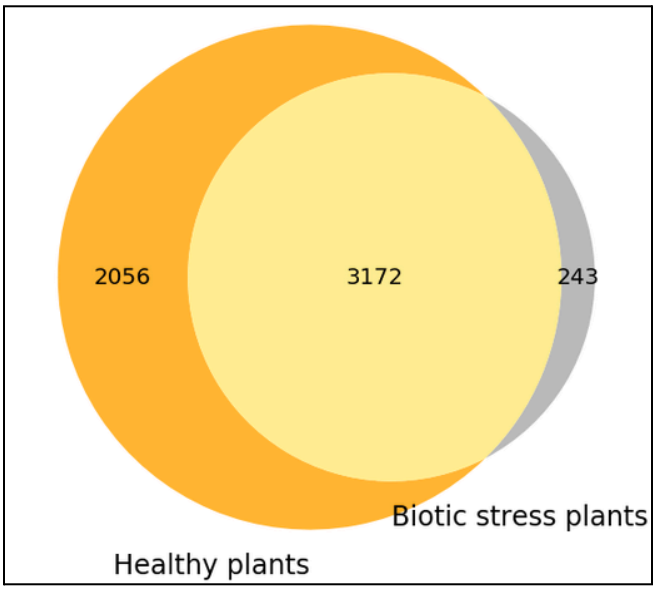


Figura 2.3. Diagramas de Venn para genes no expresados. Los valores de expresión iguales a cero se eliminaron en ambas matrices de expresión.

2.3.2.2 Paso 2,3: control de la variabilidad (normalización y estandarización)

Normalizados los valores de las matrices de expresión a TPM (transcritos por millón) con el objetivo de hacer los valores comparables (**Figura 2.1B**), las métricas obtenidas para la matriz de control son, μ=44.6, σ=333±6 y R=36543, y para la matriz de tratamientos, μ=41.2, σ=284.5±84,5 y R=30046 (**Figura 2.4A**). Para reducir la variabilidad en las matrices se definieron dos umbrales de corte basados en percentiles para filtrar los genes

subrepresentados y/o con valores de expresión exponenciales. Se filtraron los genes con valores por debajo del 1er. percentil en al menos el 70 % de las muestras, y los genes con valores por encima del 99vo. percentil en general. En la matriz control se eliminaron 2634 genes, 2262 de estos corresponden al límite inferior, es decir, están por debajo del 1er. percentil (TPM ≤ 0.1), y 372 genes están en el límite superior, es decir, por encima del 99vo. percentil (TPM ≥ 830). En la matriz de tratamientos se eliminaron 3965 genes, 3495 por debajo del límite inferior (TPM ≤ 0.1) y 470 por encima del límite superior (TPM ≥ 845). El número de genes retenidos total es de 19792 y 20274 para la matriz de control y tratamientos, según corresponde. En la matriz de control se tiene una μ=27.2±2.7, σ=50±5 y R=840, y en la matriz de tratamientos una μ=27.21±6.2, σ=62.5±4.5 y R= 845 (**Figura 2.4B**).

2.3.2.3 Paso 4: identificación de muestras atípicas

Para reducir la escala de las matrices y resaltar las distribuciones se transformaron los valores a escala $\log_2(TPM+1)$. Se obtuvo en la matriz de control una μ=3.4±0.2, σ=2.1±0.1 y R=9.28, y en la matriz de tratamientos una μ=2.9±0.8, σ=2.1±0.1 y R=9.72 (**Figura 2.4C**). En los tratamientos los histogramas con ajuste de densidad KDE resaltan las distribuciones mezcladas (**Figura 2.4C**), mostrando la forma preponderante de las distribuciones, y resaltando las muestras con valores atípicos (Ss30). Los gráficos de violín con ajuste KDE se utilizaron para observar las distribuciones individualmente y dar soporte a los resultados (**Figura 2.5**); puede observarse que después de filtrar las muestras con distribuciones atípicas (control negativos, muestras Ss30), se alcanza una μ=3.55±0.16, σ=2.075±0.075, y R=9.72, eliminado alta variabilidad causada por muchos valores de expresión contenidos en las muestras Ss30 de forma global (**Suplementario 2.1**).

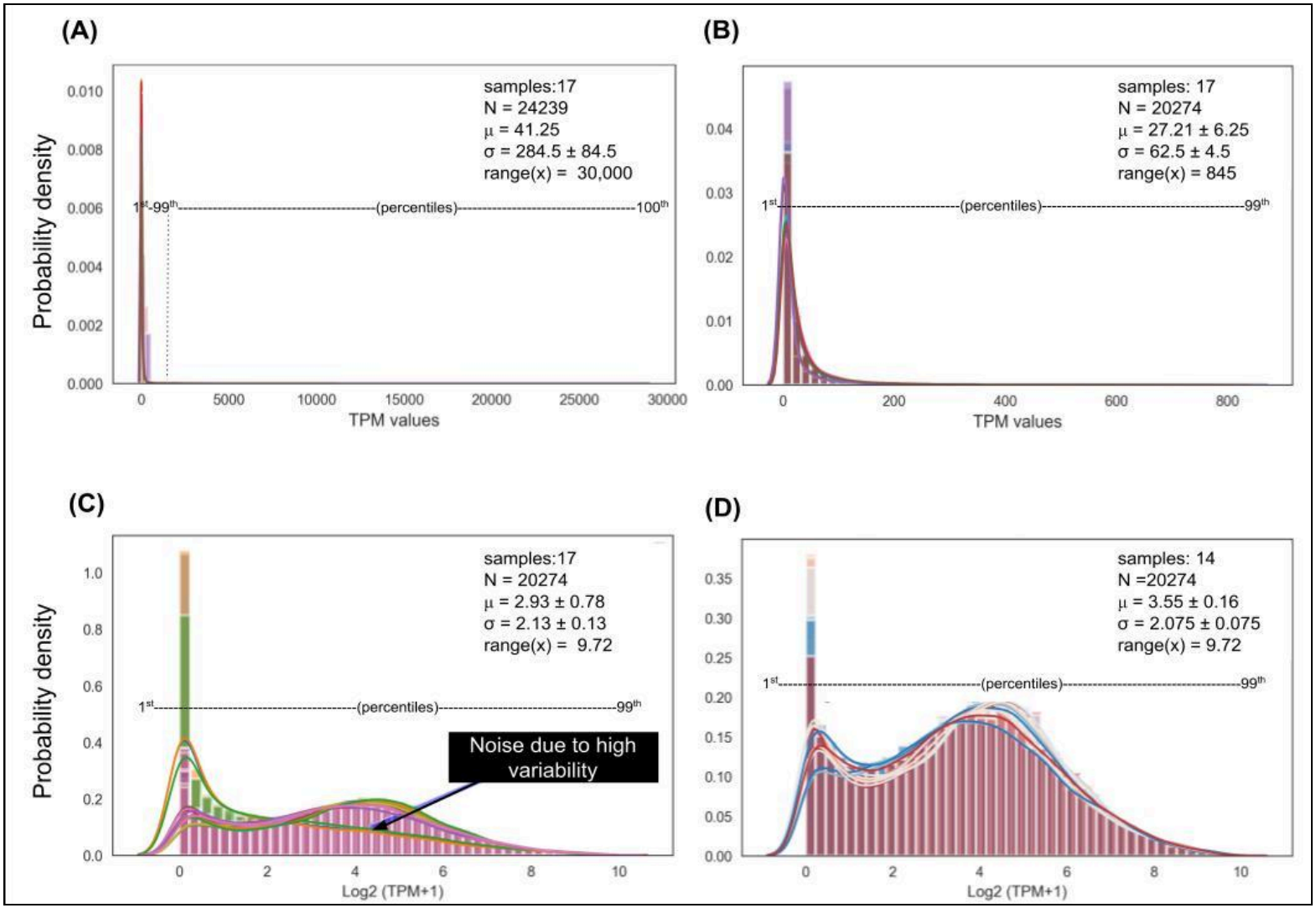


Figura 2.4. Integración de conteos en la matriz de tratamientos. (A) Distribución de la matriz en TPMs, tiene μ=41.25, σ=284.5±84.5 y R=30000. (B) Identificación de percentiles para definir umbrales de corte por debajo del 1er perc. y por encima del 99vo, tiene μ=27.21±6,25, σ=62.5±4.5 y R=845. (C) Distribución de la matriz filtrada en escala log2(TMP+1) para resaltar distribuciones atípicas (control negativos, muestras Ss30), tiene μ=2.13±0.13, σ=2.93±0.78, y R=9.72 (D) Distribución de la matriz sin las muestras atípicas (Ss30), tiene μ=3.55±0.16, σ=2.075±0.075, y R=9.72

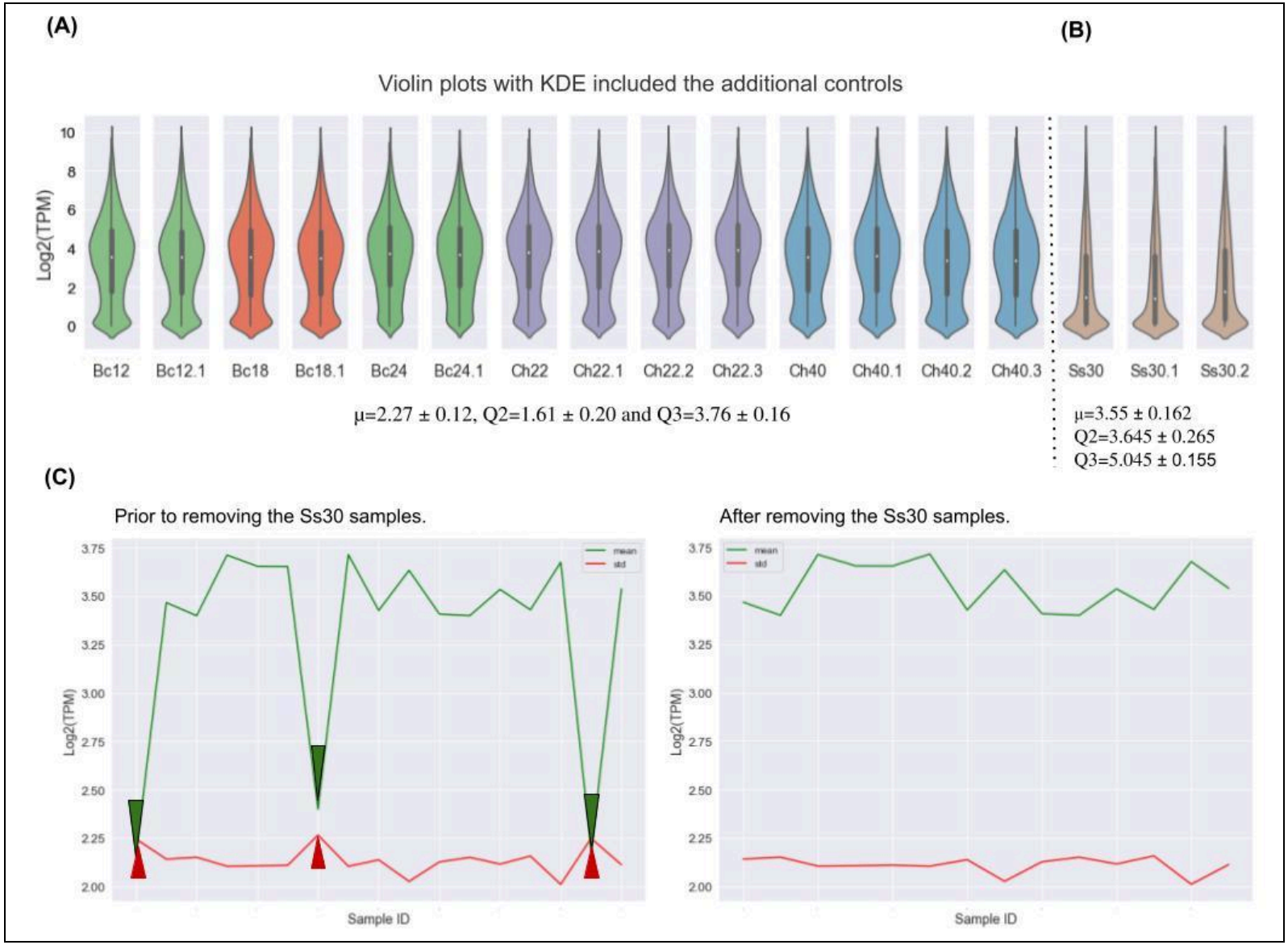


Figura 2.5. Distribución de las muestras en la red de tratamientos**.** (A) Distribución de las muestras sin incluir las muestras Ss30, (B) Distribución de las muestras Ss30, (C) Media (línea verde) y desviación estándar (línea roja) antes y después de eliminar las muestras Ss30.

2.3.3 Redes de coexpresión ponderadas WGCNA

Se construyeron dos redes ponderadas con signo por el método de *Pearson* con la herramienta WGCNA, una para la matriz de control y otra para los tratamientos (**Suplementario 2.2**). En la red con los controles se alcanzó un coeficiente r^2=0.80 y se identificaron 237 agrupamientos con una conectividad media de nodo (NCM) igual a 374 (β=28; r^2=0.78). En la red fusionada a 0.1 distancias se identificaron 23 agrupamientos (**Figura 2.6A**; **Tabla 2.3**). En la red con los tratamientos se alcanzó un coeficiente r^2=0.83, se identificaron 100 agrupamientos con una NCM igual a 270 genes por agrupamiento (β=27; r^2=0.79). En la red fusionada a 0.1 distancias se identificaron 36 agrupamientos (**Figura 2.6B**; **Tabla 2.3**).

Tabla 2.3. Índices de la red TOM y la media de conectividad de nodo (NCM).

Red	r^2	# Clústeres (std ntw)	NCM	# Clústers (Red fusionada 0.1)
Red de control	β=28; 0.78	237	374	23
Red de tratamientos sin incluir las muestras Ss30	β=27; 0.79	100	270	36
Red de tratamientos con las muestras Ss30	β=27; 0.79	276	***1041**	100

Las muestras de control negativo (Ss30) en la red de tratamientos alcanzó un coeficiente de correlación (CC) r^2=0.82, identificando 276 agrupamientos (β=27; r^2=0,79), con una NCM=1041, 4 veces más alta que la NCM alcanzada una vez eliminadas las muestras de control negativo (NCM=270), siendo muy similar a la NCM de los controles (**Tabla 2.3**; **Figura 2.6B**). Derivado de esto, se descartaron las muestras Ss30 de la red de tratamientos. En ambas redes se identificaron módulos con correlaciones positivas y negativas (**Tabla 2.4**).

2.3.4 Agrupamientos y anotación funcional

Los módulos génicos con coeficiente r^2>0.75 y valor-p<0.05 fueron considerados en este análisis (**Tabla 2.4**) (**Suplementario 2.2**). Se identificaron 3 módulos (*Coral3, Navajowhite3 y Blue3*) en la red de control y 4 en la red de tratamientos (*Chocolate, Dodgerblue1, Chocolate2 y Green3*). El número de genes en cada módulo fue entre 79 y 2024 en el control, de los cuales entre de 34-96% se clasificaron en clases GO; en la red de tratamientos los módulos oscilaron entre 72 y 818 genes, con 25-30% genes clasificados.

Las pruebas de sobrerrepresentación GO realizadas con PANTHER v17.0 sobre los módulos de la red de control dieron como resultado en *Coral3,* 16 clases en función molecular, siendo "*heterocyclic compound binding"* (GO:1901363) y "*organic cyclic compound binding"* (GO:0097159) las más representadas con genes relacionados con proteínas del metabolismo del RNA*,* factor de iniciación de la traducción, unión de cromatina, proteína G heterotrimérica, factor de transcripción de caja MADS y modulador de proteína G. En *Navajowhite3* (plantas sanas 12 hpi) se obtuvieron 2 clases, "*vesicle-mediated transport"* (GO:0016192) con genes involucrados con la capa vesicular, *y "membrane traffic regulatory protein" and guanyl-nucleotide exchange factor"*; en *Blue3* en la categoría de proceso biológico (plantas sanas 18 hpi) se obtuvieron 6 clases, siendo "*response to light stimulus"* (GO:0009416) y *"response to abiotic stimulus"* (GO:0009628) las más representadas con genes involucrados con factores de transcripción de unión a DNA, proteínas de unión a cromatina y actina no motora (**Figura 2.7A,B**) (**Suplementario 2.3**).

Tabla 2.4. Módulos de red identificados en la red de control y tratamientos.

Red	Nombre del módulo	Tamaño	Coef. r²
Módulos de la red de control	coral3 *	2024	+0.98
	blue3 *	79	+0.96
	navajowhite3 *	492	+0.79
	magenta2	204	-0.98
	darkolivegreen4	1792	-0.97
	antiquewhite	764	-0.88
Módulos de la red de plantas infectadas	chocolate *	818	+0.98
	dodgerblue1 *	72	+0.96
	green3 *	472	+0.95
	chocolate2 *	573	+0.91
	tomato2	218	-0.94
	palevioletred1	368	-0.93
	green	236	-0.90
	deepskyblue	39	-0.80

***** Se seleccionaron módulos genéticos con coeficiente r² positivo para realizar la prueba de sobrerrepresentación de Gene Ontology.

Las pruebas de sobrerrepresentación de los módulos de la red de tratamientos mostró 19 clases sobrerrepresentadas en la categoría función molecular para el módulo *Chocolate*, siendo nuevamente "*protein binding"* (GO:0005515) la más representada con genes relacionados con proteína motora de unión a microtúbulos, GTPasa pequeña, activador de cinasa, proteína SNARE, proteína ligasa de ubiquitina, proteína fosfatasa, proteína de tráfico de

membrana, proteína cinasa de serina/treonina no receptora, proteína del metabolismo del RNA, entre otras. En las clases "*protein kinase activity"* (GO:0004672), "*phosphotransferase activity"* (GO:0016773) y "*protein serine/threonine kinase activity"* (GO:0004674) --subclases del "*cyclin-dependent protein serine/threonine kinase activity"* (GO:0004693), se encontraron genes relacionados con PK no receptor de serina/treonina y receptores de señales transmembrana. En el módulo *Chocolate2* (*A thaliana/C higginsianum* 22 hpi) se tiene a la clase "*RNA binding"* (GO:0003723) con genes relacionados con factor de empalme de RNA, cofactor de transcripción, iniciación de la traducción, entre otros, además de numerosas proteínas que contienen repeticiones de penta-tricopéptidos no clasificados involucrados en modificaciones del RNA y respuesta celular a la hipoxia. En el módulo *Green3* se encontró la clase "*unfolded protein binding"* (GO:0051082) conteniendo proteínas de choque térmico *(heat shock proteins / Hsp)* clase I y II (chaperonas), como los genes *Hsp70 and Hsp90* que se activan ante diversos tipos de estrés, como por ejemplo la respuesta al calor, al peróxido de hidrógeno, al estrés salino y la hipoxia.

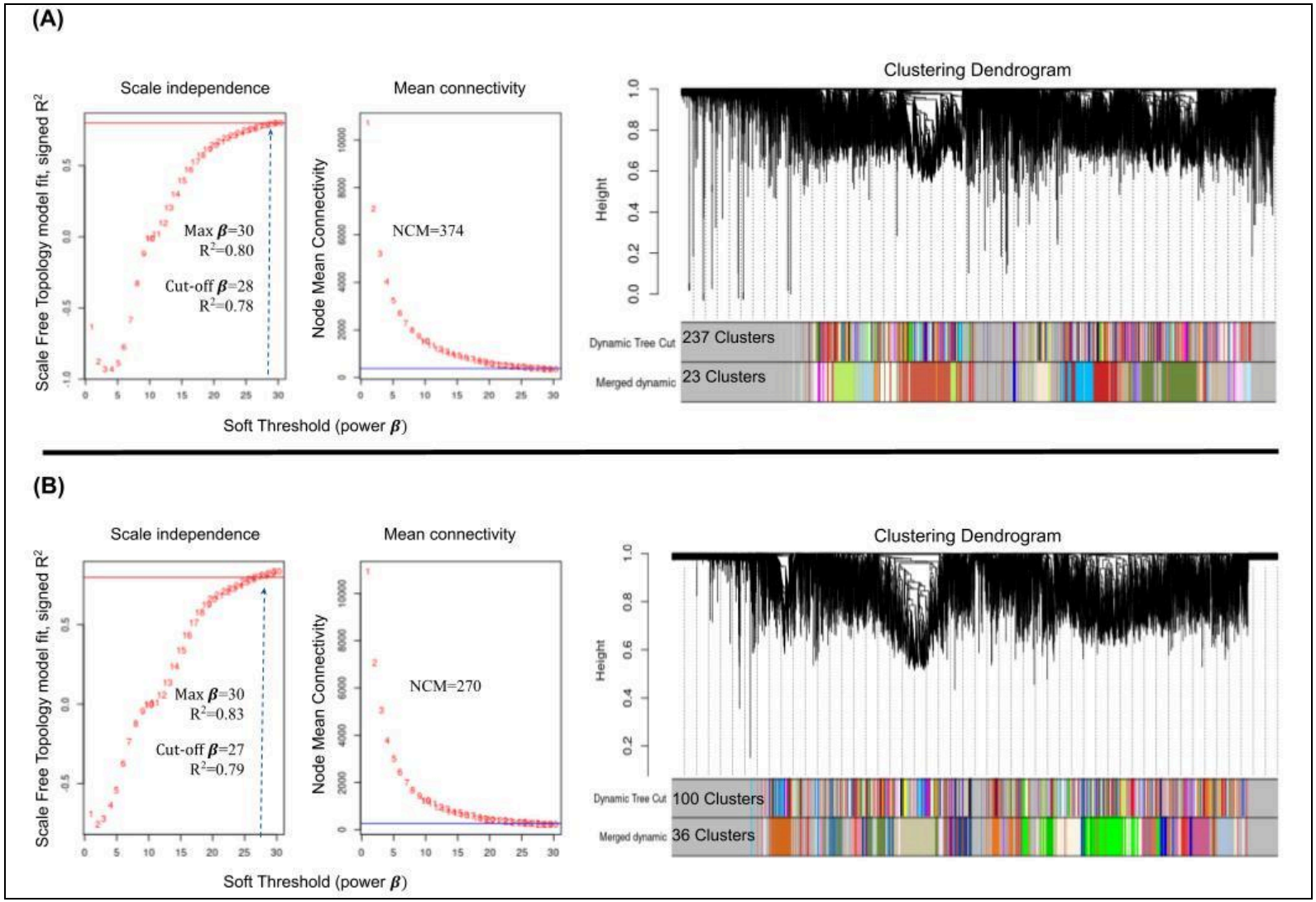


Figura 2.6. Redes de coexpresión y agrupamientos. Se construyeron dos redes de coexpresión ponderadas con signo con el método de *Pearson*. (A) Modelo SFT para la red con los controles; alcanzó un coeficiente r^2=0.80. En el umbral β = 28 se alcanzó un coeficiente r^2=0.78 y NMC=374 genes. La red fusionada a 0.1 distancias generó 23 agrupamientos. (B) En la red con los tratamientos se alcanzó en el umbral β = 27 un coeficiente r^2=0.79 y NCM de 270 genes. La red fusionada generó 36 agrupamientos.

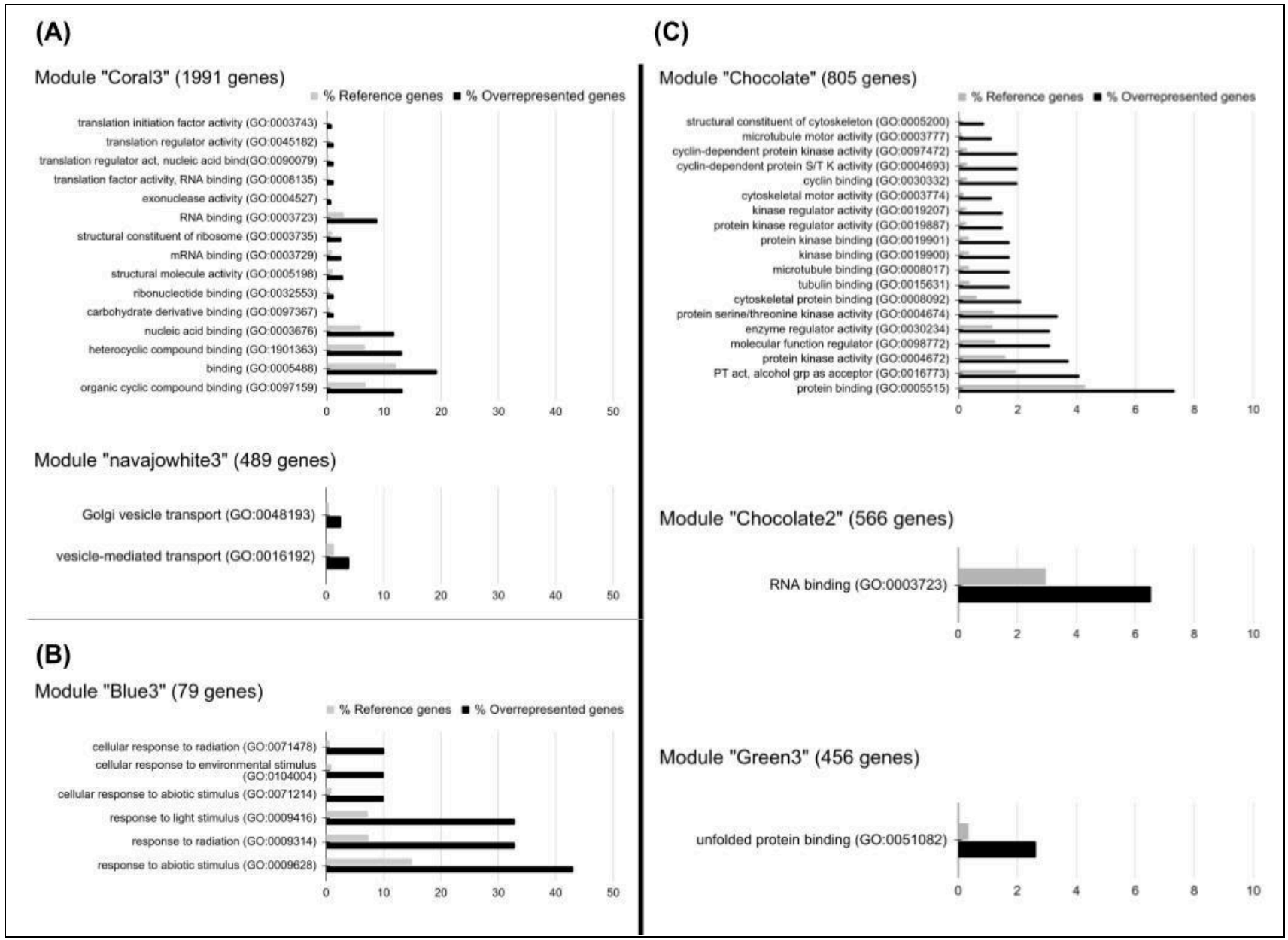

(A)
Module "Coral3" (1991 genes)
% Reference genes
% Overrepresented genes
translation initiation factor activity (GO:0003743)
translation regulator activity (GO:0045182)
translation regulator act, nucleic acid bind(GO:0090079)
translation factor activity, RNA binding (GO:0008135)
exonuclease activity (GO:0004527)
RNA binding (GO:0003723)
structural constituent of ribosome (GO:0003735)
mRNA binding (GO:0003729)
structural molecule activity (GO:0005198)
ribonucleotide binding (GO:0032553)
carbohydrate derivative binding (GO:0097367)
nucleic acid binding (GO:0003676)
heterocyclic compound binding (GO:1901363)
binding (GO:0005488)
organic cyclic compound binding (GO:0097159)
0 10 20 30 40 50
Module "navajowhite3" (489 genes)
Golgi vesicle transport (GO:0048193)
vesicle-mediated transport (GO:0016192)
0 10 20 30 40 50
(B)
Module "Blue3" (79 genes)
% Reference genes
% Overrepresented genes
cellular response to radiation (GO:0071478)
cellular response to environmental stimulus (GO:0104004)
cellular response to abiotic stimulus (GO:0071214)
response to light stimulus (GO:0009416)
response to radiation (GO:0009314)
response to abiotic stimulus (GO:0009628)
0 10 20 30 40 50
(C)
Module "Chocolate" (805 genes)
% Reference genes
% Overrepresented genes
structural constituent of cytoskeleton (GO:0005200)
microtubule motor activity (GO:0003777)
cyclin-dependent protein kinase activity (GO:0097472)
cyclin-dependent protein S/T K activity (GO:0004693)
cyclin binding (GO:0030332)
cytoskeletal motor activity (GO:0003774)
kinase regulator activity (GO:0019207)
protein kinase regulator activity (GO:0019887)
protein kinase binding (GO:0019901)
kinase binding (GO:0019900)
microtubule binding (GO:0008017)
tubulin binding (GO:0015631)
cytoskeletal protein binding (GO:0008092)
protein serine/threonine kinase activity (GO:0004674)
enzyme regulator activity (GO:0030234)
molecular function regulator (GO:0098772)
protein kinase activity (GO:0004672)
PT act, alcohol grp as acceptor (GO:0016773)
protein binding (GO:0005515)
0 2 4 6 8 10
Module "Chocolate2" (566 genes)
RNA binding (GO:0003723)
0 2 4 6 8 10
Module "Green3" (456 genes)
unfolded protein binding (GO:0051082)
0 2 4 6 8 10

Figura 2.7. Prueba de sobrerrepresentación de Gene Ontology (GO). (A) Clases GO-slim de la red de control para función molecular (MF) en el módulo Coral3, y proceso biológico (BP) en el módulo Navajowhite3. (B) Clases GO-slim de la red de control para BP en el módulo Blue3. (C) Clases GO-slim de la red de plantas infectadas para MF en los módulos Chocolate, Chocolate2 y Green3.

2.4. DISCUSIÓN

RNA-Seq ha transformado la investigación en especies modelo (Wang et al., 2009) y ha supuesto un punto de inflexión para incluso estudiar organismos que carecen de un genoma de referencia (Góngora-Castillo y Buell, 2013). En el ámbito de la transcriptómica de plantas hoy en día contamos con plataformas en línea como ATTEND-II (Obayashi et al., 2022), Expression Angler (Austin et al., 2016), AtCAST (Kakei y Shimada, 2015), Aranet (Lee et al., 2015) y PlaD (Qi et al., 2018), entre otras, que integran BDs y organizan la información de múltiples conjuntos de datos para análisis exploratorio, sin embargo, el contenido en datos suele ser limitado a ciertas especies e interacciones moleculares, además, están sujetas a métodos de análisis específicos. En general se espera que una integración de datos *personalizada* ofrezca mayor flexibilidad que cualquier plataforma, sin embargo, la integración enfrenta muchos retos debido a las diferencias en el diseño experimental, los métodos de secuenciación y la variabilidad en los datos (Xu y Tian, 2015), además, el análisis de datos públicos de plantas enfrenta otras limitaciones, como el amplio número de especies e interacciones moleculares, tanto a nivel fenotípico como genotípico (Fondi y Liò, 2015; Hughes, 2015), por lo que es común encontrar estrategias muy específicas (Ibrahim et al., 2021; Thomas et al., 2019; L. Zhang et al., 2018; Shaik y Ramakrishna, 2013) o muy generales (Corchete et al., 2020; Conesa et al., 2016).

Uno de los principales en los análisis integrales se encuentra en las diferencias entre el diseño experimental y los métodos de secuenciación, conocido como *efecto batch*; para ayudar a mitigar tales efectos se han desarrollado varias herramientas, como ComBat (Y. Zhang et al., 2020), uno de los métodos de ajuste de efectos por lotes más populares cuando los efectos provienen de fuentes conocidas, o SVASeq (Leek, 2014) y RUVSeq (Risso et al., 2014) para la heterogeneidad de fuentes desconocidas, todos mayormente diseñados para análisis de expresión diferencial (DGE); sin embargo su rendimiento en la corrección de datos para evitar la aparición de asociaciones espurias durante la agrupación jerárquica no se comprende bien (Vandenbon, 2022), por lo que la corrección de estos efectos siguen siendo un área de investigación abierta. En este estudio proponemos un marco metodológico alternativo, en parte porque la mayoría de los enfoques estudiados son específicos para DGE, y en esta tesis

buscamos una alternativa para la búsqueda de patrones génicos de consenso a partir de redes de coexpresión.

Nuestra metodología se implementó utilizando cuatro estudios independientes de *A. thaliana* (PRJNA148307, PRJNA315516, PRJNA593073, PRJNA418121) procedentes de diferentes grupos de investigación; se seleccionaron 25 conjuntos de datos de RNA-Seq de transcriptomas de hoja bajo estrés biótico, y de planta sana para control. El marco metodológico consta de 4 pasos (**Figura 2.1**) que conceptualizan pasos específicos para reducir la variabilidad entre conteos génicos de datos RNA-Seq. Con la finalidad de facilitar la adopción de la estrategia aplicamos métricas estadísticas convencionales e histogramas con ajuste de densidad KDE (Waskom, 2021) (distribuciones) para dar seguimiento a los ajustes en los datos. La matriz de expresión resultante es evaluada mediante los índices de red del modelo SFT de la red TOM contenido en WGCNA. Se comienza con la integración de los conteos de expresión en matrices de datos, seguido por la normalización a TPM (transcritos por million); la selección del método de normalización es flexible y puede optarse por un método más conveniente; la normalización TPM utilizada, trata los conteos de forma independiente, proponiendo una representación de nivel de expresión relativo, en principio equivalente; sin embargo, existe una amplia discusión sobre este y otros método *perse*, ¿Cuál es mejor método para qué?, la respuesta no es simple, ya que se ha llegado a diferentes conclusiones en diversas pruebas comparativas entre diversos tejidos celulares, números de muestras y métodos; por ejemplo, Wagner et al. (2012) y colaboradores discutieron los beneficios de TPM sobre FPKM y abogaron por el uso de TPM basado en pequeños conjuntos de datos (seis muestras) de tejidos humanos con solo dos réplicas; años después Abrams et al. (2019) y colaboradores concluyeron que TPM era el método de normalización con mejor rendimiento porque conservaba la variabilidad biológica sin introducir muchos sesgos adicionales en conjuntos de datos de líneas celulares de cáncer y muestras de cerebro humano; su conclusión se basó en el análisis de réplicas técnicas (es decir, las mismas muestras secuenciadas en diferentes laboratorios); Zhao et al. (2021) y colaboradores concluyeron que en datos RNA-Seq generados a partir de tumores cultivados en modelos PDX (patient-derived xenograft), los conteos normalizados DESeq2 y TMM (trimmed mean of M-values) funcionaron mejor que las normalizaciones TPM o FPKM/RPKM (fragments per kilobase of transcript per million fragments mapped). En general, dado que diferentes tejidos expresan diversos repertorios de RNA, los valores TPM entre tejidos no deben considerarse directamente comparables (S. Zhao et al., 2020), tampoco debe usarse TPM para comparaciones cuantitativas entre muestras cuando los

contenidos de RNA total y sus distribuciones son muy diferentes, aunque bajo ciertas circunstancias TPM puede ser útil para la comparación cualitativa, como PCA y análisis de conglomerados (S. Zhao et al., 2020). Por lo tanto, se debe de partir del hecho que cada método de normalización viene con un conjunto de suposiciones, por lo tanto, la validez de los resultados del análisis depende de sí la configuración experimental es congruente con estas suposiciones (Evans et al., 2018); en el clustering debe trabajarse la sensibilidad del método a las distribuciones heterogéneas, es decir, con características expresadas de manera alta y diferencial, que pueden afectar el análisis; por lo tanto la elección del método de dependerá de las características de los datos.

Partiendo de los supuestos de qué los datos provienen de protocolos y métodos de secuenciación similares (RNA-Seq polyA secuenciadas por Illumina), de un solo organismo y tipo de tejido celular (hoja), y qué los conteos producidos siguieron la misma estrategia; una forma práctica de identificar y eliminar variabilidad parte del prefiltrado de genes basados en sus percentiles. En el paso 3, partiendo del concepto general de qué el área debajo de las colas en una distribución representa la región de rechazo (Pukelsheim, 1994), definimos 2 umbrales de corte, el *primero* para filtrar genes que tienen un rango de variabilidad muy pequeño y tienden a agruparse en un solo grupo, genes con valores TMP por debajo del 1er percentil, que además están subrepresentados y no se expresan en >70% de las muestras; en el *segundo* umbral filtramos el grupo de genes con valores exponenciales que generan una pronunciada cola en la distribución, genes con valores TPM>99vo percentil; para este umbral al no poderse anticipar la variabilidad en la matriz de expresión, no se aplicó una segunda condición para su cumplimiento. La definición de los umbrales no es estricta, por lo que se pueden definir umbrales más suaves o más estrictos según sea su caso. Bajo esta definición retuvimos >80% de los genes expresados, logrando en la matriz de control pasar de una μ=44.6, σ=333 ± 6 y R=36543 a una μ=27.2 ± 2.7, σ=50 ± 5 y R=840; y en la matriz de tratamientos pasar de μ=41.2, σ=284.5 ± 84.5 y R=30046 a una μ=27.21 ± 6.2, σ=62.5 ± 4.5 y R=845 (**Figura 2.4A**,**B**). En el último paso, utilizamos los conteos del control negativo de las muestras con baja cobertura Ss30 (<35%) de alineamiento para resaltar la forma preponderante de las distribuciones y contrastar las muestras con alto contenido de valores atípicos (Ss30); transformando los valores a $\log_2$(TPM+1), medimos el coeficiente de correlación r^2 y la NCM (media de conectividad de nodo) de las redes antes y después de quitar el control negativo (muestras Ss30). El coeficiente de correlación r^2 no varió mucho entre las redes antes y después de quitar el control negativo, sin embargo la NCM fue significativamente

diferente, reportando una NCM 4 veces más alta (NCM=1041) en la red con las muestras de control negativo (Ss30), en contraste con la NCM=270 en la red sin este control, muy similar al NCM=374 obtenida para la red de control (**Figura 2.6B; Tabla 2.3**). El control negativo permite identificar alta variabilidad general que afectan el clustering, y la NCM puede capturar la variabilidad derivada de la cobertura, a diferencia del coeficiente de correlación r^2

En ambas redes se identificaron módulos génicos con moderadas a altas positivas y negativas; tomando para validación los módulos con coeficiente r² positivo>0.80 (alto), las pruebas de sobrerrepresentación GO con PANTHER (Thomas et al., 2022; Mi et al., 2019) en la red de control (**Figura 2.7A,B**) revelaron funciones moleculares y/o procesos biológicos relacionados con el transporte mediado por vesículas y la respuesta a la luz, el mantenimiento y desarrollo de la planta, en contraste, los módulos de la red con plantas infectadas revelaron la presencia de genes con actividades quinasa, tráfico de membrana, proteínas quinasas no receptor de serina/treonina, receptores de señales transmembrana y proteínas de choque térmico (**Figura 2.7C**), estas últimas estrechamente relacionadas con la respuesta al estrés oxidativo, chaperonina involucrada en el replegamiento de proteínas, factores de transcripción para regular la expresión génica y la metilación de histonas. Los resultados finales obtenidos, son congruentes con los fenotipos incluidos en cada red. Para finalizar, el marco metodológico permitió identificar un módulo génico *consenso* (objetivo de esta investigación) para la interacción de *A thaliana* con *C higginsianum* (Ch22) y *A thaliana* con *B cinerea* (Bc24), en un rango de las 22 a 24 horas de infección, respectivamente; aunque los hongos involucrados poseen estilos de vida diferentes y huéspedes diferentes, se logró captar funciones y procesos biológicos convergentes dentro de la amplia colección de estrategias de respuesta de la maquinaria de defensa que posee la planta, lo cual se discute en el siguiente capítulo.

2.5 CONCLUSIONES

Pensamos que nuestra propuesta metodológica es una alternativa práctica y flexible qué contribuye a la demanda de estrategias y herramientas para la integración de datos públicos, particularmente en datos de plantas, además esta diseñada para la identificación de patrones consenso mediante redes de coexpresión génica, a diferencia de las propuestas para DGEs (Y. Zhang et al., 2020; Leek, 2014; Risso et al., 2014); las etapas permiten conceptualizar la relevancia de la variabilidad en los datos, y logran identificar módulos asociados biológicamente en conjuntos de datos independientes; para su implementación deben revisarse los supuestos sobre los cuales fue diseñada.

CAPÍTULO III.

CROSSTALK DEL MECANISMO DE RESPUESTA AL ESTRÉS EN ARABIDOPSIS INFECTADA POR *B CINEREA* Y *C HIGGINSIANUM*

3.1. INTRODUCCIÓN

En este capítulo se describen los resultados de las redes de coexpresión construidas para la identificación de módulos génicos *consenso*, el análisis de enriquecimiento y análisis complementarios para describir los hallazgos encontrados en las principales vías de respuesta activas encontradas entre *A thaliana* con *B cinerea* (Bc_24) y *A thaliana con C higginsianum* (Ch_22) a las 24 y 22 hpi, respectivamente.

3.2. MATERIALES Y MÉTODOS

3.2.1 Redes de coexpresión e identificación de módulos consenso

Los métodos utilizados para crear los perfiles de expresión e integrar los conteos génicos de los transcriptomas de arabidopsis en matrices de expresión se encuentran descritos en el Capítulo II.

Con las matrices de expresión se construyeron dos redes de coexpresión ponderadas con signo (signed-ntw) utilizando el método de *Pearson,* una red para plantas sanas y otra para infectadas. Se utilizó la herramienta WGCNA v1.69-81 (Langfelder y Horvath, 2008) para R (R: The R Project for Statistical Computing, s. f.), siguiendo las pautas señaladas en el inciso 2.2.3 CONSTRUCCIÓN DE REDES DE COEXPRESIÓN PONDERADAS. Se identificaron en ambas redes módulos génicos con coeficientes de correlación (CC) $r^2>0.75$ o CC $r^2<0.75$, y valor-p<0.05; en la red de plantas infectadas se consideraron los módulos con CC $r^2>0.50$. Mediante comparaciones lógicas se identificaron los módulos génicos de las plantas infectadas diferenciados en >75% de los módulos de la red de control. El código utilizado se encuentra indicado en la **Tabla 2.1** del capítulo anterior.

3.2.2 Clusterización funcional de enriquecimiento con DAVID

Los módulos génicos con coeficientes de correlación r^2 positivo moderado-alto y altamente diferenciados del control (>57%), se anotaron con *clusterización funcional de*

enriquecimiento en la plataforma DAVID (**D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery) (Huang et al., 2007) v6.8 (DAVID 2021 update; DAVID Knowledgebase v2021q4). Se corrió prueba de enriquecimiento EASE con corrección *Bonferroni*; el EASE Score es un *valor-p* exacto de *Fisher* modificado para el análisis de enriquecimiento, que en DAVID es más conservador. El EASE Score va de 0 a 1, y el valor predeterminado es 0.1. *Fisher* exact p-value=0 representa de enriquecimiento perfecto. Bajo esta definición el valor-p =<0.05 es considerado fuertemente enriquecido en las categorías de anotación. Se utilizó la referencia genómica TAIR10 y la anotación Araport11 (Cheng et al., 2017); se corrieron 10 pruebas con diferentes niveles de estridencia estadística contra 9 BDs en función de los porcentajes de cobertura obtenidos.

3.2.3 Análisis de polimorfismos de secuencia y localización cromosomal

Con los genes de los clústeres que resultaron enriquecidos, se realizó localización cromosomal y análisis de polimorfismos utilizando código personalizado y la anotación Araport11 (Cheng et al., 2017). Por *localización cromosomal* nos referimos al cromosoma que contiene al gen, y por *polimorfismo de secuencia*, a las secuencias en las que se reporta cualquier variación génica o epigenética entre dos ecotipos o dentro de un ecotipo que pueda evaluarse por experimentación, secuenciación, amplificación PCR, análisis de restricción o inspección fenotípica. Los *polimorfismos* se encuentran en cualquier sitio (intron, exon, promoter, UTR, etc) del genoma, y en su gran mayoría están en estado desconocido, no se conoce el tipo de alelo o no se dispone de la información.

3.3. RESULTADOS

3.3.1 Módulo identificados y módulo consenso darkmagenta

En la red de plantas infectadas se identificaron 13 módulos génicos, 4 con coeficiente r^2>=0.70 para los módulos *firebrick2, mistyrose3, lavender y chocolate*, y 5 con coeficiente r^2<=0.70 (alto moderado, downregulated) para los módulos *brown2, tan3, lightsteelblue, mediumpurple1 y chocolate2*. También se identificaron 3 módulos antagónicos, *coral4, indianred y green,* estos, regulados positivamente en *A thaliana* con *C higginsianum (*Ch_22*)*, y negativamente en *A thaliana con B cinerea (*Bc_24). Se encontró 1 solo módulo de respuesta consenso denominado "*Darkmagenta"*, con coeficiente de correlación r^2=0.57 para Bc_24 y r^2=0.59 para Ch_22 (**Figura 3.1B**) (**Suplementario 3.1**) .

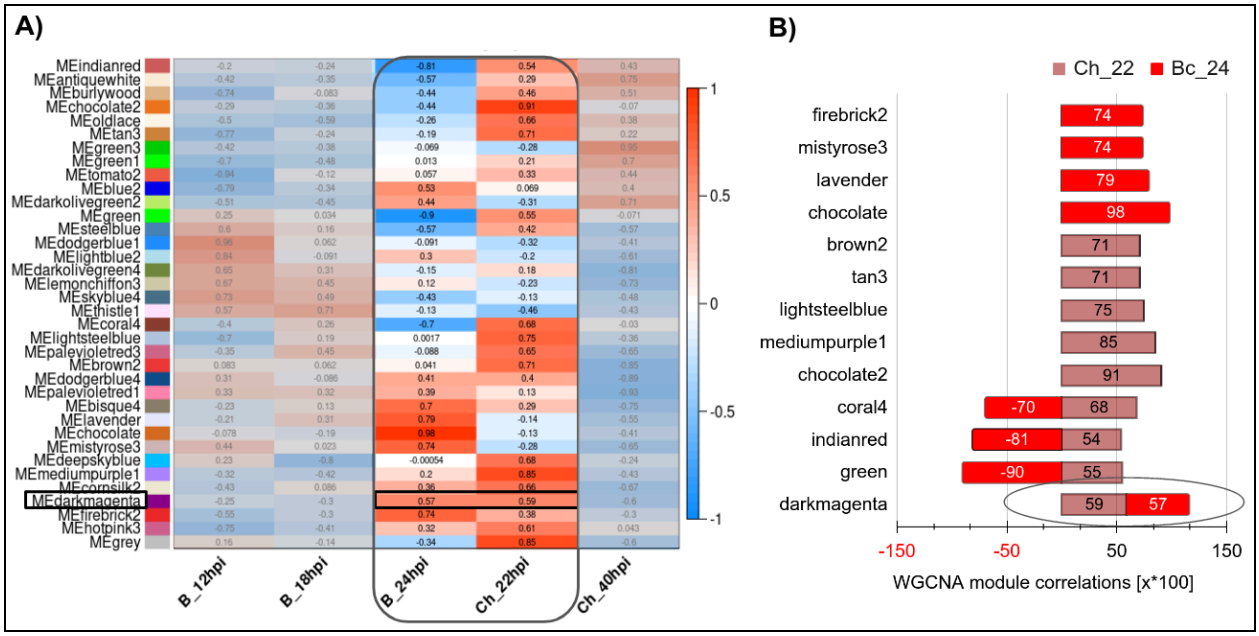


Figura 3.1. Módulo génico de consenso**.** (A) Mapa de calor con los tratamientos incluidos en el experimento. El rectángulo muestra los tratamientos en estudio. B_24hpi engloba las muestras de *A thaliana* con *B cinerea (Bc_24),* y *A thaliana* con *C higginsianum* (Ch_22). (B) Gráfico mostrando los 13 módulos encontrados con c $r^2 > 0.75$ y $r^2 < 0.75$. La elipse señala el módulo consenso encontrado para estos dos patosistemas.

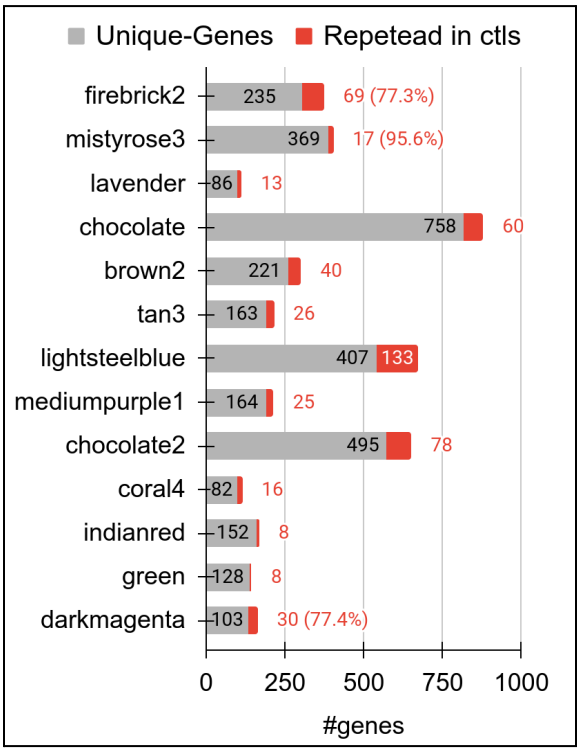


Figura 3.2. Genes únicos de los módulos de la red de tratamientos. Las barras color gris indican la cantidad de genes exclusivos del MG, las barras rojas es la diferencia de genes encontrada en alguno de los MG de la red de control**.**

Todos los módulos de la red están diferenciados en más del 77% de sus genes con respecto a los 23 módulos del control; por ejemplo, el módulo *Darkmagenta* contiene 133 genes, de los cuales 103 (77.4%) están exclusivamente en este agrupamiento (**Figura 3.2**).

3.3.2 Patosistemas del módulo consenso *darkmagenta*

Se identificó un módulo de respuesta *consenso* con coeficiente de correlación $r^2>0.57$, que corresponde a la interacción de *A thaliana* infectada por *B cinerea* (Bc_24) y *A thaliana* infectada *C higginsianum* (Ch_22) (**Figura 3.1A**) a las 24 y 22 hpi, respectivamente; etapa aproximada dentro de las muestras incluidas, al proceso de infección *"in planta appressoria". C higginsianum* es un hongo *ascomycete sordariomycete* hemibiotrofo de múltiples etapas que se establece como biótrofo dentro de las primeras 36 hpi, formando una estructura multilobulada y multiseptada de forma variable, y confinada dentro de las células epidérmicas inicialmente infectadas; en esta etapa, las células infectadas aún pueden plasmolizarse normalmente, y el plasmalema y el tonoplasto del huésped permanen funcionales; tras la colonización posterior de las células vecinas a las 72 hpi se produce un cambio en la morfología de las hifas y en la relación trófica (Damm et al., 2014); se ha reportado que a las 22 hpi (*in planta appressoria)* los apresorios perforan las superficies del hospedador usando fuerza mecánica y degradación enzimática, las hifas biótrofas envueltas por una membrana intacta del hospedador se desarrollan dentro de las células epidérmicas vivas y, finalmente el hongo cambia a necrotrofo, destruyendo los tejidos del huésped. *C higginsianum* está bien equipado con genes que codifican enzimas activas de carbohidratos (CAZymes), que potencialmente degradan la pared celular de la planta y modifican la pared celular fúngica, también codifican una gran cantidad de enzimas degradantes de pectina, aunque la mayoría se activan durante la necrotrofia, ~48 hpi (O'Connell et al., 2012).

B cinerea es un hongo *ascomycete leotomycete,* muy extendido entre las plantas, con un estilo de vida necrotrofo; la infección por *B cinerea* puede prosperar por rutas diferentes que varían según la especie de planta, el tipo de tejido y las condiciones externas; no existe un factor de virulencia único para este tipo de hongos; el desarrollo de la enfermedad tiene múltiples capas y está regulado por múltiples factores, se ha reportado que entre las 4 y 6 hpi las esporas se adhieren a la superficie de la planta y producen un tubo germinativo; entre las 12 y 18 hpi las esporas producen apresorios y cojines de infección que ayudan en la penetración del huésped; junto con la penetración del tejido huésped, las hifas superficiales continúan desarrollándose en la superficie; después del establecimiento de la infección, que

sucede entre las 32 y 48 hpi, las hifas radiantes se diferencian y facilitan la propagación de la lesión (Bi et al., 2022). Tomando como referencia lo anterior, partimos de la asumpción de que algunos genes del *módulo consenso* deben esta asociados a la respuesta de la planta a la etapa de *in planta appressorium,* donde las hifas han perforado la superficie del hospedador para posteriormente desarrollarse dentro de las células epidérmicas vivas, es decir, el proceso de infección esta en una etapa de iniciación avanzada, pero no establecida.

3.3.3 Clusterización funcional de enriquecimiento del módulo consenso

La clusterización funcional de enriquecimiento en el módulo *Darkmagenta* realizada con la herramienta DAVID, se ajustó a 9 categorías (BDs) de anotación en función de la variedad de ómicas y el porcentaje de cobertura alcanzados (**Tabla.3.1**) (**Suplementario 3.2**).

Tabla 3.1. Porcentajes de cobertura génica de las BDs utilizadas para la anotación.

Base de datos (categorías)	Porcentaje	Genes asignados
UP_KW_BIOLOGICAL_PROCESS	41,4	55
UP_KW_CELLULAR_COMPONENT	66,2	88
UP_KW_MOLECULAR_FUNCTION	57,1	76
BD GO Annotation		
GOTERM_BP_DIRECT	79,7	106
GOTERM_CC_DIRECT	98,5	131
GOTERM_MF_DIRECT	83,5	111
BD Pathways		
KEGG_PATHWAY	24,1	32
BD Protein Domains		
INTERPRO	93,2	124
BD Interactions		
BIOGRID_INTERACTION	60,9	81

Con las 9 categorías seleccionadas se realizaron 10 pruebas de anotación con diversos niveles de rigor estadístico para verificar y extraer los agrupamientos más altamente enriquecidos, y mejor preservados en las 10 pruebas; se utilizaron los resultados de las pruebas 0 y 1 con rigor *moderado/alto*, y las pruebas 3 y 6 con rigor *alto/muy alto* (**Figura 3.3**).

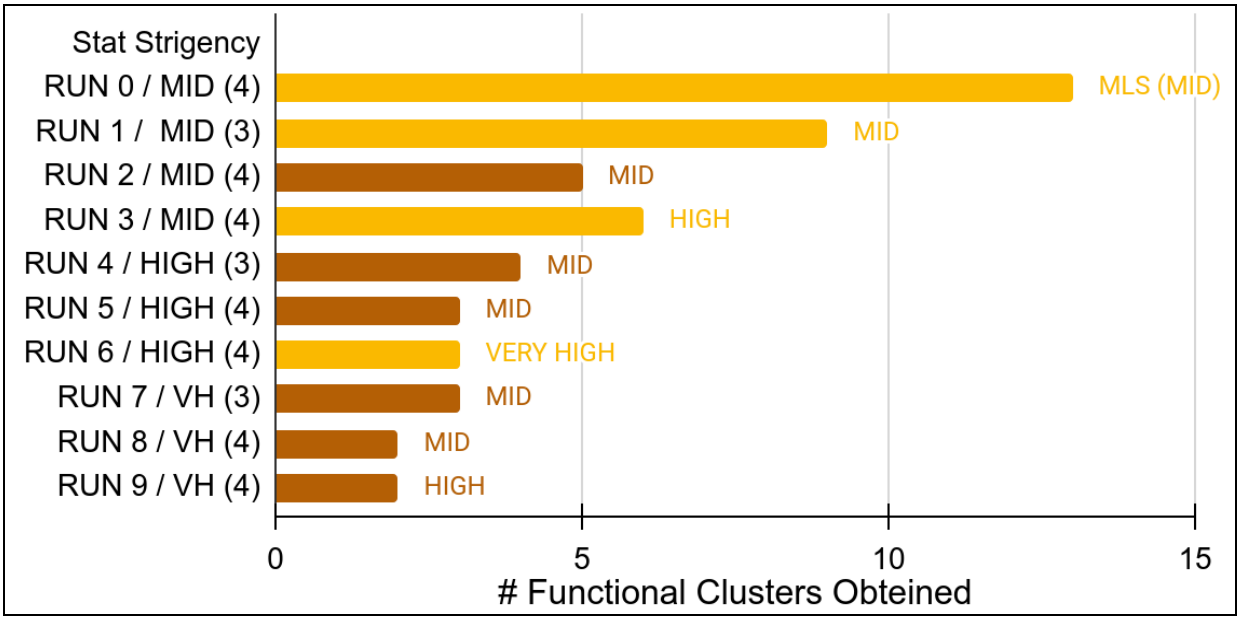


Figura 3.3 Pruebas de clusterización funcional realizadas en DAVID. Pruebas realizadas con diferentes niveles de rigor estadístico, y clusters enriquecidos obtenidos por prueba. Las corridas (RUN) utilizadas para la anotación final se indican en color amarillo. MID = Medio = 0.50 (Default); ALTO = Alto = 0.75; STO = Similarity Term Overlap (Default = 3); MLS =Multi-linkage Threshold (MID, ALTO, Default = 50%).

La prueba (RUN) 0 (rigor MID, STO 4 y MLS MID) produjo 13 clusters con 57 genes asignados, la consulta RUN 1 (rigor MID, STO 3 y MLS MID) produjo 9 clusters con 61 genes asignados, la consulta RUN 3 (rigor ALTO, STO 4 y MLS ALTO) produjo 6 clusters con 71 genes asignados, y la consulta RUN 6 (rigor MID, STO 4 y MSL ALTO) produjo 3 clústeres con 61 clusters asignados (**Figura 3.3**). Los resultados de las 4 pruebas se curaron manualmente para fusionar los agrupamientos con genes repetidos. Se obtuvieron un total de 14 agrupamientos enriquecidos que se filtraron por EASE Score>0.80 y p-value<0.05, resultando en un total de 6 clústeres enriquecidos (**Tabla 3.2**) (**Suplementario 3.2**).

3.3.3 Clusterización funcional de enriquecimiento del módulo *darkmagenta*

Seis clústeres enriquecidos se identificaron en el módulo consenso *Darkmagenta* con EASE Score sobre el valor estadístico de estridencia estándar y p-value<0.05. Los clústeres se etiquetaron como: *WD40/YVTN repeat-like-containing, Sterol metabolism, Glycosyltransferase, intracellular protein transport, Zinc finger, y Methyltransferase (***Tabla 3.2**). Un total de 33 genes se asignaron en los 6 clústeres (**Tabla 3.3**). A continuación se describen (**Suplementario 3.2**).

3.3.3.1 *Clúster 1. Interpro IPR015943: WD40/YVTN repeat-like-containing*

WD40/YVTN repeat-like-containing es una superfamilia que tiene un dominio similar a una repetición WD40/YVTN, el motivo repetido WD40/YVTN consta de ~40 residuos que tiene secuencias distintas pero comparten una estructura similar. Los genes AT5G23430, AT3G13340, AT3G18060, AT5G51980, VPS11, AT1G79990 y VCS forman parte de este clúster (**Tabla 3.2**). **AT1G79990** esta implicado en el transporte mediado por vesículas del retículo endoplásmico (RE) al aparato de Golgi y el transporte de proteínas intracelulares. **VPS11** se involucra en la organización de endosomas, transporte de proteínas intracelulares, fusión de orgánulos, organización de vacuolas, acoplamiento de vesículas y en la exocitosis; VPS11 se requiere para la biogénesis de vacuolas en el embrión, por lo que es esencial para la embriogénesis (Tan et al., 2017) (**Figura 3.4**). **VCS** (VARICOSA) responde a la auxina (AUX), se le ha encontrado en el gránulo de estrés citoplasmático y el núcleo. VCS es una proteína necesaria para el desarrollo de la hoja; como componente del complejo decapping (VCS, DP1 y DCP2) participa en la degradación de mRNAs (Deyholos et al., 2003). **AT5G23430** (KATANIN P80 SUBUNIT 4, KTN80.4) confiere precisión al corte de microtúbulos (MT) mediante la selección específica de complejos de katanina en células vegetales. Esta enzima que corta MT desencadena la reorientación dinámica de las matrices de MT corticales que desempeñan funciones cruciales durante la morfogénesis de las células vegetales, como el alargamiento celular, la biosíntesis de la pared celular y la señalización hormonal. El corte de MT ocurre específicamente en sitios de nucleación de cruce o ramificación en células vivas de arabidopsis (Wang et al., 2017). Otros genes, como **AT5G51980** (ATC3H63) y **AT3G13340** se involucran con la regulación de múltiples procesos. ATC3H63 se involucra en la transcripción, el desarrollo, el fotoperiodismo, la floración, y respuesta al estímulo de luz, mientras AT3G13340 se involucra en el proceso metabólico de macromoléculas, modificación de proteínas, y regulación de la transcripción de la plantilla de DNA. También responde al estímulo lumínico y desarrollo del sistema de brotes. **AT3G18060** (AIP1-2) se expresa bajo el control de la maquinaria de patrones dependiente de WER/WEREWOLF (regulador dependiente de la posición del patrón de células epidérmicas) y la vía de señalización del etileno. Es un modulador de la polaridad mediada por actina.

3.3.3.2 *Clúster 2.* KW-1207 / KW-0752 / GO:0016126: *Sterol metabolism*

Los esteroles juegan un papel crucial como componentes de la membrana y precursores de las hormonas esteroides (por ejemplo, *brasinoesteroides*). Dentro de las membranas los esteroles regulan la permeabilidad y la fluidez de la membrana al interactuar con otros lípidos y proteínas. Encontramos a los genes SMT1, SMO1-1, 3BETAHSD/D1 (**Tabla**

3.2). **3betaHSD/D** parece afectar la actividad del transportador de AUX, posiblemente alterando la composición de esteroles en las membranas (B. Kim et al., 2012). La familia **SMO1** de esterol 4alfa-metiloxidasa es esencial para la embriogénesis regulada por AUX y citoquininas. SMO1-1 trabaja junto con SMO1-2 para mantener la composición correcta de esteroles y equilibrar las actividades de auxina y citoquinina durante la embriogénesis (**Figura 3.4**).

3.3.3.3 *Clúster 3.* GO:0016757 / KW-0328: *Glycosyltransferase*

Las glicosiltransferasas son una subclase importante de enzimas que catalizan la biosíntesis de enlaces glucosídicos en oligosacáridos, polisacáridos y glicoconjugados al transferir un residuo de azúcar desde un sustrato donador a un sustrato aceptor. Encontramos a los genes ALG3, FUT13, CALS1, UK/UPRT1, AT4G38040, GUT2, PARVUS, SETH2, AT5G45660, AT1G34270 (**Tabla 3.2**). **CALS1** (*callose synthase 1*) participa en la regulación de la acumulación de calosa en los canales plasmodesmales, estrategia común para alterar la permeabilidad plasmodesmal tanto bajo la infección por patógenos como bajo el estrés de la herida mecánica. El complejo de genes CALS1 y CALS8 son clave en este proceso, están involucrados en vías de señalización tanto conocidas como nuevas que controlan las respuestas al estrés biótico y abiótico (Cui y Lee, 2016). **FUT13** (*fucosyltransferase 13*) esta involucrado en la síntesis de la pared celular, y **GUT2** (*Exostosin family protein*) esta involucrado en la síntesis de hemicelulosa glucuronoxilano, un componente principal de las paredes celulares secundarias. **ALG3** responde al estímulo luminoso y glicosilación de proteínas (**Figura 3.4**). Plantas con mutaciones en esta proteína tienen perfiles de glicosilación anormales y exhiben respuestas anormales a los MAMP (Trempel et al., 2016). Genes involucrados en otras funciones como **PARVUS** (*Nucleotide-diphospho-sugar transferases superfamily protein*) actúan dentro de la respuesta al ion cadmio, proceso catabólico de xilano y xiloglucano. **SETH2** (*UDP-Glycosyltransferase superfamily protein*) codifica un homólogo de arabidopsis de una proteína conservada involucrada en el primer paso de la ruta biosintética de GPI. **AT1G34270** y **AT4G38040** son proteínas de la familia exotoxina localizadas en el aparato de Golgi. **UK/UPRT1** (uridine kinase/uracil phosphoribosyltransferase 1) juegan un rol dual en la codificación de uridina quinasa y uracilo fosforribosiltransferasa, que forman UMP a través de la vía de recuperación de pirimidina en arabidopsis. **AT5G45660** actúa dentro de la modificación de macromoléculas, regulación del desarrollo postembrionario, y respuesta a la luz. Y **AT1G34270** miembro de la familia de las glicosiltransferasas 47, esta involucrado en la glicosilación de proteínas.

3.3.3.4 *Cluster 4.* KW-0968 / GO:0006886: *Intracellular protein transport*

En esta categoría encontramos a proteínas asociadas con vesículas citoplasmáticas que median el transporte vesicular entre los orgánulos de los sistemas secretor y endocítico. Estas vesículas de transporte se clasifican por la identidad de la cubierta proteica utilizada en su formación y también por la carga que contienen. Encontramos en este clúster a los genes PLA2-ALPHA, PAT2, AT1G60070, AT4G13730, AT1G14910, VPS11, AT1G79990 (**Tabla 3.2**). La interacción proteica entre **PLA2-ALPHA** (*fosfolipasa secretada AtsPLA(2)-alfa*) y AtMYB30 conduce a la represión de la actividad transcripcional de AtMYB30 y a regulación negativa de la respuesta hipersensible (HR) de la planta, caracterizada por una muerte celular rápida localizada en el sitio de inoculación, una de las reacciones de resistencia más eficientes al ataque de patógenos en las plantas. Las plantas mutantes de PLA2-ALPHA son más resistentes a la inoculación bacteriana, mientras que la sobreexpresión conduce a una disminución de la resistencia, confirmando que PLA2-ALPHA es un regulador negativo de la defensa mediada por AtMYB30 (Froidure et al., 2010) (**Figura 3.4**). El gen **PAT2** actúa dentro del aparato de Golgi para el transporte de vacuolas, organización de vacuolas líticas, y regulación del pH intracelular, es una forma negativa dominante de arabidopsis AP-3 β-adaptina (PAT2) que mejora la homeostasis del pH intracelular, sugiriendo que tiene un rol de esta adaptina en el tráfico de canales iónicos o transportadores al tonoplasto (Niñoles et al., 2013). **AT4G13730** localizada en el núcleo se involucra en la activación de la GTPasa y el transporte de proteínas intracelulares. **AT1G14910** (PICALM1B) localizado en la región extracelular se involucra en el ensamblaje de la cubierta de clatrina (CME), la endocitosis dependiente de CME, y la gemación de vesículas de la membrana (evaginación) (**Figura 3.4**). Interactúa con el dominio SNARE de VAMP72 y CME en la membrana plasmática. En las células vegetales, la recuperación de las proteínas de membrana depende en gran medida de la endocitosis mediada por CME (Fujimoto et al., 2020).

3.3.3.5 Clúster 5. IPR011011 / IPR019787: Zinc finger

Los dominios de dedos de zinc (Znf) son motivos proteicos relativamente pequeños que contienen múltiples protuberancias similares a dedos que hacen contactos en tándem con su molécula objetivo. Algunos de estos dominios se unen al zinc, pero muchos otros se unen a metales como el hierro, o a ningún metal, ahora se reconoce que también se unen a sustratos de DNA, RNA, proteínas y/o lípidos. Los Znf muestran una versatilidad considerable en los modos de unión, incluso entre miembros de la misma clase, que sugiere que los motivos Znf

son andamios estables que han desarrollado funciones especializadas. Por ejemplo, las proteínas que contienen Znf funcionan en la transcripción de genes, la traducción, el tráfico de RNAm, la organización del citoesqueleto, el desarrollo epitelial, la adhesión celular, el plegamiento de proteínas, la remodelación de la cromatina y la detección de zinc, por nombrar solo algunos. Los motivos de unión a zinc son estructuras estables y rara vez experimentan cambios conformacionales al unirse a su objetivo. Encontramos en este clúster a los genes EMB1135, ATX2, AT5G12350, AT1G50620 (**Tabla 3.2**). **EMB1135** (*Embryo Defective* 1135, FGT1, FORGETTER 1) responde a la aclimatación al calor para adaptarse al estrés ambiental y mitigar sus efectos adversos, parte de dicha adaptación es mantener una memoria activa del estrés por calor durante varios días que promueva una respuesta más eficiente al estrés recurrente. Mutantes EMB1135 muestran un mantenimiento reducido de la expresión génica inducida por calor. EMB1135 interactúa con los remodeladores de cromatina de las familias SWI/SNF e ISWI que también muestran una memoria de estrés por calor reducida (Brzezinka et al., 2016). **ATX2** (*Arabidopsis Trithorax 1*) y **AT1G50620** actúan en la modificación de peptidil-lisina, la regulación de la transcripción con plantilla de DNA, y la regulación del proceso de desarrollo. **AT5G12350** es un regulador de la familia de condensación cromosómica (RCC1) que contiene el dominio de dedos de zinc FYVE. En muchas especies de plantas, las raíces mantienen ángulos de crecimiento específicos conocidos como ángulos de punto de ajuste gravitrópicos (GSA) que contribuyen a la adquisición eficiente de agua y nutrientes. Los genes AtLAZY1/LAZY1-LIKE (LZY) están involucrados en el control de GSA al regular el flujo de auxina hacia la dirección de la gravedad en arabidopsis, se ha encontrado que las proteínas de dominio similar a RCC1 (RLD), identificadas como interactores LZY, son reguladores esenciales del transporte de AUX. La interacción Chemokine Ligand (CCL) de LZY con el dominio *Brevis radix* (BRX) y RLD es importante para el reclutamiento de RLD desde el citoplasma a la membrana plasmática por LZY (Furutani et al., 2020) (**Figura 3.4**).

3.3.3.6 *Clúster 6.* KW-0489: *Methyltransferase*

Este es un gran grupo de enzimas que metilan sus sustratos y se pueden dividir en varias subclases de acuerdo a sus características estructurales. La clase más común de metiltransferasas es la Clase I, que contienen un pliegue de *Rossmann* para unirse a la *S-adenosil metionina* (SAM); la Clase II que contienen un dominio SET, ejemplificado por histonas *metiltransferasas* de dominio SET, y la Clase III, que están asociadas a la membrana. Las metiltransferasas también se pueden agrupar en diferentes tipos que utilizan diferentes sustratos en las reacciones de transferencia de metilo, incluyen proteínas metiltransferasas,

DNA/RNA metiltransferasas, metiltransferasas de productos naturales y metiltransferasas no dependientes de SAM. SAM es el donante de metilo clásico para las metiltransferasas, sin embargo, en la naturaleza se ven ejemplos de otros donantes de metilo. Las reacciones enzimáticas que producen se encuentran en muchas vías y están implicadas en enfermedades genéticas, cáncer y enfermedades metabólicas. Encontramos en este clúster a los genes AT3G15530, SUVH4, AT2G34300, SMT1, ATX2, AT5G06050 (**Tabla 3.2**). **ATX2** también proteína de la superfamilia Znf actúa dentro de la modificación de peptidil-lisina, la regulación de la transcripción con plantilla de DNA y la regulación del proceso de desarrollo. **AT3G15530** localizada en el citoplasma responde a la defensa frente a hongos y bacterias, respuesta al frío, sustancias inorgánicas, estrés oxidativo, y estrés por heridas. **SUVH4** responde al mantenimiento de la metilación del DNA. SUVH4 es una proteína de dominio SET implicada en el control epigenético de la expresión génica. Los complejos KYP/SUVH4 estar putativamente involucrados en el proceso independiente de la telomerasa conocido como alargamiento alternativo de los telómeros. Pruebas realizadas en arabidopsis para medir la respuesta a varios estresores mostró una reacción disminuida ante el estrés osmótico (Luhua et al., 2013), también hay evidencia de su participación en el silenciamiento génico transcripcional (TGS) que puede servir como una inmunidad innata contra los virus de DNA invasores en los eucariotas. KYP se une a la cromatina viral y controla su metilación para combatir la infección por virus (Castillo-González et al., 2015) (**Figura 3.4**). **AT2G34300** proteína de la superfamilia de metiltransferasas dependiente de S-adenosil-L-metionina localizada en múltiples organelos, se expresa durante las etapas de crecimiento y desarrollo. **SMT1** (*sterol methyltransferases 1*) es un esterol vegetal miembro de la superfamilia de metiltransferasa de unión a SAM de clase I, encontrado altamente expresado en tejido vascular, hojas maduras y en regiones en proceso de expansión celular. Se ha encontrado que mutantes de SMT1 influyen en la estructura y el tráfico de la membrana. Mutantes SMT1, hyd2, cpi y cvp1 exhiben una mala distribución de la proteína PIN de localización polar, un transportador de salida de AUX. Mutantes de SMT1 son también letales para los embriones (Carland et al., 2010). **AT5G06050** proteína putativa de la familia de las metiltransferasas, esta activa en el citoplasma y la región extracelular.

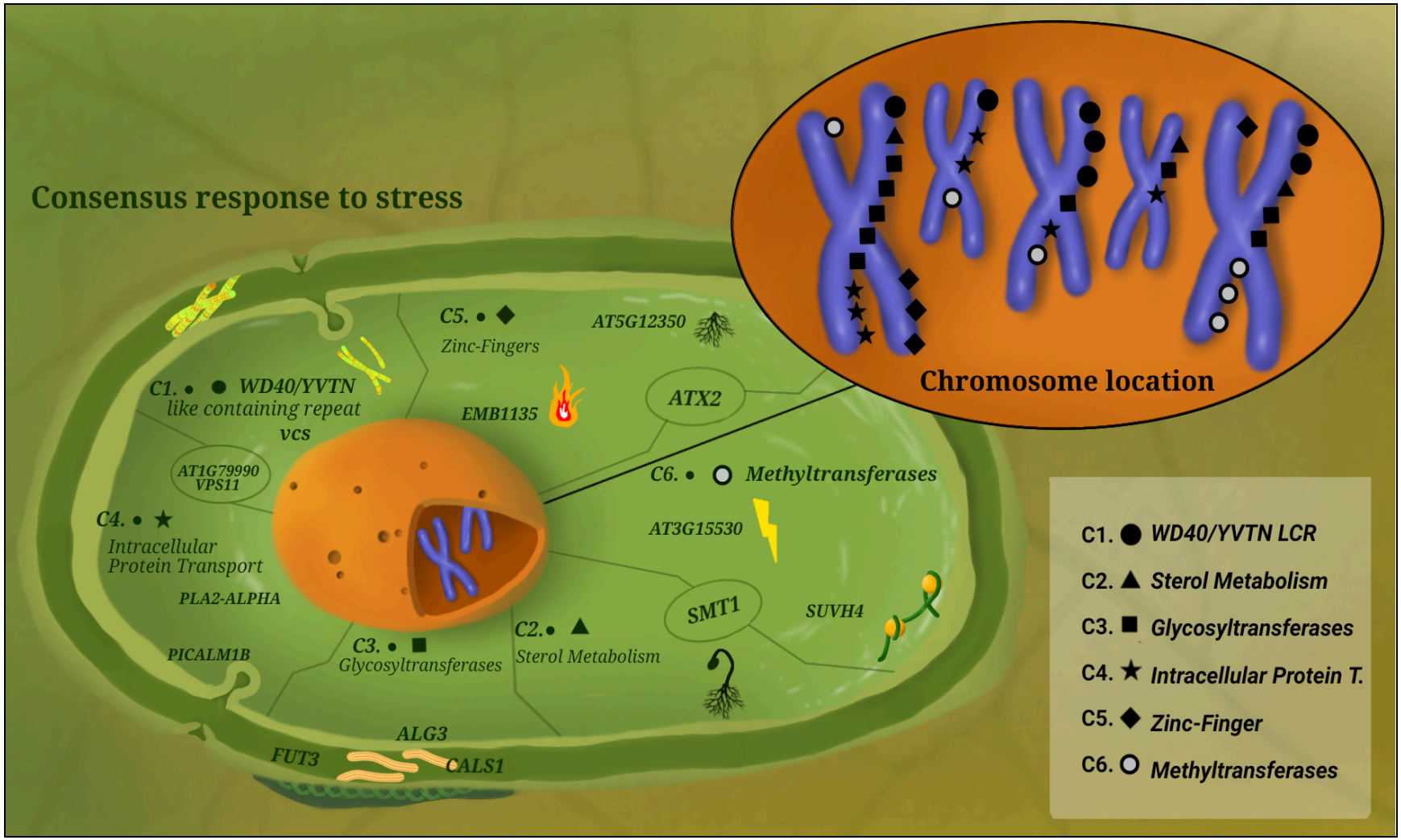


Figura 3.4 Respuesta consenso en *A thaliana* ante estrés causado por *B cinerea* y *C higginsianum* a las 24 y 22 hpi. La imagen de la célula muestra los seis clústeres enriquecidos en el módulo consenso *Darkmagenta*, etiquetados con símbolos específicos indicados en el cuadro con las leyendas. Los círculos dibujados entre los clústeres indican genes que comparten más de una función enriquecida, los cuales además tienen altos polimorfismos de secuencias. En la parte superior se muestra la localización de los genes en los cromosomas, los cuales se especifican de acuerdo al símbolo asignado a cada clúster.

3.5 FUNCIONES COMPARTIDAS Y POLIMORFISMOS DE SECUENCIA

Dentro de los 33 genes enriquecidos, encontramos a los genes AT1G79990, VPS11, SMT1 y ATX2 compartiendo dominios funcionales entre sí (**Tabla 3.5**). Los genes **AT1G79990** (coatomer subunit beta-2) y **VPS11** comparten funciones involucradas en los dominios funcionales de *C1. WD40/YVTN repeat-like-containing y C4. Intracellular protein transport* comparten; **AT1G79990** esta implicado en el transporte mediado por vesículas y el transporte de proteínas intracelulares, y **VPS11** en la organización del transporte de proteínas intracelulares, acoplamiento de vesículas y la exocitosis (Tan et al., 2017). El gen **SMT1** comparte funciones entre los clústeres *C2. Sterol metabolism y C6. Methyltransferase*, reporta actividades relacionadas con el control del nivel de esteroles en las plantas; los mutantes de esteroles influyen en la estructura y el tráfico de la membrana, y en su ausencia exhiben una mala distribución de la proteína PIN de localización polar, un transportador de salida de AUX (*Auxina*); se ha encontrado que los defectos morfológicos celulares en fk, smt1 e hyd1 están asociados con huecos en la pared celular y engrosamientos aberrantes de la pared celular con depósitos ectópicos de lignina y callosa (Carland et al., 2010). El gen **ATX2** (*A. Trithorax 1*) comparte funciones de los clústeres *C5. Zinc finger y C6. Methyltransferase*, **ATX2** actúa en la regulación de la transcripción y la regulación del proceso de desarrollo, la regulación epigenética y demetilación de la histona H3-lisina 4; estos estados de metilación de Lys (Lisina) se han clasificado como marcas represivas y activadoras, según su efecto sobre la expresión génica; **ATX2** actúa en la regulación epigenética del represor floral FLC y FT para prevenir la transición del desarrollo vegetativo al reproductivo (Pien et al., 2008).

Tabla 3.2. Agrupamientos enriquecidos del módulo consenso *Darkmagenta*.

Clúster and E. Score	Category	Term	*p* value	P Benjamini	#Genes (merged)	Gene Names	Stats. Stringency
Cluster 1 E.Score: 1.5527	IPR015943	WD40/YVTN repeat-like-containing	0.0029	0.5307	7	AT5G23430, AT3G13340, AT3G18060, AT5G51980, VPS11, AT1G79990, VCS	M
Cluster 2 E.Score: 1.3511	KW-1207 KW-0752 GO:0016126	Sterol metabolism	0.0125 0.0275 0.0277	0.4613 0.7459 0.9995	3	SMT1, SMO1-1, 3BETAHSD/D1	H
Cluster 3 E.Score: 1.3489	GO:0016757 KW-0328	Glycosyltransferase	0.0113 0.0346	0.8358 0.7467	10	ALG3, FUT13, CALS1, UK/UPRT1, AT4G38040, GUT2, PARVUS, SETH2, AT5G45660, AT1G34270	VH
Cluster 4 E.Score: 1.1112	KW-0968 GO:0006886	intracellular protein transport	0.0066 0.0428	0.1592 0.9999	7	PLA2-ALPHA, PAT2, AT1G60070, AT4G13730, AT1G14910, VPS11, AT1G79990	H
Cluster 5 E.Score: 1.0173	IPR011011 IPR019787	Zinc finger, FYVE/PHD-type	0.0190 0.0477	0.9928 0.9999	4	EMB1135, ATX2, AT5G12350, AT1G50620	H
Cluster 6 E.Score: 0.8795	KW-0489	Methyltransferase	0.0403	0.7991	6	AT3G15530, SUVH4, AT2G34300, SMT1, ATX2, AT5G06050	H

*Solo clusters con E.Score>0.75 y valores p<0.05 son considerados. Abrev. M=Moderate; H=High; VH=Very High.

Tabla 3.3. Anotación de los 33 genes asignados a los 6 clusters enriquecidos.

Clúster	Gen ID	Nombre	Polym	Chr	Descripción
1	AT1G79990	AT1G79990	50	Chr1	Transducin/WD40 repeat-like superfamily protein(VCS)
1	AT2G05170	VPS11	5	Chr2	Transducin/WD40 repeat-like superfamily protein(AT3G13340)
1	AT3G13300	VCS	35	Chr3	transducin family protein / WD-40 repeat family protein(AT3G18060)
1	AT3G13340	AT3G13340	27	Chr3	Transducin/WD40 repeat-like superfamily protein(AT5G23430)
1	AT3G18060	AT3G18060	6	Chr3	Transducin/WD40 repeat-like superfamily protein(AT5G51980)
2	AT5G23430	AT5G23430	54	Chr5	3beta-hydroxysteroid-dehydrogenase/decarboxylase (3BETAHSD/D1)

2	AT5G51980	AT5G51980	16	Chr5	sterol-4alpha-methyl oxidase 1-1(SMO1-1)
3	AT1G47290	3BETAHSD/D1	18	Chr1	callose synthase 1(CALS1)
3	AT4G12110	SMO1-1	5	Chr4	Nucleotide-diphospho-sugar transferases superfamily protein(PARVUS)
3	AT5G13710	SMT1	26	Chr5	Exostosin family protein(GUT2)
3	AT1G05830	ATX2	96	Chr1	Exostosin family protein(AT1G34270)
3	AT1G50620	AT1G50620	4	Chr1	fucosyltransferase 13(FUT13)
3	AT1G79350	EMB1135	31	Chr1	UDP-Glycosyltransferase superfamily protein(SETH2)
3	AT5G12350	AT5G12350	9	Chr5	Exostosin family protein(AT4G38040)
3	AT1G05570	CALS1	169	Chr1	uridine kinase/uracil phosphoribosyltransferase 1(UK/UPRT1)
3	AT1G14910	AT1G14910	30	Chr1	adenine phosphoribosyltransferase(AT5G45660)
3	AT1G19300	PARVUS	1	Chr1	asparagine-linked glycosylation 3(ALG3)
4	AT1G27440	GUT2	4	Chr1	ENTH/ANTH/VHS superfamily protein(AT1G14910)
4	AT1G34270	AT1G34270	3	Chr1	Adaptor protein complex AP-1, gamma subunit(AT1G60070)
4	AT1G60070	AT1G60070	37	Chr1	Phospholipase A2 family protein(PLA2-ALPHA)
4	AT1G71990	FUT13	3	Chr1	protein affected trafficking 2(PAT2)
4	AT2G06925	PLA2-ALPHA	9	Chr2	Ypt/Rab-GAP domain of gyp1p superfamily protein(AT4G13730)
5	AT2G34300	AT2G34300	21	Chr2	RING/FYVE/PHD zinc finger superfamily protein(AT1G50620)
5	AT2G47760	ALG3	66	Chr2	RING/FYVE/PHD zinc finger superfamily protein(EMB1135)
5	AT3G15530	AT3G15530	4	Chr3	Regulator of chromosome condensation (RCC1) family with FYVE zinc finger domain-containing protein(AT5G12350)
6	AT3G45100	SETH2	24	Chr3	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein(AT2G34300)
6	AT3G55480	PAT2	20	Chr3	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein(AT3G15530)
6	AT4G13730	AT4G13730	27	Chr4	Putative methyltransferase family protein(AT5G06050)
6	AT4G38040	AT4G38040	9	Chr4	histone-lysine N-methyltransferase, H3 lysine-9 specific SUVH4-like protein(SUVH4)
1-4	AT5G06050	AT5G06050	9	Chr5	coatomer subunit beta-2(AT1G79990)
1-4	AT5G13960	SUVH4	26	Chr5	vacuolar protein sorting 11(VPS11)
2-6	AT5G40870	UK/UPRT1	14	Chr5	sterol methyltransferase 1(SMT1)
5-6	AT5G45660	AT5G45660	2	Chr5	trithorax-like protein 2(ATX2)

*Listado de anotaciones de PANTHER

Tabla 3.4. Clústeres con genes compartidos y polimorfismos de secuencias.

Clústeres Anotados	#Genes compartidos	Nombre del gen	#Polimorfismos de secuencia
C1. WD40/YVTN repeat-like-containing	2	AT1G79990	50
C4. Intracellular protein transport		VPS11	5
C2. Sterol metabolism C6. Methyltransferase	1	SMT1	26
C5. Zinc-finger C6. Methyltransferase	1	ATX2	96

3.6 POLIMORFISMOS DE SECUENCIA

Tres de los cuatro genes que comparten funciones en los clústeres enriquecidos (**Figura 3.4**) también contienen un alto grado de polimorfismos de secuencia que va de 26 a 96 isoformas por gen, exceptuando al gen VPS11, que solo contiene solo 5. Aunque no se realizó un análisis detallado a nivel poliformismo sobre todos los genes del módulo Darkmagenta, encontramos que >50% de los genes asignados a los clústeres enriquecidos (33) tienen >20 polimorfismos por gen (**Figura 3.5**).

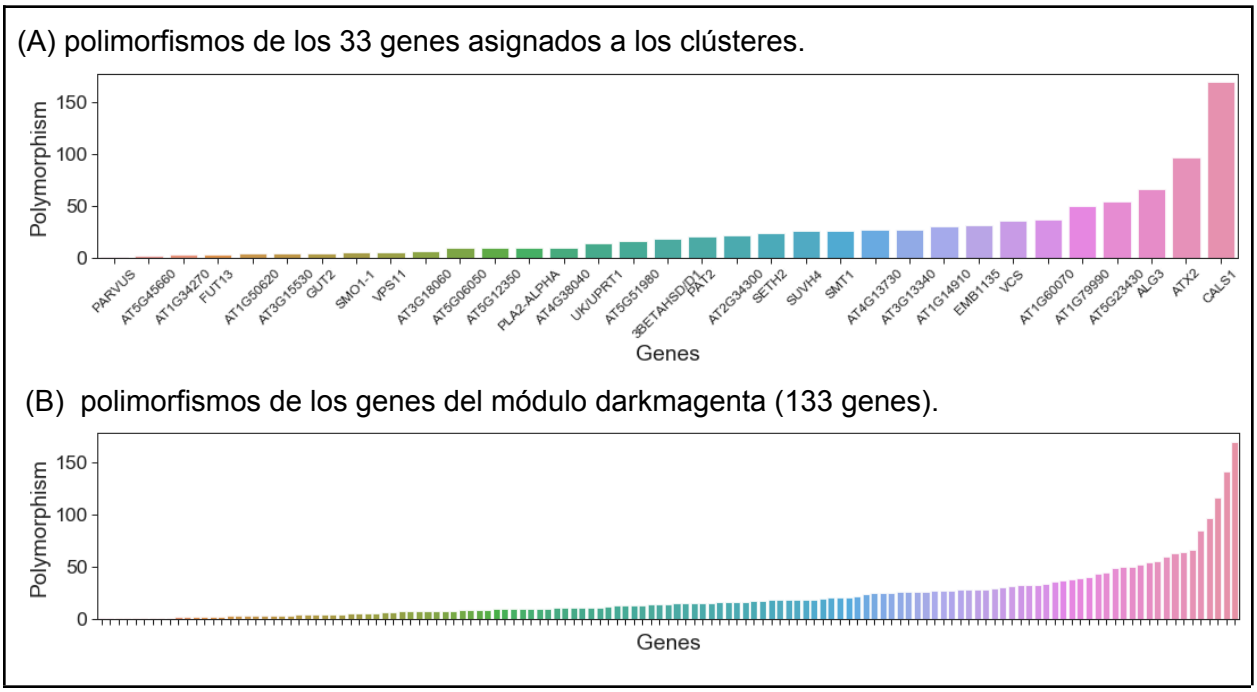


Figura 3.5 Polimorfismos de los genes en el módulo *Darkmagenta*.

Entre los genes con más alta frecuencia de polimorfismos (**Tabla 3.5**) se encuentran:

1. **CALS1** (*Callose Synthase 1*), miembro de la familia de las glicosiltransferasas presente en el plasma de la membrana, involucrado en la respuesta celular a compuestos que contiene oxígeno, el ácido salicílico (SA), y la transducción de señales.

2. **ATX2** presente en los clústeres *Zinc finger y Methyltransferase*, involucrado en la regulación epigenética del represor floral FLC y FT (Pien et al., 2008).

3. **ALG3** actuando en la glicosilación de proteínas. Las plantas con mutaciones en esta proteína tienen perfiles de glicosilación anormales y exhiben respuestas anormales a los

MAMP, lo que da como resultado una subglucosilación de varios PRR y una señalización comprometida de MAMP/DAMP (Trempel et al., 2016).

4. **AT5G23430** (*Katanin P80 Subunit* 4), localizado en el citoesqueleto de microtúbulos, actuando en la organización y seccionamiento de microtúbulos corticales.

5. **AT1G60070** (*Complex Gamma Subunit* 1), involucrado en Golgi para el transporte de vacuolas, proteínas intracelulares, y transporte mediado por vesículas.

6. **VCS** (VARICOSE), el cual forma un complejo de decapado de RNAm con DCP1 (At1g08370) y DCP2 (At5g13570), colocalizado en locis citoplasmáticos. Los mutantes nulos de DCP1, DCP2 y VCS comparten un fenotipo letal similar en la etapa de cotiledón de la plántula, con morfología celular epidérmica alterada. VCS se requiere para el desarrollo de las hojas.

7. **EMB1135** (*Embryo Defective* 1135) involucrado en la regulación epigenética de la expresión génica en la respuesta celular al calor. Es un un coactivador altamente conservado del regulador del desarrollo Notch, y media la memoria de la cromatina inducida por el estrés. Esta memoria de estrés por calor requiere EMB1135 (Brzezinka et al., 2016).

8. **AT1G14910** (PICALM1B) localizado en la región extracelular, esta involucrado en el ensamblaje de la cubierta de clatrina (CME), la endocitosis dependiente de clatrina, y la gemación de vesículas de la membrana (evaginación de membrana). Interactúa con el dominio SNARE de VAMP72 y CME en la membrana plasmática. En doble mutante de PICALM1 se exhibe un desarrollo vegetativo deteriorado que indica el requisito de reciclaje de VAMP72 para el crecimiento normal de la planta (Fujimoto et al., 2020).

Tabla 3.5. Ranking de polimorfismos encontrados en las anotaciones.

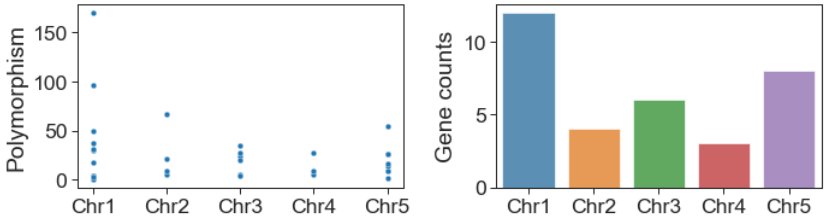
TAIR ID	Nombre Gen	Polim.	Chr	Description
AT1G05570	CALS1	169	Chr1	callose synthase 1(CALS1)
AT1G05830	ATX2	96	Chr1	trithorax-like protein 2(ATX2)
AT2G47760	ALG3	66	Chr2	asparagine-linked glycosylation 3(ALG3)
AT5G23430	AT5G23430	54	Chr5	Transducin/WD40 repeat-like superfamily protein(AT5G23430)
AT1G79990	AT1G79990	50	Chr1	coatomer subunit beta-2(AT1G79990)
AT1G60070	AT1G60070	37	Chr1	Adaptor protein complex AP-1, gamma subunit(AT1G60070)

AT3G13300	VCS	35	Chr3	Transducin/WD40 repeat-like superfamily protein(VCS)
AT1G79350	EMB1135	31	Chr1	RING/FYVE/PHD zinc finger superfamily protein(EMB1135)
AT1G14910	AT1G14910	30	Chr1	ENTH/ANTH/VHS superfamily protein(AT1G14910)
AT3G13340	AT3G13340	27	Chr3	Transducin/WD40 repeat-like superfamily protein(AT3G13340)
AT4G13730	AT4G13730	27	Chr4	Ypt/Rab-GAP domain of gyp1p superfamily protein(AT4G13730)
AT5G13960	SUVH4	26	Chr5	histone-lysine N-methyltransferase, H3 lysine-9 specific SUVH4-like protein(SUVH4)
AT5G13710	SMT1	26	Chr5	sterol methyltransferase 1(SMT1)

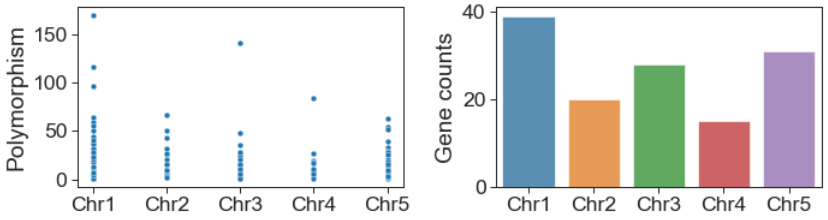
3.7 LOCALIZACIÓN CROMOSOMAL DE LOS GENES ASIGNADOS

Sesenta y seis por ciento (66%), es decir, 20 genes de los 33 genes asignados a los 6 clústeres se encuentran en los cromosomas 1 y 3 (**Figura 3.6A**). La relación de este patrón contra el módulo genético completo (133 genes) (**Figura 3.6B**), y los genes codificantes de proteínas del genoma completo de arabidopsis (**Figura 3.6C**) es proporcional, por lo que no encontramos ninguna relación significativa.

(A) Los 33 genes asignados en los 6 clústeres.



(B) Los 133 genes del módulo génico darkmagenta.



(C) Los más de 27 mil genes del genoma de arabidopsis.

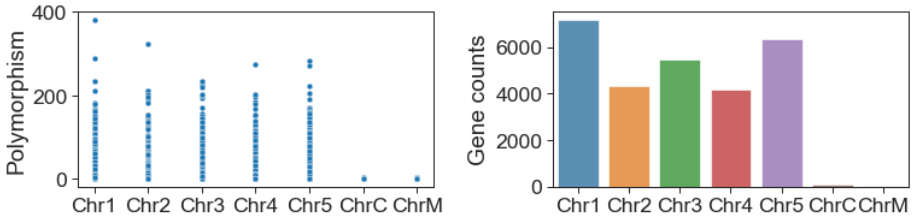


Figura 3.6 Localización cromosomal de los genes. (A) Los 33 genes asignados a los 6 clústeres enriquecidos. (B) Los 133 genes del módulo génico *Darkmagenta*. (C) Los más de 27 mil genes del genoma de arabidopsis.

La discusión y conclusiones finales de esta sección se presentan en el siguiente capítulo.

CAPÍTULO IV

DISCUSIÓN, CONCLUSIONES GENERALES Y PERSPECTIVAS

4.1 DISCUSIÓN

El mecanismo de respuesta de las plantas al estrés es muy complejo y requiere que se perciban y activen varias vías integradas de respuesta. La detección de factores de estrés tanto biótico como abióticos inicia varias vías de señalización complejas, algunos de los eventos de señalización temprana incluyen, la alteración de la concentración de Ca^{2+} intracelular, la producción de moléculas de señalización secundarias (fosfoinosítido y especies reactivas de oxígeno/ROS), y la activación de cascadas de quinasas. La detección de estrés en las plantas es determinante para el establecimiento de una estrategia de defensa exitosa.

Treinta y tres genes del clúster consenso de arabidopsis en respuesta a la infección causada por *B cinerea a las 24 hpi* y *C higginsianum* a las 22 hpi, reveló seis vías de respuesta a la infección con soporte en la información disponible en 9 BDs ómicas, encontramos que los *WD40/YVTN repeat-like-containing, Sterol metabolism, Glycosyltransferase, intracellular protein transport, Zinc finger, y Methyltransferases,* son procesos activos en la respuesta a la infección en este umbral de respuesta, que ubicamos en el contexto de *in-planta appressorium*. Entre los genes encontrados, hay genes codificantes de callosa sintasa, otros que modifican la permeabilidad bajo la infección por patógenos y estrés mecánico, como **CALS1** (Fernandez-Calvino et al., 2011), y otros genes que participan en la síntesis de la pared celular, como **GUT2** y **FUT13**, que participan en el proceso biosintético de *glucuronoxilano* para la formación de las paredes celulares y la biogénesis de la pared celular secundaria (Hooper et al., 2017)**.** Adicionalmente se observaron en el mismo módulo, otros 100 genes con alta conectividad no asignados a ninguna función. El estudio de estos genes abonará a encontrar más respuestas relacionadas con los mecanismos de defensa comunes y su regulación.

El gen **ALG3** (*Asparagine-Linked Glycosylation* 3) responde a la señalización de Ca^{2+} atenuada después de la exposición a MAMP; el calcio como segundo mensajero convierte las señales extracelulares en reacciones, su incremento en el citosol es una de las primeras respuestas de las plantas ante la exposición de MAMP, que generalmente se traduce en "estallido oxidativo" como respuesta a las condiciones de estrés. Se requiere una glicosilación adecuada de los receptores MAMP para obtener respuestas óptimas a los MAMP, por lo que es

importante para la señalización inmunitaria del huésped. Se ha encontrado en arabidopsis que la glicosilación alterada de las proteínas exportadas, incluidos los receptores inmunitarios de superficie, comprometen el Ca^{2+} y la señalización posterior a los PAMP (Trempel et al., 2016).

Otros genes involucrados en el transporte intracelular de proteínas como **PLA2-ALPHA** (*Phospholipase A2-Beta*), uno de los cuatro genes PLA2 en arabidopsis, participa en la apertura estomática en respuesta a la luz y se expresa en células guardianes, también es un regulador negativo de la defensa mediada por AtMYB30, importante para la translocación de proteínas al núcleo para la regulación génica asociada a la defensa, su sobreexpresión conduce a una disminución de la resistencia (Froidure et al., 2010). El gen **PAT2** mejora la homeostasis del pH intracelular, es parte del complejo AP-3 que parece estar asociado con la clatrina, asociado con la región de Golgi, así como con estructuras más periféricas; PAT2 facilita la brotación de vesículas desde la membrana de Golgi y puede estar directamente involucrada en el tráfico hacia la vacuola, sin embargo *no se le ha asociado con la respuesta de defensa*. **AT4G13730** actúa en el transporte de proteínas intracelulares, y **PICALM1B** participa en la endocitosis dependiente de clatrina (CME), y la evaginación de membrana, descrita como la principal vía endocítica que apoya las funciones básicas de las células; muchas proteínas de membrana se endocitan para activar mecanismos de respuesta de la planta y regular componentes de la pared celular, posiblemente también actué para producir compuestos antimicrobianos y/o secretar proteínas de defensa en la respuesta a la infección en *A thaliana* por *B cinerea* y *C higginsianum,* aunque no se le ha reportado.

En el clúster *Zinc finger* (Znf) encontramos al gen **EMB1135**, el cual responde a la aclimatación al calor y estrés ambiental, permitiendo mitigar los efectos adversos. Dicha adaptación mantiene una memoria activa del estrés por calor durante varios días que promueva una respuesta más eficiente al estrés recurrente (Brzezinka et al., 2016). Otros genes encontrados en el clúster de las *Methyltransferases* como **AT3G15530** se involucra en la respuesta de defensa al hongo, crecimiento, vía de señalización mediada por hormonas, regulación de la transcripción con plantilla de DNA, respuesta al alcohol, a las bacterias, al frío, a sustancias inorgánicas, al estrés oxidativo, y respuesta a las heridas. Se expresa en el núcleo, la vacuola y la membrana plasmática. El gen **SUVH4** se ha encontrado involucrado en el silenciamiento génico transcripcional, que puede servir como inmunidad innata contra los virus de DNA. Otros genes como **SMO1** son esenciales para el mantenimiento de la composición correcta de esteroles, o **VPS11** que es esencial para la embriogénesis y procesos

de desarrollo postembrionario y la respuesta a la luz; los encontramos más asociados a *funciones esenciales de mantenimiento, que involucrados en la respuesta a la infección.*

El análisis de los *polimorfismos* de los genes asignados a los 6 clústeres (33 genes) reveló una alta frecuencia de polimorfismos, se encontró qué más del 50% contienen al menos 20 polimorfismos de secuencia. Genes como CALS1, ATX2, ALG3, AT5G23430 (*Katanin* P80 *Subunit* 4) y AT1G79990 (*Complex Gamma Subunit* 1), encabezan la lista con más de 50 polimorfismos reportados. Los resultados indican que la variabilidad encontrada en las secuencias podría tener una implicación directa en la activación de múltiples vías de respuesta, ya que varios de los genes reportados (AT1G79990, VPS11, SMT1, ATX2) se encuentran en más de una vía de las señaladas. Sin embargo, se requieren más estudios para establecer una correlación directa. Con respecto a la localización cromosomal de los genes, 66% (20 genes) se encuentran en los cromosomas 1 y 3, sí bien esto podría sugerir un loci de respuesta para estos cromosomas, la localización de los 133 genes del cúster *Darkmagenta,* y los más de 27 mil genes CDS del genoma, tienen una distribución similar, por lo que es posible que este dato sea más atribuible al tamaño relativo de los cromosomas.

Hasta hoy en día no se tenía un modelo sobre la respuesta consenso de defensa de arabidopsis trás la infección con diversos hongos; Rodrigo et.al. (2012) realizó un estudio similar tras la infección con 7 diferentes virus, y también existen diversas BDs (ATTEND-II, Expression Angler, AtCAST, Aranet y PlaD) que presentan modelos de predicción sobre estas interacciones, sin embargo contienen diversas limitaciones para comparar y extraer información. Los estudios a nivel multisistema permitirán alcanzar una perspectiva más amplia sobre los mecanismos de respuesta de las plantas, lo cual será imperante para dar cobertura a los problemas actuales que enfrentamos de globalización y cambio climático extremo, que potencia la diseminación de enfermedades en cultivos relevantes para la seguridad alimentaria. Con el avance de NGS se pueden construir estos modelos incluso para organismos no-modelo, que han estado limitados principalmente debido a las inherentes limitaciones que pueden recabarse a partir de un organismo no-modelo (Alfred y Baldwin, 2015), por ello, fue necesario partir de una especie como arabidopsis, que ofrece importantes ventajas para la investigación básica en genética y biología molecular, desde el punto de vista técnico contiene un genoma estable para el desarrollo de cualquier estudio basado en datos y, desde el punto de vista biológico tiene la ventaja atribuible a los genomas pequeño, una corregulación más evidente de múltiples genes relacionados, debido a que la densidad de genes generalmente aumenta con la disminución del tamaño del genoma (Sela et al., 2016).

4.2 CONCLUSIONES GENERALES

El patrón de la expresión génica consenso encontrado en la respuesta de *A thaliana* al estrés causado por *B cinerea* y *C higginsianum*, contiene genes tan esenciales para la supervivencia de la planta como SMO1 y VPS11, involucrados con el mantenimiento de la composición correcta de esteroles y el transporte celular para la embriogénesis, la regulación del desarrollo postembrionario y respuesta a la luz. Sin embargo, hemos identificado en el módulo consenso 33 genes asignados a 6 vías de respuesta activas durante este umbral de respuesta a la infección, con genes que responden al estrés en diversas categorías; en relación directa con patógenos encontramos al gen CALS1, participando en la regulación de la formación de calosa, a los genes FUT13 y GUT2 participando en la síntesis de la pared celular, al gen PLA2-ALPHA, regulador negativo de la defensa mediada por AtMYB30, cuya sobreexpresión conduce a una disminución de la resistencia; también encontramos a los genes AT4G13730 y PICALM1B que participan en la CME y la evaginación de membrana, al gen AT3G15530 que responde a la defensa fúngica y bacteriana, pero que sin embargo esta ligado a muchos otros tipos de respuesta tanto bióticos como abióticos, y el gen SUVH4 que se involucra en el control epigenético de la expresión, conduciendo al silenciamiento génico transcripcional. Otros genes encontrados, reportan diversas funciones relacionadas con estrés abiótico, como el gen EMB1135 que responde a la aclimatación al calor y estrés ambiental, manteniendo una memoria activa del estrés por calor durante varios días.

4.3 LIMITACIONES EN EL ESTUDIO

1. El estudio con organismos modelo se encuentra inherentemente restringido por la información que puede recabarse a partir del modelo vegetal utilizado, no es posible estudiar todos los procesos biológicos a partir del análisis de una sola especie.
2. Los resultados encontrados están limitados a los tratamientos incluidos en este estudio.
3. El método de normalización utilizado (TPMs) tiene sus propias limitaciones, sin embargo, se consideraron los supuestos conocidos para su utilización, y se invita a revisarlos o consultar cualquier otro método, antes de decidir el método a utilizar.

4.4 PERSPECTIVAS

Considerando las características de los datos, qué son fundamentalmente el primer eslabón para determinar la técnica de análisis, el desafío a resolver fue establecer pautas para evaluar los datos y apegarnos a los supuestos estadísticos planteados; se propuso una forma

alternativa para integrar datos públicos, práctica, sencilla y efectiva, que derivó en la propuesta metodológica planteada en el capítulo II; sería ideal integrar a la metodología planteada otras técnicas para el manejo de la variabilidad a nivel gen, tomando en consideración otros metadatos, una consecuencia derivada de la avalancha de ómicas alojadas en los repositorios públicos, que son cada vez más numerosos y diversos, y exigen seguir produciendo alternativas que fomenten la utilización de estas minas de datos. Los enfoques no solo requieren ser robustos, sino además flexibles para ser útiles bajo múltiples contextos de análisis, el desafío esta vigente y continuará creciendo para cubrir las demandas vigentes de despliegue de este tipo alternativas para completar el ciclo de análisis.

En la respuesta al estrés fúngico encontrada en arabidopsis en este estudio qué incluye genes de la respuesta al estrés por infección por hongo, pero también otros ligados a múltiples estresores bióticos, profundizar en el estudios más integral de plantas infectadas con otros hongos distintos a los señalados, podrá consolidar estos hallazgos, con la finalidad de establecer un core de genes activado ante el estrés causado por múltiples estresores fúngicos.

BIBLIOGRAFÍA

Abrams, Z. B., Johnson, T. S., Huang, K., Payne, P. R. O., and Coombes, K. (2019). A protocol to evaluate RNA sequencing normalization methods. BMC Bioinformatics, 20(24), 679. https://doi.org/10.1186/s12859-019-3247-x

Alfred, J., and Baldwin, I. T. (2015). New opportunities at the wild frontier. *eLife*, *4*, e06956. https://doi.org/10.7554/eLife.06956

Austin, R. S., Hiu, S., Waese, J., Ierullo, M., Pasha, A., Wang, T. T., Fan, J., Foong, C., Breit, R.,Desveaux, D., Moses, A., and Provart, N. J. (2016). New BAR tools for mining expression data and exploring cis-elements in *Arabidopsis thaliana*. The Plant Journal: For Cell and Molecular Biology, 88(3), 490–504. https://doi.org/10.1111/tpj.13261

Babarinde, I. A., Li, Y., and Hutchins, A. P. (2019). Computational methods for mapping, assembly and quantification for coding and non-coding transcripts. Computational and Structural Biotechnology Journal, 17, 628–637.https://doi.org/10.1016/j.csbj.2019.04.012

Badet, T., Voisin, D., Mbengue, M., Barascud, M., Sucher, J., Sadon, P., Balagué, C., Roby, D., and Raffaele, S. (2017). Parallel evolution of the POQR *prolyl oligo peptidase* gene conferring plant quantitative disease resistance. PLOS Genetics, 13, e1007143. https://doi.org/10.1371/journal.pgen.1007143

Bi, K., Liang, Y., Mengiste, T., and Sharon, A. (2022). Killing softly: A roadmap of *Botrytis cinerea* pathogenicity. Trends in Plant Science, 0(0).https://doi.org/10.1016/j.tplants.2022.08.024

Boulent, J., Foucher, S., Théau, J., and St-Charles, P.-L. (2019). Convolutional neural networks for the automatic identification of plant diseases. Frontiers in Plant Science, 10. https://www.frontiersin.org/article/10.3389/fpls.2019.00941

Boutrot, F., and Zipfel, C. (2017). Function, discovery, and exploitation of plant pattern recognition receptors for broad-spectrum disease resistance. Annual Review of Phytopathology, 55, 257–286. https://doi.org/10.1146/annurev-phyto-080614-120106

Brzezinka, K., Altmann, S., Czesnick, H., Nicolas, P., Gorka, M., Benke, E., Kabelitz, T., Jähne, F., Graf, A., Kappel, C., and Bäurle, I. (2016). *Arabidopsis* FORGETTER1 mediates stress-induced chromatin memory through nucleosome remodeling. ELife, 5, e17061. https://doi.org/10.7554/eLife.17061

Bürger, M., and Chory, J. (2019). Stressed out about hormones: how plants orchestrate immunity.Cell host y microbe. 26(2), 163–172. https://doi.org/10.1016/j.chom.2019.07.006

Carland, F., Fujioka, S., and Nelson, T. (2010). The sterol methyltransferases SMT1, SMT2, and SMT3 influence *Arabidopsis* development through nonbrassinosteroid products. Plant Physiology, 153(2), 741–756. https://doi.org/10.1104/pp.109.152587

Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., and Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: Predictions from modular yeast co-expression networks. BMC Genomics, 7(1), 40. https://doi.org/10.1186/1471-2164-7-40

Castillo-González, C., Liu, X., Huang, C., Zhao, C., Ma, Z., Hu, T., Sun, F., Zhou, Y., Zhou, X., Wang, X.-J., and Zhang, X. (2015). Geminivirus-encoded TrAP suppressor inhibits the histone methyltransferase SUVH4/KYP to counter host defense. ELife, 4, e06671. https://doi.org/10.7554/eLife.06671

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

Childs, K. L., Davidson, R. M., and Buell, C. R. (2011). Gene coexpression network analysis as a source of functional annotation for rice genes. PLOS ONE, 6(7), e22196. https://doi.org/10.1371/journal.pone.0022196

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome Biology, 17(1), 13.https://doi.org/10.1186/s13059-016-0881-8

Coolen, S., Proietti, S., Hickman, R., Davila Olivas, N. H., Huang, P.-P., Van Verk, M. C., Van Pelt, J. A., Wittenberg, A. H. J., De Vos, M., Prins, M., Van Loon, J. J. A., Aarts, M. G. M., Dicke, M., Pieterse, C. M. J., and Van Wees, S. C. M. (2016). Transcriptome dynamics of *Arabidopsis* during sequential biotic and abiotic stresses. The Plant Journal, 86(3), 249–267. https://doi.org/10.1111/tpj.13167

Cooper, J., and Dobson, H. (2007). The benefits of pesticides to mankind and the environment. Crop Protection, 26(9), 1337–1348. https://doi.org/10.1016/j.cropro.2007.03.022

Corchete, L. A., Rojas, E. A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N. C., and Burguillo, F. J. (2020). Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Scientific Reports, 10(1), 1. https://doi.org/10.1038/s41598-020-76881-x

Cummings, K. conceptos de genetica Klug Cummings. Pdfcoffee.Com. Recuperado el 4 de enero de 2023

Crandall, S. G., Gold, K. M., Jiménez-Gasco, M. del M., Filgueiras, C. C., and Willett, D. S. (2020). A multi-omics approach to solving problems in plant disease ecology. PLoS

ONE, 15(9), e0237975. https://doi.org/10.1371/journal.pone.0237975

Cui, W., and Lee, J.-Y. (2016). *Arabidopsis* callose synthases CalS1/8 regulate plasmodesmal permeability during stress. Nature Plants, 2(5), 16034. https://doi.org/10.1038/nplants.2016.34

Damm, U., O'Connell, R. J., Groenewald, J. Z., and Crous, P. W. (2014). The *Colletotrichum* destructivum species complex—Hemibiotrophic pathogens of forage and field crops. Studies in Mycology, 79, 49–84. https://doi.org/10.1016/j.simyco.2014.09.003

Das, S., Meher, P. K., Rai, A., Bhar, L. M., and Mandal, B. N. (2017). Statistical approaches for gene selection, hub gene identification and module interaction in gene co-expression network analysis: An application to aluminum stress in *soybean* (Glycine max L.). PLOS ONE, 12(1), e0169605. https://doi.org/10.1371/journal.pone.0169605

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. Nature Biotechnology, 34(5), 518–524. https://doi.org/10.1038/nbt.3423

Deyholos, M. K., Cavaness, G. F., Hall, B., King, E., Punwani, J., Van Norman, J., and Sieburth, L. E. (2003). VARICOSE, a WD-domain protein, is required for leaf blade development. Development (Cambridge, England), 130(26), 6577–6588. https://doi.org/10.1242/dev.00909

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M., Jaffrézic, F., and on behalf of The French StatOmique Consortium. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics, 14(6), 671–683. https://doi.org/10.1093/bib/bbs046

Dodds, P. N., and Rathjen, J. P. (2010). Plant immunity: Towards an integrated view of plant–pathogen interactions. Nature Reviews Genetics, 11(8), 8. https://doi.org/10.1038/nrg2812

Doehlemann, G., Ökmen, B., Zhu, W., and Sharon, A. (2017). Plant pathogenic fungi. Microbiology Spectrum, 5(1). https://doi.org/10.1128/microbiolspec.FUNK-0023-2016

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. Science, 323(5910), 133–138. https://doi.org/10.1126/science.1162986

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25), 14863–14868. https://doi.org/10.1073/pnas.95.25.14863

Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics, 19(5), 776–792. https://doi.org/10.1093/bib/bbx008

Fernandez-Calvino, L., Faulkner, C., Walshaw, J., Saalbach, G., Bayer, E., Benitez-Alfonso, Y., and Maule, A. (2011). *Arabidopsis* plasmodesmal proteome. *PLOS ONE*, *6*(4), e18880. https://doi.org/10.1371/journal.pone.0018880

Finotello, F., and Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. Briefings in Functional Genomics, 14(2), 130–142. https://doi.org/10.1093/bfgp/elu035

Fondi, M., and Liò, P. (2015). Multi-omics and metabolic modeling pipelines: Challenges and tools for systems microbiology. Microbiological Research, 171, 52–64. https://doi.org/10.1016/j.micres.2015.01.003

Froidure, S., Canonne, J., Daniel, X., Jauneau, A., Brière, C., Roby, D., and Rivas, S. (2010). AtsPLA2-alpha nuclear relocalization by the *Arabidopsis* transcription factor AtMYB30 leads to repression of the plant defense response. Proceedings of the National Academy of Sciences of the United States of America, 107(34), 15281–15286. https://doi.org/10.1073/pnas.1009056107

Fujimoto, M., Ebine, K., Nishimura, K., Tsutsumi, N., and Ueda, T. (2020). Longin R-SNARE is retrieved from the plasma membrane by ANTH domain-containing proteins in *Arabidopsis*. Proceedings of the National Academy of Sciences, 117(40), 25150–25158. https://doi.org/10.1073/pnas.2011152117

Furutani, M., Hirano, Y., Nishimura, T., Nakamura, M., Taniguchi, M., Suzuki, K., Oshida, R., Kondo, C., Sun, S., Kato, K., Fukao, Y., Hakoshima, T., and Morita, M. T. (2020). Polar recruitment of RLD by LAZY1-like protein during gravity signaling in root branch angle control. Nature Communications, 11, 76. https://doi.org/10.1038/s41467-019-13729-7

Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., Law, C., Davis, S., Carey, V., Morgan, M., Zimmer, R., and Waldron, L. (2021). Toward a gold standard for benchmarking gene set enrichment analysis. Briefings in Bioinformatics, 22(1), 545–556. https://doi.org/10.1093/bib/bbz158

Ghazalpour, A., Doss, S., Kang, H., Farber, C., Wen, P.-Z., Brozell, A., Castellanos, R., Eskin, E., Smith, D. J., Drake, T. A., and Lusis, A. J. (2008). High-resolution mapping of gene expression using association in an outbred mouse stock. PLOS Genetics, 4(8), e1000149. https://doi.org/10.1371/journal.pgen.1000149

Gill, H., and Garg, H. (2014). Pesticides: Environmental impacts and management strategies. InTech, 188–230. Godichon-Baggioni, A. (2019). Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. Journal of Statistical Planning and Inference, 203, 1–19. https://doi.org/10.1016/j.jspi.2019.01.001

Góngora-Castillo, E., and Buell, C. R. (2013). Bioinformatics challenges in de novo transcriptome assembly using short read sequences in the absence of a reference genome sequence. Natural Product Reports, 30(4), 490–500. https://doi.org/10.1039/c3np20099j

Goulson, D. (2014). Pesticides linked to bird declines. Nature, 511(7509), 7509. https://doi.org/10.1038/nature13642

Grigoriev, I. V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., Smirnova, T., Nordberg, H., Dubchak, I., and Shabalov, I. (2014). MycoCosm portal: Gearing up for 1000 fungal genomes. Nucleic Acids Research, 42(D1), D699–D704. https://doi.org/10.1093/nar/gkt1183

Han, Y., Gao, S., Muegge, K., Zhang, W., and Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. Bioinformatics and Biology Insights, 9(Suppl 1), 29–46. https://doi.org/10.4137/BBI.S28991

Hawksworth, D. (1991). The fungal dimension of biodiversity: Magnitude, significance, and conservation. *Mycological Research*, *95*, 641–655. https://doi.org/10.1016/S0953-7562(09)80810-1

Hooper, C. M., Castleden, I. R., Tanz, S. K., Aryamanesh, N., and Millar, A. H. (2017). SUBA4: The interactive data analysis centre for *Arabidopsis* subcellular protein locations. *Nucleic Acids Research*, *45*(D1), D1064–D1074. https://doi.org/10.1093/nar/gkw1041

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., Laurance, M. F., Zhao, W., Qi, S., Chen, Z., Lee, Y., Scheck, A. C., Liau, L. M., Wu, H., Geschwind, D. H., Febbo, P. G., Kornblum, H. I., Cloughesy, T. F., Nelson, S. F., and Mischel, P. S. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proceedings of the National Academy of Sciences, 103(46), 17402–17407. https://doi.org/10.1073/pnas.0608396103

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2007). The DAVID gene functional classification tool: A novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biology, 8(9), R183. https://doi.org/10.1186/gb-2007-8-9-r183

Hughes, M. (2015). Systems biology tools for integrated omics analysis. Genetic Engineering and Biotechnology News, 35(3), 18–19. https://doi.org/10.1089/gen.35.03.11

Iancu, O. D., Kawane, S., Bottomly, D., Searles, R., Hitzemann, R., and McWeeney, S. (2012). Utilizing RNA-Seq data for de novo coexpression network inference. Bioinformatics, 28(12), 1592–1597. https://doi.org/10.1093/bioinformatics/bts245

Ibrahim, H. M. M., Kusch, S., Didelon, M., and Raffaele, S. (2021). Genome-wide alternative splicing profiling in the fungal plant pathogen *Sclerotinia sclerotiorum* during the

colonization of diverse host families. Molecular Plant Pathology, 22(1), 31–47. https://doi.org/10.1111/mpp.13006

Jamil, I. N., Remali, J., Azizan, K. A., Nor Muhammad, N. A., Arita, M., Goh, H.-H., and Aizat, W. M. (2020). Systematic multi-omics integration (MOI) approach in plant systems biology. Frontiers in Plant Science, 11, 944. https://doi.org/10.3389/fpls.2020.00944

Jones, J. D. G., and Dangl, J. L. (2006). The plant immune system. Nature, 444(7117), 7117. https://doi.org/10.1038/nature05286

Kakei, Y., and Shimada, Y. (2015). AtCAST3.0 update: A web-based tool for analysis of transcriptome data by searching similarities in gene expression profiles. Plant and Cell Physiology, 56(1), e7. https://doi.org/10.1093/pcp/pcu174

Kchouk, M., Gibrat, J.-F., and Elloumi, M. (2017). Generations of sequencing technologies: From first to next generation. Biology and Medicine, 09. https://doi.org/10.4172/0974-8369.1000395

Kim, B., Kim, G., Fujioka, S., Takatsuto, S., and Choe, and S. (2012). Overexpression of *3β-hydroxysteroid dehydrogenases/c-4 decarboxylases* causes growth defects possibly due to abnormal auxin transport in *Arabidopsis*. Molecules and Cells, 34(1), 77–84. https://doi.org/10.1007/s10059-012-0102-6

Kim, K.-H., Kabir, E., and Jahan, S. A. (2017). Exposure to pesticides and the associated human health effects. The Science of the Total Environment, 575, 525–535. https://doi.org/10.1016/j.scitotenv.2016.09.009

Kolenc, Ž., Pirih, N., Gretic, P., and Kunej, T. (2021). Top trends in multiomics research: evaluation of 52 published studies and new ways of thinking terminology and visual displays. OMICS: A Journal of Integrative Biology. https://doi.org/10.1089/omi.2021.0160

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., and Huala, E. (2012). The *Arabidopsis* information resource (TAIR): Improved gene annotation and new tools. Nucleic Acids Research, 40(D1), D1202–D1210. https://doi.org/10.1093/nar/gkr1090

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-55. Recuperado el 4 de enero de 2023

Larson, M. G. (2006). Descriptive statistics and graphical displays. Circulation, 114(1), 76–81. https://doi.org/10.1161/CIRCULATIONAHA.105.584474

Lee, T., Yang, S., Kim, E., Ko, Y., Hwang, S., Shin, J., Shim, J. E., Shim, H., Kim, H., Kim, C., and Lee, I. (2015). AraNet v2: An improved database of co-functional gene networks for

the study of *Arabidopsis thaliana* and 27 other nonmodel plant species. Nucleic Acids Research, 43(D1), D996–D1002. https://doi.org/10.1093/nar/gku1053

Leek, J. T. (2014). svaseq: Removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Research, 42(21), e161. https://doi.org/10.1093/nar/gku864

Leinonen, R., Sugawara, H., Shumway, M., and on behalf of the International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. Nucleic Acids Research, 39(suppl_1), D19–D21. https://doi.org/10.1093/nar/gkq1019

Liang, S., Ma, A., Yang, S., Wang, Y., and Ma, Q. (2018). A review of matched-pairs feature selection methods for gene expression data analysis. Computational and Structural Biotechnology Journal, 16, 88–97. https://doi.org/10.1016/j.csbj.2018.02.005

Liang, X., Shang, S., Dong, Q., Wang, B., Zhang, R., Gleason, M. L., and Sun, G. (2018). Transcriptomic analysis reveals candidate genes regulating development and host interactions of *Colletotrichum fructicola*. BMC Genomics, 19(1), 557. https://doi.org/10.1186/s12864-018-4934-0

Lin, C.-T., Xu, T., Xing, S.-L., Zhao, L., Sun, R.-Z., Liu, Y., Moore, J. P., and Deng, X. (2019). Weighted gene co-expression network analysis (WGCNA) reveals the hub role of protein ubiquitination in the acquisition of desiccation tolerance in *boea hygrometrica*. Plant and Cell Physiology, 60(12), 2707–2719. https://doi.org/10.1093/pcp/pcz160

Luhua, S., Hegie, A., Suzuki, N., Shulaev, E., Luo, X., Cenariu, D., Ma, V., Kao, S., Lim, J., Gunay, M. B., Oosumi, T., Lee, S. C., Harper, J., Cushman, J., Gollery, M., Girke, T., Bailey-Serres, J., Stevenson, R. A., Zhu, J.-K., and Mittler, R. (2013). Linking genes of unknown function with abiotic stress responses by high-throughput phenotype screening. Physiologia Plantarum, 148(3), 322–333. https://doi.org/10.1111/ppl.12013

Macho, A. P., and Zipfel, C. (2014). Plant PRRs and the activation of innate immune signaling. Molecular Cell, 54(2), 263–272. https://doi.org/10.1016/j.molcel.2014.03.028

Maza, E., Frasse, P., Senin, P., Bouzayen, M., and Zouine, M. (2013). Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments. Communicative and Integrative Biology, 6(6), e25849. https://doi.org/10.4161/cib.25849

Niñoles, R., Rubio, L., García-Sánchez, M. J., Fernández, J. A., Bueso, E., Alejandro, S., and Serrano, R. (2013). A dominant-negative form of *Arabidopsis AP-3 β-adaptin* improves intracellular pH homeostasis. The Plant Journal: For Cell and Molecular Biology, 74(4), 557–568. https://doi.org/10.1111/tpj.12138

Obayashi, T., Hibara, H., Kagaya, Y., Aoki, Y., and Kinoshita, K. (2022). ATTED-II v11: A plant gene coexpression database using a sample balancing technique by subagging of

principal components. Plant and Cell Physiology, 63(6), 869–881. https://doi.org/10.1093/pcp/pcac041

O'Connell, R. J., Thon, M. R., Hacquard, S., Amyotte, S. G., Kleemann, J., Torres, M. F., Damm, U., Buiate, E. A., Epstein, L., Alkan, N., Altmüller, J., Alvarado-Balderrama, L., Bauser, C. A., Becker, C., Birren, B. W., Chen, Z., Choi, J., Crouch, J. A., Duvick, J. P., Vaillancourt, L. J. (2012). Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. Nature Genetics, 44(9), 1060–1065. https://doi.org/10.1038/ng.2372

Oh, S., Geistlinger, L., Ramos, M., Blankenberg, D., van den Beek, M., Taroni, J. N., Carey, V. J., Greene, C. S., Waldron, L., and Davis, S. (2022). GenomicSuperSignature facilitates interpretation of RNA-seq experiments through robust, efficient comparison to public databases. Nature Communications, 13(1), 1. https://doi.org/10.1038/s41467-022-31411-3

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

Pien, S., Fleury, D., Mylne, J. S., Crevillen, P., Inzé, D., Avramova, Z., Dean, C., and Grossniklaus, U. (2008). *Arabidopsis Trithorax1* dynamically regulates flowering locus C activation via histone 3 *Lysine* 4 *Trimethylation*. The Plant Cell, 20(3), 580–588. https://doi.org/10.1105/tpc.108.058172

Provart, N. J., Alonso, J., Assmann, S. M., Bergmann, D., Brady, S. M., Brkljacic, J., Browse, J., Chapple, C., Colot, V., Cutler, S., Dangl, J., Ehrhardt, D., Friesner, J. D., Frommer, W. B., Grotewold, E., Meyerowitz, E., Nemhauser, J., Nordborg, M., Pikaard, C., … McCourt, P. (2016). 50 years of *Arabidopsis* research: Highlights and future directions. New Phytologist, 209(3), 921–944. https://doi.org/10.1111/nph.13687

Pukelsheim, F. (1994). The three *sigma* rule. The American Statistician, 48(2), 88–91. https://doi.org/10.2307/2684253

Qi, H., Jiang, Z., Zhang, K., Yang, S., He, F., and Zhang, Z. (2018). PlaD: A transcriptomics database for plant defense responses to pathogens, providing new insights into plant immune system. Genomics, Proteomics and Bioinformatics, 16(4), 283–293. https://doi.org/10.1016/j.gpb.2018.08.002

Ranf, S. (2018). Pattern recognition receptors—versatile genetic tools for engineering broad-spectrum disease resistance in crops. Agronomy, 8(8), 8. https://doi.org/10.3390/agronomy8080134

Renesh, B. (2020). reneshbedre/bioinfokit: Bioinformatics data analysis and visualization toolkit | Zenodo. https://zenodo.org/record/3965241

Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. Molecular Cell, 58(4), 586–597. https://doi.org/10.1016/j.molcel.2015.05.004

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nature biotechnology, 32(9), 896–902. https://doi.org/10.1038/nbt.2931

Rodrigo, G., Carrera, J., Ruiz-Ferrer, V., Toro, F. J. del, Llave, C., Voinnet, O., and Elena, S. F. (2012). A meta-analysis reveals the commonalities and differences in *Arabidopsis thaliana* response to different viral pathogens. PLOS ONE, 7(7), e40526. https://doi.org/10.1371/journal.pone.0040526

Russo, P. S. T., Ferreira, G. R., Cardozo, L. E., Bürger, M. C., Arias-Carrasco, R., Maruyama, S. R., Hirata, T. D. C., Lima, D. S., Passos, F. M., Fukutani, K. F., Lever, M., Silva, J. S., Maracaja-Coutinho, V., and Nakaya, H. I. (2018). CEMiTool: A Bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinformatics, 19(1), 56. https://doi.org/10.1186/s12859-018-2053-1

Saijo, Y., Loo, E. P., and Yasuda, S. (2018). Pattern recognition receptors and signaling in plant–microbe interactions. The Plant Journal, 93(4), 592–613. https://doi.org/10.1111/tpj.13808

Sanchez-Bayo, F., and Goka, K. (2014). Pesticide residues and bees – a risk assessment. PLoS ONE, 9(4), e94482. https://doi.org/10.1371/journal.pone.0094482

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., and Nelson, A. (2019). The global burden of pathogens and pests on major food crops. Nature Ecology and Evolution, 3(3), 3. https://doi.org/10.1038/s41559-018-0793-y

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., … Sherry, S. T. (2020). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 49(D1), D10–D17. https://doi.org/10.1093/nar/gkaa892

Sela, I., Wolf, Y. I., and Koonin, E. V. (2016). Theory of prokaryotic genome evolution. Proceedings of the National Academy of Sciences, 113(41), 11399–11407. https://doi.org/10.1073/pnas.1614083113

Shahan, R., Zawora, C., Wight, H., Sittmann, J., Wang, W., Mount, S. M., and Liu, Z. (2018). Consensus coexpression network analysis identifies key regulators of flower and fruit development in wild strawberry. Plant Physiology, 178(1), 202–216. https://doi.org/10.1104/pp.18.00086

Shaik, R., and Ramakrishna, W. (2013). Genes and co-expression modules common to drought and bacterial stress responses in arabidopsis and rice. PLoS ONE, 8(10), e77261. https://doi.org/10.1371/journal.pone.0077261

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science, 302(5643), 249–255. https://doi.org/10.1126/science.1087447

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. Bioinformatics and Biology Insights, 14. https://doi.org/10.1177/1177932219899051

Sutherst, R. W., Baker, R. H. A., Coakley, S. M., Harrington, R., Kriticos, D. J., and Scherm, H. (2007). Pests under global change—meeting your future landlords? En J. G. Canadell, D. E. Pataki, and L. F. Pitelka (Eds.), Terrestrial Ecosystems in a Changing World (pp.211–226). Springer. https://doi.org/10.1007/978-3-540-32730-1_17

Tan, X., Wei, J., Li, B., Wang, M., and Bao, Y. (2017). AtVps11 is essential for vacuole biogenesis in embryo and participates in pollen tube growth in arabidopsis. Biochemical and Biophysical Research Communications, 491(3), 794–799. https://doi.org/10.1016/j.bbrc.2017.07.059

Thomas, J., Kim, H. R., Rahmatallah, Y., Wiggins, G., Yang, Q., Singh, R., Glazko, G., and Mukherjee, A. (2019). RNA-seq reveals differentially expressed genes in rice (*Oryza sativa*) roots during interactions with plant-growth promoting bacteria, *Azospirillum brasilense*. PLOS ONE, 14(5), e0217309. https://doi.org/10.1371/journal.pone.0217309

Trempel, F., Kajiura, H., Ranf, S., Grimmer, J., Westphal, L., Zipfel, C., Scheel, D., Fujiyama, K., and Lee, J. (2016). Altered glycosylation of exported proteins, including surface immune receptors, compromises calcium and downstream signaling responses to microbe-associated molecular patterns in *Arabidopsis thaliana*. BMC Plant Biology, 16, 31. https://doi.org/10.1186/s12870-016-0718-3

Tsushima, A., Gan, P., Kumakura, N., Narusaka, M., Takano, Y., Narusaka, Y., and Shirasu, K.(2019). Genomic plasticity mediated by transposable elements in the plant pathogenic fungus *Colletotrichum higginsianum*. Genome biology and evolution, 11, 1487–1500. https://doi.org/10.1093/gbe/evz087

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. Briefings in Bioinformatics, 19(4), 575–592. https://doi.org/10.1093/bib/bbw139

Van Someren, E. J. W. (2006). Mechanisms and functions of coupling between sleep and temperature rhythms. En Progress in Brain Research (Vol. 153, pp. 309–324). Elsevier. https://doi.org/10.1016/S0079-6123(06)53018-3

Wang, C., Liu, W., Wang, G., Li, J., Dong, L., Han, L., Wang, Q., Tian, J., Yu, Y., Gao, C., and Kong, Z. (2017). KTN80 confers precision to microtubule severing by specific targeting of katanin complexes in plant cells. The EMBO Journal, 36(23), 3435–3447. https://doi.org/10.15252/embj.201796823

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews. Genetics, 10(1), 57–63. https://doi.org/10.1038/nrg2484

Wu, L. F., Hughes, T. R., Davierwala, A. P., Robinson, M. D., Stoughton, R., and Altschuler, S. J. (2002). Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. Nature Genetics, 31(3), 3. https://doi.org/10.1038/ng906

Wu, M., Yi, H., and Ma, S. (2021). Vertical integration methods for gene expression data analysis. Briefings in Bioinformatics, 22(3). https://doi.org/10.1093/bib/bbaa169

Xu, D., and Tian, Y. (2015). A Comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, 4, Article17. https://doi.org/10.2202/1544-6115.1128

Zhang, C., Zhang, B., Lin, L.-L., and Zhao, S. (2017). Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics, 18(1), 583. https://doi.org/10.1186/s12864-017-4002-1

Zhang, C., Zhang, B., Vincent, M. S., and Zhao, S. (2016). Bioinformatics tools for RNA-seq gene and isoform quantification. Journal of Next Generation Sequencing and Applications, 03(03). https://doi.org/10.4172/2469-9853.1000140

Zhang, J., Li, W., Xiang, T., Liu, Z., Laluk, K., Ding, X., Zou, Y., Gao, M., Zhang, X., Chen, S., Mengiste, T., Zhang, Y., and Zhou, J.-M. (2010). Receptor-like cytoplasmic kinases integrate signaling from multiple plant immune receptors and are targeted by a *Pseudomonas syringae* effector. Cell Host and Microbe, 7(4), 290–301. https://doi.org/10.1016/j.chom.2010.03.007

Zhang, L., Huang, X., He, C., Zhang, Q.-Y., Zou, X., Duan, K., and Gao, Q. (2018). Novel Fungal pathogenicity and leaf defense strategies are revealed by simultaneous transcriptome analysis of *Colletotrichum fructicola* and strawberry infected by this fungus. Frontiers in Plant Science, 9. https://www.frontiersin.org/article/10.3389/fpls.2018.00434

Zhang, Y., Parmigiani, G., and Johnson, W. E. (2020). ComBat-seq: Batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics, 2(3), lqaa078.

https://doi.org/10.1093/nargab/lqaa078

Zhao, S., Ye, Z., and Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA (New York, N.Y.), 26(8), 903–909. https://doi.org/10.1261/rna.074922.120

Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., and McShane, L. M. (2021). TPM, FPKM, or normalized counts? a comparative study of quantification measures for the analysis of RNA-seq data from the nci patient-derived models repository. Journal of Translational Medicine, 19(1), 269. https://doi.org/10.1186/s12967-021-02936-w

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Górczak, K., Klamecka, K., Figlerowicz, M., and Siatkowski, I. (2015). The impact of normalization methods on RNA-seq data analysis. BioMed Research International, 2015, e621690. https://doi.org/10.1155/2015/621690