UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

GRADO EN BIOTECNOLOGÍA

Búsqueda de motivos genómicos con posible relevancia biológica en tres aislados de hongos del género *Plectosphaerella*

TRABAJO DE FIN DE GRADO

Autor: Andrea Álvarez Pérez

Tutores: Soledad Sacristán Benayas, Julio Luis Rodríguez Romero

UNIVERSIDAD POLITÉCNICA DE MADRID

Escuela Técnica Superior De

Ingeniería Agronómica, Alimentaria y de Biosistemas

GRADO EN BIOTECNOLOGÍA

Búsqueda de motivos genómicos con posible relevancia biológica en tres aislados de hongos del género *Plectosphaerella*

TRABAJO DE FIN DE GRADO

Andrea Álvarez Pérez

Madrid, 2021

Tutores: Soledad Sacristán Benayas, Julio Luis Rodríguez Romero

Departamento de Biotecnología-Biología Vegetal

TITULO DEL TFG- Búsqueda de motivos genómicos con posible relevancia biológica en tres aislados de hongos del género *Plectosphaerella*

Memoria presentada por Andrea Álvarez Pérez para la obtención del título de Graduado en Biotecnología por la Universidad Politécnica de Madrid

Fdo: Andrea Álvarez Pérez

VºBº Tutor y Director del TFG

Soledad Sacristán Benayas

Profesora Titular de Universidad

Dpto de Biotecnología-Biología Vegetal

ETSIAAB – Universidad Politécnica de Madrid

VºBº Cotutor

Julio Luis Rodríguez Romero

Profesor Contratado Doctor I3

Dpto de Biotecnología-Biología Vegetal

ETSIAAB – Universidad Politécnica de Madrid

Madrid, 24, Junio, 2021

ÍNDICE

ÍNDICE TABLAS

ÍNDICE FIGURAS

ABREVIATURAS

CAZymas…… *Carbohidrate Active enZymes*

CHV1…… *Cryphonectria hypovirus 1*

Col-0…….genotipo silvestre

CThTV…… *Curvularia* thermal tolerance *virus*

Cyp79b2b3 ……. doble mutante *cyp79b2 cyp79b3*

DEGs…… differentially expressed genes

DNA …… deoxyribonucleic acid

FPKM …… Fragments per kilobase of exon per million

hpi……. horas post inoculación

MAPK……. Mitogen-Activated Protein Kinases

NCBI…… National Center for Biotechnology Information

P0831……aislado epífito de *Plectosphaerella*

PAMPs……....Pathogen-associated molecular pattern

PcBMM……. aislado patogénico de *Plectosphaerella*

Pc2127 …… aislado no patogénico de *Plectosphaerella*

PPR……. Peptide Pattern Recognition

PRR……Pattern-recognition receptors

PTI…… PAMP-triggered immunity

RdRP ……. RNA Polimerasa dependiente de RNA

RNA……ribonucleic acid

RNAseq…… tecnología de secuenciación de RNA

ROS …… Especies Reactivas del Oxígeno

SRA……... Sequence Read Archive

SSP…… Small Secreted Peptides

TAIR……. The Arabidopsis Information Resource

TRV……. *Turnip Ringspot Virus*

ABSTRACT

This study aims to determine possible sequences in the genomes of three isolates of the fungus *Plectosphaerella spp* that may have biological relevance in the fungus-plant interaction. These sequences may give us information about the characteristics and lifestyle of each one. The isolates we worked with were the pathogenic PcBMM, the non-pathogenic Pc2127 and the epiphyte P0831. In particular, we focused on the following sequences:

1. Mycovirus. The presence of mycovirus may affect the virulence and lifestyle of the fungal host. In this study, a bioinformatic processing was performed to search for mycovirus in the transcriptome data of the three *Plectosphaerella* isolates, using similar data of the fungus *Colletotrichum tofieldiae* as a positive control. In addition, this method was contrasted with the one used by Gilbert, K., *et al* (2019), and the differences between both were evaluated in order to propose a strategy that was as reliable and precise as possible for the detection of mycovirus in transcriptomes.

2. Sequences candidate as mimetics of two groups of Arabidopsis phytoregulatory peptides. These peptides may influence the pathogenic or non-pathogenic nature of the fungus, being able to participate in various defense mechanisms of the plant. We have focused on the SCOOP family, from which mimetic peptides have been previously identified in other fungi, such as *Fusarium spp*, and contain common motifs responsible for triggering the immune response. Also, we have looked for common motifs with other putative SSP peptides of Arabidopsis. By setting various criteria, we have proposed certain candidate genes that could mimic the activity of these peptides in Arabidopsis.

3. *Carbohidrate Active Enzymes* (CAZymes). CAZymes may be important molecules in the interaction of fungi with plants. *Plectosphaerella* CAZymes were analyzed using two bioinformatic tools, CUPP and dbCAN2 and divided into six groups, including all five families and the presence or not of CBM modules. Then, the resulting CAZymes were classified as secreted or not by the tool SECRETOOL. Finally, we have done a phylogenetic analysis of the total and secreted CAZymes present in the genomes of the three *Plectosphaerella* isolates. This analysis is a first approach to study whether there is a relationship between the abundance of CAZymas with extracellular signal peptide and the lifestyle of each *Plectosphaerella* isolate. Although there are no big differences in the general data, there are some instances different between the three isolates that could be further studied.

1. CAPÍTULO 1: INTRODUCCIÓN Y OBJETIVOS

1.1 Objeto de estudio: *Plectosphaerella sp*.

El objeto de estudio de este trabajo son hongos del género *Plectosphaerella*, perteneciente a la división *Ascomycota*, clase *Sordariomycetes*. Se trata de hongos filamentosos capaces de crecer en ambientes muy diversos, como saprófitos o parásitos necrótrofos de diferentes huéspedes animales y vegetales, entre los que se encuentra *Arabidopsis thaliana* [1]. Este trabajo profundiza en el estudio de tres aislados cuya caracterización se ha venido llevando a cabo en trabajos anteriores, y que poseen distintas capacidades que les permiten la colonización del tejido vegetal o el establecimiento de interacciones epífitas [1 - 3]. El aislado PcBMM de la especie *Plectosphaerella cucumerina* (Lindf.) es patogénico tanto en plantas de Arabidopsis de genotipo silvestre (Col-0) como en el mutante inmunodeprimido de *cyp79b2b3*, defectivo en la síntesis de metabolitos secundarios de defensa derivados del triptófano [2]. El aislado Pc2127, también de la especie *P. cucumerina*, es una variante no patogénica en Col-0 pero capaz de infectar al mutante inmunodeprimido *cyp79b2b3*. Por último, el aislado P0831, cuya especie dentro del género *Plectosphaerella* aún no ha sido determinada, coloniza de manera epífita tanto las plantas de Arabidopsis Col-0 como las del mutante inmunodeprimido *cyp79b2b3*. En el estudio llevado a cabo por Muñoz-Barrios, A., *et al* (2020) [1], se describió la secuenciación y la anotación de los genomas de los tres aislados con el objetivo de encontrar determinantes moleculares de su estilo de vida y patogénesis. Además, en este trabajo se realizó un perfil de expresión de RNAseq de todo el genoma de los tres aislados de *Plectosphaerella* en condiciones *in vitro* e *in planta*, en plantas de genotipo Col-0 a 10, 16 y 24 hpi (horas post inoculación) y del mutante *cyp79b2b3* a 10 y 16 hpi. Se encontró que los genomas de los tres hongos eran muy similares, estando las mayores diferencias en las respuestas transcripcionales de cada aislado tras la inoculación en planta y siendo llamativa la baja cantidad de genes expresados por P0831 en comparación con los otros dos aislados. El estudio de genes de expresión diferencial (DEGs) mostró que durante la colonización de la planta se van activando clústeres de genes que podrían estar determinando el carácter epífito o patogénico del aislado [1].

El presente trabajo pretende profundizar en el análisis de los genomas y transcriptomas de estos aislados, realizando una búsqueda de determinadas secuencias que pueden determinar el tipo de interacción que establecen con la planta.

1.2. Micovirus

Los micovirus, también conocidos como virus fúngicos, son virus que infectan a hongos. No fueron descubiertos hasta la década de 1960, en *Agaricus bisporus,* hongo basidiomiceto de la familia *Agaricales* también conocido como "champiñón común" [4]. Este descubrimiento tan tardío se debió principalmente a que la mayoría de los hongos infectados con micovirus no exhiben ninguna característica de una infección viral "típica", como la lisis celular o la transmisión de enfermedades extracelulares. La mayoría de los micovirus descritos tienen genomas de RNA, ya sea bicatenario (dsRNA) o monocatenario (ssRNA), sobre todo de sentido positivo (+ssRNA). Se cree que los micovirus no son infecciosos como partículas libres, ya que carecen de ruta extracelular de infección y se transmiten intracelularmente durante la división celular, la esporogénesis y la fusión celular del hongo huésped [5]. Aunque se han descubierto micovirus en los principales filos de hongos, se estima que la mayoría de ellos todavía están por descubrir [4].

Muchos micovirus pueden tener un impacto significativo en el hongo, y sus efectos varían desde causar un crecimiento irregular o pigmentación anormal, hasta cambiar la reproducción sexual del huésped. Por otra parte, otros muchos producen virulencia reducida o hipovirulencia en hongos patógenos de plantas. A este respecto, el ejemplo más exitoso es el de la enfermedad del chancro del castaño causada por el hongo *Cryphonectria parasítica,* que se ha controlado con éxito en Europa durante más de 40 años utilizando aislados de *C. parasitica* infectados con *Cryphonectria hypovirus 1* (CHV1) [6]. Recientemente se ha reportado que la infección por el micovirus SsHADV-1 convierte al hongo fitopatógeno *Sclerotinia sclerotiorum* en un endófito asintomático [7], y es también muy conocido el caso del micovirus CThTV, necesario para que el hongo endófito *Curvularia protuberata* establezca una interacción mutualista que protege a la planta huésped de temperaturas extremas [8]. Por tanto, los micovirus pueden modular o hasta determinar el estilo de vida de los hongos que los hospedan.

El efecto de los micovirus sobre la virulencia puede estar relacionado con la modulación de la expresión de genes del hongo hospedador. Estudios de micovirus de *C. parasitica*, *Fusarium graminearum* y *S. sclerotiorum* mostraron que el nivel de expresión de genes fúngicos difería entre los aislados fúngicos infectados y los libres de virus. También se ha visto que el micovirus SsHADV-1 reprime la expresión de genes claves para la patogenicidad de *S. sclerotiorum* [7]. Por tanto, sería posible que las diferencias en la patogénesis y la expresión de genes encontradas entre los aislados de *Plectosphaerella* objeto de este estudio se debieran a la presencia de micovirus. Por el momento no hay estudios publicados sobre la existencia de micovirus en *Plectosphaerella,* y la posible presencia de micovirus en este hongo, su caracterización y análisis podría ser posteriormente utilizado como estrategia de biocontrol.

Actualmente, se han utilizado enfoques transcriptómicos para la identificación y detección de micovirus [9 - 10]. El método seguido en este trabajo para la búsqueda de micovirus consiste en una adaptación del análisis utilizado en el estudio de Ruiz-Padilla, A., *et al* (2021) [10] en el hongo patógeno *Botrytis cinerea*, y se basa de un procesado bioinformático de muestras de RNAseq. La ventaja de este método es que permite identificar de una gran variedad de micovirus de distintos tipos de genoma, incluyendo virus de RNA y de DNA.

1.3. Péptidos fitorreguladores

Los SSP (*Small Secreted Peptides*) son pequeños péptidos secretados que se traducen directamente como un péptido final o a partir de un precursor de proteína inactivo que debe procesarse para tener actividad [11]. En los últimos años se ha puesto de manifiesto el papel de los SSPs como componentes importantes en varios procesos de las plantas, desde la regulación de la diferenciación y el desarrollo celular, hasta el crecimiento del tubo polínico, embriogénesis y la regulación de funciones fisiológicas. Su actividad biológica se observa a concentraciones muy bajas, por lo que se puede definir como de naturaleza hormonal [11]. Se ha demostrado que los SSPs también pueden estar implicados en los mecanismos de defensa de las plantas [12 - 13].

Como parte de sus mecanismos de defensa, las plantas tienen la capacidad de reconocer pequeñas secuencias de moléculas que se repiten en grupos de patógenos, llamados PAMPs (*Pathogen-Associated Molecular Pattern*) a través de receptores PRR (*Pattern-Recognition Receptors)* de su superficie en lo que se denomina PTI (*PAMP-Triggered* Inmunity). Las respuestas intracelulares asociadas a la PTI incluyen las cascadas de MAPK (*Mitogen-Activated Protein Kinase*), la producción de especies reactivas de oxígeno (ROS), la producción de etileno o el aumento de calcio citosólico. Sin embargo, esta respuesta puede ser suprimida por los patógenos mediante efectores proteicos que mimetizan o interfieren con funciones celulares de la planta para facilitar la colonización [14]. Evidencias recientes apuntan a la posibilidad de que los patógenos puedan generar péptidos que actúen como efectores, mimetizando la función de los SSPs en las plantas [15]. Por otro lado, las plantas pueden tener receptores que detecten estos péptidos y desencadenen reacciones de defensa [14]. Las herramientas bioinformáticas pueden ayudar a identificar estos péptidos miméticos en los genomas de los patógenos y desentrañar su función en patogénesis [15]. Es posible que la existencia de estos péptidos en los genomas de los aislados de *Plectosphaerella* determine su patogénesis.

1.3.1. Péptidos SCOOP

Gracias al análisis de transcriptomas y a las predicciones bioinformáticas, se identificó en Arabidopsis la familia de péptidos SCOOP. La caracterización funcional de uno de sus miembros, PROSCOOP12, mostró que este pequeño gen podría actuar como moderador en la respuesta a diferentes agresiones de patógenos y en el desarrollo radicular mediante el control de la producción de especies reactivas de oxígeno (ROS) [12]. Recientemente se ha descubierto que

los péptidos SCOOP son el ligando del receptor MIK2, que juega un papel fundamental en la respuesta a diversos estreses ambientales, incluida la relacionada con la integridad de la pared celular, la tolerancia al estrés salino y la resistencia al patógeno fúngico del suelo *Fusarium oxysporum* [16 - 17]. El reconocimiento de los péptidos SCOOP por MIK2 elicita reacciones típicas de la respuesta PTI [17].

Se han encontrado motivos característicos de la familia SCOOP (SCOOP-LIKE) en distintas especies de hongos [16 - 17]. Estos motivos SCOOP-LIKE activan las respuestas inmunitarias dependientes de MIK2-BAK1 / SERK4, lo que apunta al papel de MIK2 en la percepción de patógenos. En el trabajo de Rhodes, J., *et al.*, 2021 [17], han encontrado motivos SCOOP-LIKE en distintas formas especiales de *F. oxysporum* y en otras especies de hongos fitopatógenos, como *F. langsethiae, Trichoderma atroviridis, Verticillium dahliae* y *Magnaporthe oryzae*, e incluso en el hongo no fitopatógeno *Neurospora crassa*, lo que sugiere que estos motivos están altamente conservados dentro de la división Ascomycota. Por otra parte, en el trabajo posterior de Hou, S. *et al*. (2021) [16] han encontrado motivos SCOOP-LIKE en una amplia gama de patógenos del género *Fusarium*. Este gran grupo de 22 péptidos puede dividirse en siete subgrupos, donde el motivo conservado SXS es el responsable de la actividad del péptido. En *F. graminearum*, dicho motivo está localizado en el extremo N-terminal de una proteína no caracterizada que contiene el dominio de unión GAL4 al DNA [16]. El género *Plectospharella* pertenece a la misma familia que *Verticillium*, y ambos son muy cercanos a *Fusarium* [1]. Por tanto, es muy probable que haya motivos SCOOP-LIKE por descubrir en los genomas de los aislados objeto de este estudio.

1.3.2. Otros putativos SSPs en Arabidopsis

En trabajos anteriores del grupo de inmunidad innata de las plantas y resistencia a hongos necrótrofos, durante la caracterización del gen YODA (YDA), se descubrieron una serie de putativos SSPs que podrían estar implicados en reacciones de defensa frente a patógenos [13]. El gen YDA es una MAPK que regula varios procesos de desarrollo en Arabidopsis y también regula la respuesta inmune [13]. Los mutantes defectivos *yda* tienen comprometida su resistencia a patógenos, mientras que las plantas con expresión constitutiva de YDA (plantas CA-YDA) muestran resistencia de amplio espectro a hongos, bacterias y oomicetos. En estas plantas CA-YDA se sobreexpresa constitutivamente una serie de genes, entre los cuales se encuentran un conjunto de putativos SSPs. Plantas defectivas en la expresión de dos de estos SSPs muestran una mayor susceptibilidad que las de genotipo silvestre (Col-0) a *P. cucumerina* PcBMM [13]. Se desconoce la función de la mayoría de estos SSPs, que son descritos como proteínas no caracterizadas en las bases de datos principales como el NCBI o TAIR, aunque varios de ellos se clasifican como proteínas transmembrana. Tampoco se conoce cuáles son realmente sus secuencias maduras ni los residuos biológicamente activos. Ya que se ha visto que dos de estos

SSPs están implicados en la defensa frente a *Plectosphaerella*, es posible que existan miméticos de estos SSPs en los aislados objeto de este estudio.

1.4. CAZymas

Las CAZymas (*Carbohidrate Active enZymes*) son proteínas que comprenden diferentes actividades catalíticas en carbohidratos, participando en su síntesis, metabolismo y en el reconocimiento de complejos. Debido a su función, normalmente tienen una alta especificidad. Las CAZymas han evolucionado a partir de un número limitado de antecesores adquiriendo nuevas especificidades a nivel de sustrato y producto. Esta variedad tan vertiginosa de sustratos y enzimas convierte a las CAZymas en un tema particularmente desafiante para la caracterización experimental y para la anotación funcional en genomas [18].

Desde la década de 1990 se han definido más de 360 familias de CAZymas y se han clasificado en cinco grupos principales: glucosiltransferasas [GT], glucósido hidrolasas [GH], polisacárido liasas [PL], esterasas de carbohidratos [CE] y enzimas para las actividades auxiliares [AA] [19]. Los diferentes grupos se han definido siguiendo criterios de clasificación como la identidad de secuencia de aminoácidos y el plegamiento tridimensional [19]. Cualquiera de estos grupos puede incluir dominios de unión a carbohidratos [CBM], módulos no catalíticos asociados con enzimas activas en la hidrólisis de la pared celular. La presencia de CBMs también constituye otro método de clasificación de CAZymas.

Los hongos secretan una serie muy amplia de CAZymas, lo que refleja la utilización de sustratos especializados relacionados con el hábitat. A pesar de su importancia, la presencia de CAZymas no se utiliza como criterio taxonómico, ya que no se ha establecido una relación general entre los perfiles de CAZymas y la filogenia de los hongos. Varios estudios apoyan la idea de que exista una relación entre la cantidad de CAZymas y el tipo de colonización de la planta por parte del hongo. En su estudio, Hacquard, S., *et al* (2016) [20] demostró que la inducción de CAZymas que actúan sobre hemicelulosa y celulosa durante el periodo de colonización de la planta es el doble que en las condiciones *in vitro*, por lo que juegan un papel fundamental en la penetración y acomodación del hongo en la planta. Por otro lado, el estudio de Muñoz-Barrios, A., *et al*, (2020) [1] reveló que las CAZymas secretadas eran 5 veces más abundantes en el aislado patogénico PcBMM de *Plectosphaerella*, lo que podría indicar una relación entre la cantidad de CAZymas y la patogenicidad del hongo. En este aislado también destaca la abundancia de los módulos CWDE (*Cell Wall-Degradating Enzymes*), encargados de degradar los polímeros de la pared celular vegetal, que es superior que en los otras dos aislados [1]. Es posible, por tanto, que la presencia de determinadas CAZYmas se pueda relacionar con el estilo de vida y patogénesis de los distintos aislados objeto de estudio.

1.5. Objetivos

Este trabajo está orientado a la utilización de herramientas bioinformáticas para la búsqueda y análisis de determinadas secuencias que puedan tener relevancia biológica en la interacción de distintos aislados de hongos del género *Plectosphaerella* con la planta. En concreto, se han comparado los genomas, transcriptomas y proteomas de los aislados PcBMM, Pc2127 y P0831, con diferentes estilos de vida en la interacción con la planta *A. thaliana* (patogénico y no patogénico). Los objetivos concretos del trabajo son:

1. Buscar posibles secuencias de micovirus presentes en el transcriptoma del hongo. La hipótesis planteada es que la presencia de micovirus pueda modular la expresión génica y el tipo de interacción planta-hongo.
2. Determinar la presencia de posibles motivos miméticos de péptidos fitorreguladores. La hipótesis que se plantea en este trabajo es que estos candidatos puedan actuar como efectores e incluso que la planta sea capaz de reconocer algunos de estos motivos conservados en el patógeno, pudiendo ser considerados como PAMPs.
3. Analizar filogenéticamente los genes que codifican para CAZymas en *Plectosphaerella*, agrupados por grupos y por la presencia o no de péptido señal extracelular, con el fin de confirmar una relación entre la abundancia de módulos o grupos concretos con la patogenicidad del hongo.

2. CAPÍTULO 2: MATERIAL Y MÉTODOS

Todos los programas utilizados, los scripts realizados y sus descripciones están disponibles en el repositorio *on line* Github dentro del proyecto *andreaalvarezp/Scripts_TFG*.

2.1 Micovirus

2.1.1 Búsqueda de micovirus en datos de RNAseq

Análisis de secuencias de RNAseq de Plectosphaerella

a. Datos iniciales

Se partió de archivos de datos de RNAseq de cada uno de los tres aislados de *Plectosphaerella*. Dichos datos habían sido obtenidos por secuenciación mediante la tecnología Illumina *single end* con un tamaño de lectura de 50 pares de bases y un formato de salida de lectura FASTQ. Este formato almacena el identificador de la secuencia Illumina, la secuencia biológica y la puntuación de calidad codificadas con caracteres ASCII. Para cada aislado se contó con lecturas de RNAseq del hongo cultivado *in vitro*, cultivado en los ecotipos Col-0 a 10, 16 y 24 hpi, y en *cyp79b2b3* a 10 y 16 hpi, con un total de 3 réplicas para cada condición y tiempo. Los datos brutos de secuencias SRA están depositados en el NCBI BioProject PRJNA614936 [1].

Para la búsqueda de micovirus en dichas secuencias se adaptó un procesado bioinformático previamente elaborado [10, 21] utilizando el clúster de supercomputación del CBGP a través de MobaXterm (https://mobaxterm.mobatek.net/), terminal que proporciona todas las herramientas necesarias de manera remota a través de un cliente SSH y comandos Unix a mi escritorio (Figura 1a).

b. Pretratamiento de los datos

El pretratamiento de los datos se realizó con BBTools [22], herramienta bioinformática multiproceso escrito en Java que permitió eliminar adaptadores de la secuenciación, secuencias repetidas, ribosómicas y normalizar la cobertura de los datos a través de los distintos módulos que proporciona. Durante el pretratamiento de los datos no se aplicaron filtros de calidad, ya que el RNA de *A. thaliana* está sobrerrepresentado con respecto al de *Plectosphaerella*. El interés reside en analizar el genoma del hongo, por lo que, para evitar que las lecturas correspondientes al mismo se eliminaran por no pasar los filtros de calidad, se dejaron los valores por defecto de los parámetros

c. Ensamblaje *de novo* de lecturas cortas

El proceso de ensamblaje *de novo* fue llevado a cabo por el programa *Trinity* [23]. Este paso consiste en la reconstrucción de los transcritos de longitud completa a partir de lecturas cortas con considerables tasas de error debidas a la secuenciación, es decir, realiza una reconstrucción del

transcriptoma. *Trinity* es un software que combina tres módulos independientes que se ejecutan de manera secuencial para procesar grandes volúmenes de lecturas de RNAseq.

d. Identificación de secuencias virales u homólogas

Se llevaron a cabo dos grandes filtrados de los datos. Para ello se utilizó la herramienta BLAST, cargada desde el módulo DIAMOND [24]. Se trata de un software alineamiento de secuencias para búsquedas de proteínas y DNA traducido, diseñado para el análisis de grandes secuencias.

El primer paso de la búsqueda consistió en hacer BLAST-X de las lecturas ensambladas con *Trinity* contra una base de datos personalizada de virus con un E (evalue) < 10^{-3}. En dicha base de datos se incluyeron virus de RNA y de DNA pequeños, *Cressdnaviricota*, virus de ssDNA circulares, *unclassified* DNA virus y virus de ssDNA. Los resultados pasados a tablas de Excel se examinaron manualmente para poder identificar posibles secuencias candidatas con una alta homología con algún virus. A partir de aquí se siguieron dos procesados distintos de los datos con el objetivo de comparar ambos y ver qué resultados ganan en coherencia y precisión.

El primer *pipeline* consistió en realizar un segundo BLAST-X utilizando como base de datos NCBI *nr* del marzo del 2021. Los resultados obtenidos se recogieron en tablas de Excel para examinarlas manualmente. El segundo procesado consistió en realizar un reensamblado de los *contigs* utilizando CAP3 (Contig Assembly Program) [25]. A continuación, se realizó un segundo BLAST-X contra la misma base de datos de virus personalizada como un segundo filtrado, y en última instancia se llevó a cabo un BLAST-X contra la misma base de datos del NCBI *nr* usada en el paso anterior. Los resultados obtenidos se recogieron en tablas de Excel para filtrar los datos manualmente.

Análisis de secuencias de RNAseq de Colletotrichum tofieldiae – Validación del pipeline

Para poder validar los resultados obtenidos en *Plectosphaerella*, es conveniente tener un control positivo que nos diga que el *pipeline* se ha seguido correctamente. Por ello, se utilizaron datos de RNAseq del hongo endófito *Colletotrichum tofieldiae* [20] y se les aplicó el mismo protocolo que a *Plectosphaerella* (Figura 1a). Se eligió *C. tofieldiae* porque otro estudio previo había identificado secuencias de RdRP (RNA Polimerasa dependiente de RNA) del virus de la mancha anular del nabo (*Turnip Ringspot Virus*, TRV) en secuencias de RNA del hongo [9]. Los datos de SRA SRP059724 se obtuvieron del NCBI BioProject PRJNA287627. Para el análisis de datos de *C. tofieldiae* se partieron de tres pools de datos de RNAseq obtenidos de experimentos en distintas condiciones: *in vitro*, e inoculado en plantas de Arabidopsis Col-0 en dos condiciones distintas, con fosfato (plusP) o sin fosfato (minusP) en el medio de cultivo. En el caso de *C. tofieldiae*, la secuenciación en Illumina se realizó de forma *paired-end*, por lo que se tenía la lectura *foward* y *reverse*, ambas contenidas en un mismo archivo y diferenciadas por la terminación de sus identificadores.

2.1.2 Comprobación de los resultados obtenidos en *Colletotrichum tofieldiae* por un método alternativo

Como validación de nuestros resultados para establecerlo como un método fiable de búsqueda de micovirus, se repitió el mismo procesado de datos llevado a cabo por Gilbert *et al*, (2019) [9] para *C. tofieldiae*. El procedimiento sigue la misma línea general, sin embargo, difiere en algunos pasos, como en la eliminación de las secuencias del transcriptoma que alinean con el genoma del hongo y la inclusión de herramientas de análisis de dominios como paso previo a la utilización de BLASTX (Figura 1b).

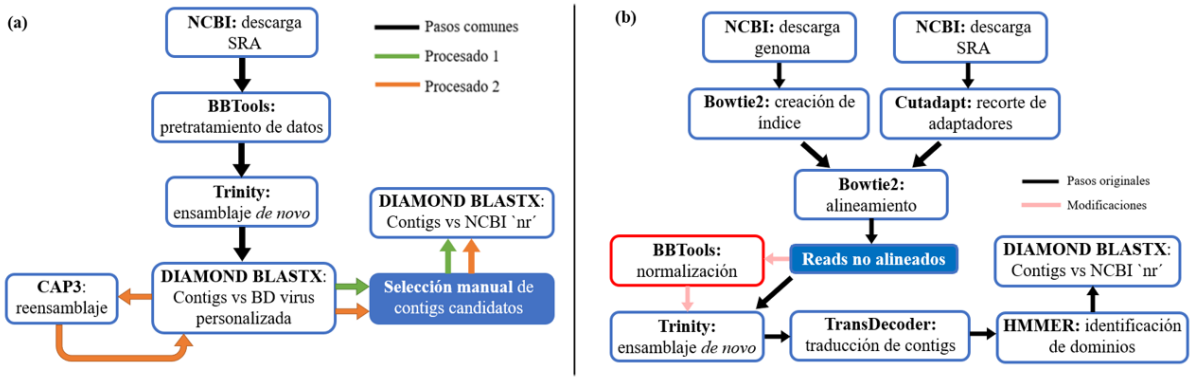


Figura 1: Protocolo empleado para el trabajo. Los colores de las flechas muestran que línea de trabajo sigue cada uno de los procesados llevados a cabo durante el análisis**. (a)** Protocolo original**. (b)** Protocolo Gilbert, K., *et al* (2019) con los pasos modificados en rojo.

En primer lugar se llevó a cabo la separación de los archivos SRA que contienen *paired-ends* en archivos independientes de lecturas *forward* y *reverse*. Para ello se utilizó el módulo *reformat.sh*, incluido dentro de la herramienta BBTools [22]. Los adaptadores, en este caso, se recortaron utilizando la herramienta Cutadapt [26].

Las secuencias de RNAseq se alinearon con el genoma completo de *C. tofieldiae*, con el objeto de filtrar las secuencias que no pertenecen al hongo y que por tanto pudieran ser de micovirus. El genoma de *C. tofieldiae* se obtuvo del NCBI GenBank Assembly Accession GCA_001618715.1 y se creó un índice del genoma descargado a partir de bowtie2build, módulo comprendido en la herramienta Bowtie2 [27]. El alineamiento de las secuencias de RNA y genómicas se realizó también con Bowtie2 [27], y se utilizó la opción *--un* para poder conservar aquellas secuencias que no alineaban, eliminando la mayor parte del transcriptoma del hongo de los datos. Antes de someter al conjunto de secuencias resultante a un ensamblaje *de novo* con la herramienta *Trinity* [23], tuvo que introducirse una modificación en el procesado e incluir así una normalización de los *reads* utilizando el módulo *bbnorm.sh* de la herramienta BBTools [22], ya que de otra forma el ensamblaje *de novo* no podía ejecutarse. Posteriormente se sometió a un análisis con la herramienta TransDecoder (http://transdecoder.github.io), que identifica regiones codificantes candidatas dentro de secuencias de transcripción.

Las proteínas predichas por TransDecoder se alinearon frente a una base de datos de RdRPs de virus de RNA con HMMER *hmmscan* (http://hmmer.org/). HMMER [28] se utiliza en este caso para buscar homólogos de secuencia y realizar alineamientos. La base de datos de de RdRPs de virus incluía las familias de RdRP: RdRP_1, RdRP_2, RdRP_3, RdRP_4, RdRP_5 y Mitovir_RNA_Pol. Para ello se descargó la lista de dominios de HMM completos de la página principal de Pfam, una base de datos de familias de proteínas representadas por múltiples alineamientos de secuencia y modelos ocultos de Markov [29], se utilizó el comando *hmmfetch* para buscar los perfiles de las familias que me interesan y con *hmmpress* se creó con ellas una base de datos que pudiera ser utilizada en *hmmscan*. Este alineamiento permitió la identificación de secuencias de RdRPs virales que contengan una similitud limitada con los datos, ejecutándose con un valor de E (e-value) de 10.0. Los resultados sacados de HMMER se compararon contra la base de datos del NCBI *nr* de marzo de 2021 para confirmar la identidad de las secuencias extraídas. Además, se realizó otro BLAST-X de una base de datos creada con las secuencias de las poliproteínas de TRV (Accessions YP_003193665.1 y YP_003193666.1) contra los ensamblados resultantes de *Trinity,*

2.2 Péptidos fitorreguladores

2.2.1 Búsqueda de homólogos a los péptidos SCOOP

El primer paso fue crear una base de datos a partir de los proteomas de los tres aislados de *Plectosphaerella*, modificando los identificadores para evitar repeticiones (Figura 2). Esta base de datos se utilizó para buscar secuencias homólogas a los péptidos SCOOP mediante el módulo BLAST-P de la herramienta BLAST+ (Basic Local Alignment Search Tool) [30]. La tarea se ejecutó desde el cluster de supercomputación del CBGP. En este análisis se incluyeron los péptidos SCOOP10 y SCOOP12 de Arabidopsis [12], los péptidos SCOOP-LIKE encontrados en diferentes especies de *Fusarium* por Hou, S. *et al* (2021) [16], la secuencia completa proteína FGSG_07177 de *F. graminearum* que contiene esta secuencia en su extremo N-terminal, secuencias con homología a SCOOP de *Trichoderma atroviridis*, *Verticilium dahliae*, *Magnaporthe oryzae* y *Neurospora crassa* [17], los motivos SCOOP de los péptidos PROSCOOP6 y PROSCOOP11 que inducían la producción de ROS en hoja [17] y los genes A0A0D2XZ19 y A0A0M9EVJ7, que son epítopos activos de Fusarium [17]. Una vez se realizó el BLAST-P, se creó un archivo FASTA con las secuencias completas de las proteínas de los tres aislados de *Plectosphaerella* donde había alguna coincidencia con cada uno de los péptidos SCOOP y SCOOPL, a las que se añadió la secuencia original del péptido SCOOP correspondiente (Figura 2). Este archivo FASTA se utilizó como input de la herramienta MEME (Multiple Em for Motif Elicitation versión 5.3.3) [31]. Esta herramienta, además, produce LOGOS de secuencia descargables para cada motivo descubierto. Se fijaron los parámetros de búsqueda en 10 motivos con una longitud entre 3 y 10 aminoácidos, y los motivos encontrados se alinearon con la

herramienta de alineamiento múltiple ClustalOmega [32]. Los alineamientos resultantes se representaron utilizando ESPript3 (https://espript.ibcp.fr) [33].

2.2.2 Búsqueda de homólogos a putativos SPPs

Para realizar este análisis se utilizaron las secuencias de aminoácidos de putativos SPPs de Arabidopsis suministradas por el grupo de investigación (en adelante SPPs). Antes de analizar la presencia de homólogos de estos putativos SSPs en *Plectosphaerella*, se analizó la existencia de secuencias homólogas a cada uno de ellos en otros genes de Arabidopsis mediante un BLAST-P [30] desde la página del NCBI contra el proteoma de Arabidopsis. Estos homólogos en Arabidopsis se incorporaron en análisis posteriores de los proteomas de los tres aislados. Sin embargo, la información que aportaron no fue relevante ya que no mejoró la robustez de los resultados, por lo que no se incluyeron en los resultados presentados en este trabajo.

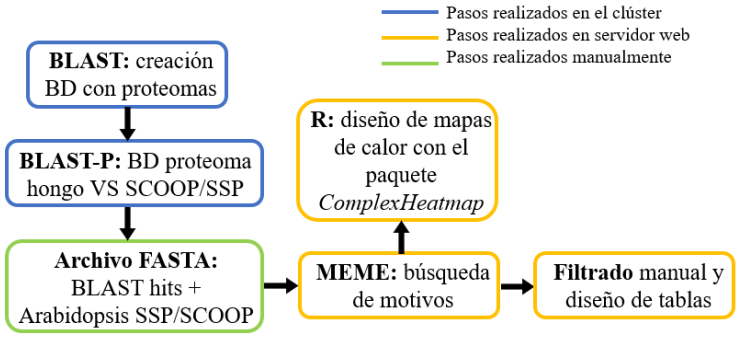


Figura 2: Esquema general del procesado de datos para la búsqueda e identificación de motivos SSPs en *Plectosphaerella*.

Para la búsqueda de motivos homólogos a los SSPs en *Plectosphaerella* se realizó un BLAST-P [30] con cada uno de los SSPs contra a una base de datos con las secuencias en formato FASTA de los proteomas de los tres aislados del apartado anterior (Figura 2). Con los resultados se creó un archivo FASTA donde se incluyó además la secuencia original del SSP correspondiente y las secuencias de las proteínas homólogas completas de Arabidopsis. Este archivo se analizó mediante la herramienta MEME (Multiple Em for Motif Elicitation versión 5.3.3) [31] fijando la búsqueda en 15 motivos con una longitud entre 3-10 aminoácidos.

Para el análisis de expresión de los genes donde se encuentran dichos motivos SSPs se realizaron mapas de calor o *heatmaps*. Para ello se recurrió al paquete *ComplexHeatMaps* implementado en R [34]. Los datos utilizados de FPKM y de expresión relativa fueron obtenidos de un experimento de RNAseq previo [1]. Mediante la función *scale()* se escalaron los datos para que estuvieran confinados en un rango [-2, 2]. La función *scale()* afecta a las columnas de la matriz, por lo que para poder escalar en relación a genes se operó con la matriz traspuesta utilizando el comando *t(scale()),*y una vez escalados se deshizo la trasposición de la misma forma, *t()*. Se contrastaron cuatro métodos de *clustering* jerárquico aglomerativo: el método de distancia mínima (*linkage single*), el método de distancia máxima (*complete*), el método de distancia promedio (*linkage*

average) y el método de Ward (Ward), eligiéndose el método de Ward. Este método une, en cada etapa, los dos clústeres para los que se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias dentro de cada clúster, es decir, aplica el criterio de la mínima varianza, por lo que sabremos que los genes que se agrupan en los mismos clústeres presentan una mínima variación en cuanto a su perfil de expresión [34]. Para que se pueda aplicar de manera recursiva, este método requiere que se realice en base a la distancia euclídea, que tiene en cuenta el desplazamiento individual de cada una de las variables y es la más adecuada para el manejo de datos de expresión (Figura 3).

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \qquad , \forall\, i = 1, 2, \ldots, n$$

Figura 3: Fórmula para calcular la varianza interna de cada clúster o la suma de cuadrados internos del clúster (WCSS). Se trata de la suma de las distancias euclídeas al cuadrado entre cada observación (x_i) y el centroide (μ_k) de su clúster, siendo i el número de clúster.

El diagrama de codo (Figura 4) busca seleccionar la cantidad ideal de grupos o clústeres (k) a partir de la minimización de la suma de cuadrados de las distancias dentro de cada clúster (WCSS, *Within Clusters Summed Squares*). Intenta optimizar el reparto de las observaciones en k clústeres de forma que la suma de las varianzas internas de todos ellos sea lo menor posible [35]. Para cuantificar la varianza interna de cada clúster ($W(C_k)$) se utiliza, como se ha mencionado anteriormente, la distancia euclídea (Figura 3). Si representamos el número de clústeres k respecto a la WCSS (diagrama de codo, Figura 4), el número óptimo de k corresponde a la posición en abscisas donde la pendiente de la curva se suaviza, es decir, donde se encuentra el "codo".

.

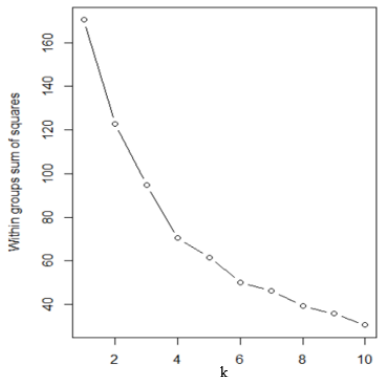


Figura 4: Diagrama de codo que enfrenta el número de clústeres frente a la suma de cuadrados internos del clúster (WCSS). La pendiente se suaviza por primera vez cuando k = 4, por lo que podemos establecer como parámetro la existencia de cuatro clústeres.

2.3 Análisis de CAZymas

2.3.1 Identificación de CAZymas en proteoma de *Plectosphaerella*

Se identificaron secuencias de CAZymas tanto en el genoma como en el proteoma de los tres aislados de *Plectosphaerella* utilizando dos herramientas distintas, CUPP y dbCAN2, que engloban varios métodos de anotación de CAZymas. CUPP [19] es una herramienta de anotación funcional y agrupación no basada en alineamientos que utiliza patrones de péptidos únicos

conservados para realizar agrupaciones automatizadas de proteínas y formar grupos. La herramienta dbCAN2 [36] es un servidor web creado en 2012 para proporcionar un servicio público para la anotación CAZymas automatizada para genomas recién secuenciados. Para ello utiliza tres métodos de búsqueda. El primero es HMMER [28] contra la base de datos dbCAN, que utiliza cadenas de Markov ocultas HMM (*Hidden Markov Models*) para explorar el espacio de soluciones más probable de manera más rápida. El segundo es DIAMOND BLAST [24] contra la base de datos de secuencias de CAZymas, CAZy, y el tercero es Hotpep [36], que anota CAZymas mediante la búsqueda contra la biblioteca PPR *(Peptide Pattern Recognition*) de motivos peptídicos cortos conservados presente en diferentes familias CAZymas. La salida de dbCAN2 se filtró de tal manera que solo se manejaron aquellas secuencias de CAZymas que habían sido detectadas por al menos dos de los métodos que utiliza, tal y como recomienda el servidor, con el objetivo de ganar robustez en los resultados. En el caso de CUPP, solo opera con cinco grupos, excluye los dominios CBM en su clasificación, mientras que dbCAN2 se ajustó a seis grupos: AA, CE, CBM, GH, GT y PL.

La herramienta CUPP se utilizó desde el servidor web, mientras que dbCAN2 se ejecutó de manera remota desde el cluster de supercomputación del CBGP. El análisis con ambas herramientas se realizó con los tres aislados por separado.

2.3.2 Identificación de CAZymas secretadas

La determinación de CAZymas secretadas extracelularmente (Tabla Suplementaria S4) se llevó a cabo con la herramienta SECRETOOL [37], servidor web que comprende un grupo de módulos que permiten hacer predicciones de secretomas a partir de archivos de secuencias de aminoácidos. A partir de los resultados de dbCAN2 y CUPP se generó un archivo FASTA por cada una de los aislados que se ejecutó en SECRETOOL.

Con estos datos se construyeron un total de 6 árboles filogenéticos, uno por cada grupo de CAZymas. Para ello se realizó un alineamiento múltiple de las secuencias con ClustalOmega [32] y los árboles se construyeron con IQ-TREE [38]. Ambos fueron ejecutados como módulos dentro del cluster de supercomputación del CBGP. La elección del modelo de sustitución aminoacídica óptimo para cada árbol se realizó con el parámetro *-m TEST* de IQ-TREE y con un bootstrap de 1000 para ganar robustez. La representación posterior de los árboles se llevó a cabo con el servidor web iTOL (Interactive Tree Of Life, https://itol.embl.de), herramienta en línea para la visualización, manipulación y anotación de árboles filogenéticos y de otro tipo [39]. Para representar los diagramas circulares de alrededor de los árboles se crearon dos bases de datos por árbol: una con los genes totales etiquetados según el aislado al que pertenecen y otra únicamente con aquellos genes que codifican para CAZymas secretadas, cada una de clases marcada con una etiqueta y un código de colores característico.

3. CAPÍTULO 3: RESULTADOS

3.1 Búsqueda de micovirus en datos de RNAseq

En la utilización del protocolo inicial el resultado fue negativo en la detección de micovirus. A continuación se presentan los resultados intermedios de dicho protocolo para evaluar el método de detección y su eficiencia.

Análisis de secuencias de RNAseq de Plectosphaerella

Se siguió el protocolo descrito en la Figura 1a. El pretratamiento de los datos con BBTools [22], que comprendía el agrupamiento de lecturas, eliminación de adaptadores, secuencias cortas, secuencias ribosómicas, y la normalización, redujo en un el tamaño de los archivos iniciales en un 80% (Tabla 1):

Tabla 1: Tamaño de los archivos de datos de RNAseq antes y después de su pretratamiento con BBTools.

Aislado *Plectosphaerella*	Tamaño inicial (GB)	Tamaño tras el pretratamiento (GB)	Porcentaje del total conservado
PcBMM	13.7	2.7	19.70%
Pc2127	11.5	2.3	20.00%
P0831	12.5	2.8	20.00%

Del paso de ensamblaje con *Trinity* se obtuvieron una serie de *contigs* cuyo número se redujo en más de un 90% tras realizar un reensamblaje con CAP3 (Tabla 2).

Tabla 2: Número de contigs resultantes tras el ensamblaje con Trinity y un paso de reensamblaje con CAP3. Entre paréntesis se representa el número de contigs que se han mantenido tras el reensamblaje.

Aislado *Plectosphaerella*	Nº contigs tras Trinity	Nº contigs tras CAP3
PcBMM	51309	3940 (7.68%)
Pc2127	42227	3292 (7.79%)
P0831	39567	3334 (8.43%)

Los pasos posteriores de ensamblaje y búsqueda de micovirus con DIAMOND BLAST [24] dieron una serie de *hits* de los cuales se sacaron los archivos FASTA utilizados como puntos de partida para los dos tipos de procesados que se siguieron. En la Tabla 3 se muestra la comparación de los resultados de ambos procesados. El número de *hits* tras el primer BLAST indican el número de zonas del transcriptoma de *Plectosphaerella* donde podríamos encontrar algún micovirus, y se debieron a que BLAST encontró similitud entre esa región concreta del transcriptoma y alguno de los micovirus que incluimos en la base de datos personalizada. Estas zonas están distribuidas en el número indicado de *scaffolds* correspondientes del genoma, ya que en un mismo *scaffold* de *Plectosphaerella* podemos encontrar más de un acierto.

Tabla 3: Comparación del número de *hits* de cada uno de los dos procesados utilizados para la identificación de micovirus en *Plectosphaerella*. El procesado 1 incluye un solo BLAST contra el NCBI *nr*, mientras que el procesado dos incluye pasos extra de filtrado

			Procesado 1	Procesado 2
Aislado *Plectosphaerella*	**Nº hits tras primer BLAST**	**Nº scaffolds correspondientes**	**Hits finales**	**Hits finales**
PcBMM	8265	4775	8525	6745
Pc2127	6955	3999	7077	6685
P0831	7069	4042	7310	6829

El procesado 2 (Tabla 3) incluía un reensamblaje de *scaffolds* con CAP3 [23] y un filtrado más contra la base de datos de virus personalizada antes de realizar la búsqueda final contra el NCBI. Este paso de reensamblaje redujo el número de *contigs* más de un 90% (Tabla 2). Se realizó una búsqueda manual mediante las palabras clave "vir" y "RdRP", no encontrándose ningún resultado correspondiente a micovirus. Los *hits* correspondieron a genes de la planta y del hongo que tienen homología muy lejana con secuencias virales.

Análisis de secuencias de RNAseq de Colletotrichum tofieldiae – Validación del pipeline

Para validar el método con un set de datos similar, se repitió el protocolo de la Figura 1a utilizando datos de RNAseq del hongo *C. tofieldiae,* en el que Gilbert, K., *et al,* (2019) utilizando otro protocolo (Figura 1b), habían identificado secuencias de RdRP (RNA Polimerasa dependiente de RNA) del virus de la mancha anular del nabo (*Turnip Ringspot Virus*, TRV). En este caso tampoco se encontraron micovirus., Al igual que con *Plectosphaerella*, el número de *contigs* sigue reduciéndose más de un 90% tras el reensamblaje con CAP3 (Tabla 4).

Tabla 4: Número de contigs resultantes tras el ensamblaje con Trinity y un paso de reensamblaje con CAP3. Entre paréntesis se representa el número de contigs que se han mantenido tras el reensamblaje.

Condiciones muestras *Colletotrichum tofieldiae*	**Nº contigs tras Trinity**	**Nº contigs tras CAP3**
in vitro	16241	794 (4.88%)
plusP	40969	4184 (10.21%)
minusP	41046	4161 (10.14%)

Además, podemos ver que el porcentaje de *hits* finales resultado del último BLAST se reduce a un 4.88% del total en las condiciones *in vitro*, a un 10.21% en condiciones con fósforo (plusP) y a un 10.14% en condiciones sin fósforo (minusP), por lo que hay una mayor reducción en los archivos más extensos, que son los que incluyen secuencias de la planta (plusP, minusP).

Como muestra la Tabla 5, el procesado 2, que incluye pasos extra en el filtrado, da un menor número de lecturas finales. Este resultado es similar al obtenido con *Plectosphaerella*. (Tabla 3).

Tabla 5: Comparación del número de *hits* de cada uno de los dos procesados utilizados para la identificación de micovirus en *Colletotrichum tofieldiae*. El procesado 1 incluye un solo BLAST contra el NCBI *nr*, mientras que el procesado dos incluye pasos extra de filtrado.

			Procesado 1	**Procesado 2**
Condiciones muestras *Colletotrichum tofieldiae*	**Nº hits tras primer BLAST**	**Nº scaffolds correspondientes**	**Hits finales**	**Hits finales**
in vitro	1372	840	872	825
plusP	8634	4821	5204	4503
minusP	8442	4764	5123	4467

3.1.1 Comprobación de los resultados obtenidos en *Colletotrichum tofieldiae* por un método alternativo

La aplicación exacta del método planteado por Gilbert, K., et al. (2019) [9] para la detección de micovirus con el fin de validar los resultados de nuestro método no se pudo concluir. A la hora de realizar el ensamblaje de las lecturas con *Trinity* [23], la no normalización previa de los datos llevó a errores que no se pudieron solventar con los parámetros proporcionados en el estudio original. Por ello, se decidió modificar dicho protocolo e introducir un paso de normalización de los datos con BBTools [22], concretamente utilizando el módulo *bbnorm.sh*, aplicable a *paired-end reads* (Figura 1b).

Tras eliminar las lecturas de RNAseq que alineaban con genoma del hongo, la cantidad de lecturas resultante con las que continuar el procesado se redujo en un 79.48% en el caso de las condiciones *in vitro* (Tabla 6). Esto quiere decir que, de todas las lecturas de RNAseq iniciales, el 20.52% no pertenecían al hongo, y por lo tanto son secuencias candidatas a ser identificadas como micovirus. Por otro lado, solo se redujo un 0.63% en el caso de las condiciones plusP y un 0.90% en el caso de las condiciones minusP. Esto se debe a que en estas condiciones contamos también con el transcriptoma de la planta, a la que corresponde la mayor parte de los transcritos ensamblados. Para mejorar la precisión del análisis en estas condiciones habría que realizar un segundo alineamiento contra el genoma de *Arabidopsis thaliana*.

El paso extra de normalización redujo el número de lecturas de manera considerable, un 70.20% en el caso *in vitro,* 90.08% en plusP y un 92.85% en minusP. Estas diferencias en los porcentajes entre las condiciones *in vitro* e *in planta* se deben a que la presencia del transcriptoma de Arabidopsis, en lugar de varios genomas pequeños de virus.

A partir de aquí solo se continuó con las condiciones *in vitro,* donde se realizó dicho el ensamblaje con *Trinity* y la identificación de secuencias virales con *TransDecoder* y HMMER. Se identificaron dos secuencias correspondientes a la RdRP. Como comprobación extra se realizó un BLAST-X de las dos poliproteínas de TRV contra los archivos de los ensamblajes de *Trinity* de las condiciones *in vitro,* dando lugar a dos *hits* correspondientes con la primera poliproteína, con una identidad de secuencia de 57% y 50%.

Tabla 6: Número de lecturas antes y después del alineamiento. Se utilizó la etiqueta *–un* para quedarnos con las lecturas que no alinearon con el genoma de *Colletotrichum tofieldiae*.

Condiciones muestras *Colletotrichum tofieldiae*	**Nº lecturas iniciales**	**Nº lecturas tras alinear con Bowtie2**	**Nº lecturas tras normalización**
in vitro	63299700	12993573 (20.52%)	3872562 (29.80%)
plusP	281135542	279373138 (99.37%)	27722242 (9.92%)
minusP	428534074	424690704 (99.10%)	30372126 (7.15%)

3.2. Presencia de posibles motivos miméticos de péptidos fitorreguladores

3.2.1. Homólogos a los péptidos SCOOP

Se han encontrado motivos homólogos a SCOOP12 en un gen de cada uno de los aislados de *Plectosphaerella* (Figura 5). En los aislados PcBMM y Pc2127, los genes que contienen estos motivos (PcBMM_AIM001598 y Pc2127_AIM00928, respectivamente) son homólogos entre sí con un 100% de identidad de secuencia, y tienen similitud mayor con el SCOOP-LIKE de *F. oxysporum f.sp lycopersici* (Fol) [16] que con la secuencia SCOOP12 de Arabidopsis (Figura 5a). Ambos homólogos en *Plectosphaerella* tienen conservados el motivo SXS, importante para la actividad del péptido, y el motivo AGG del extremo C-terminal. También se ha encontrado una alta homología entre el gen Pc2127_AIM010202 y el péptido SCOOP6#2 de Arabidopsis, que provoca la producción de ROS y respuesta inmune [17], aunque en este caso no está conservado el motivo SXS ni el motivo AGG (Figura 5b).

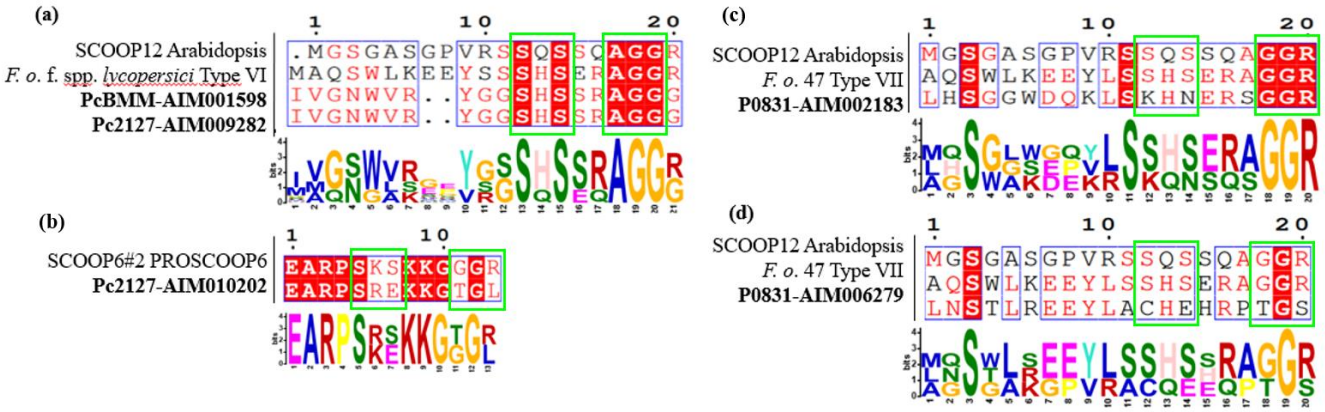


Figura 5: Secuencias de *Plectosphaerella* con motivos comunes a péptidos SCOOP de Arabidopsis o a SCOOP-LIKE de *Fusarium*. Los alineamientos se realizaron con ESPript y los logos con MEME.

En P0831, los genes que contienen motivos homólogos a SCOOP12 (P0831_AIM006279 y P0831_AIM002183) son diferentes a los genes encontrados en los otros dos aislados. En el primer caso, la mayor homología se ha encontrado con *F. oxysporum* Fo47 [16], y no tiene conservado el motivo asociado a la actividad del péptido SXS, pero sí parte del motivo GGR del extremo C-terminal (Figura 5c), mientras que en el segundo caso (Figura 5d), la homología reside únicamente con *F. oxysporum* Fo47 [16], no con SCOOP12.

Se analizó la expresión in planta de los genes de *Plectosphaerella* con homología a motivos SCOOP en los datos obtenidos de un experimento previo en Arabidopsis [1] (Figura 6). El gen PcBMM_AIM001598 con homología a SCOOP12 comienza a expresarse *in planta* a 16 hpi, tanto en plantas del genotipo silvestre Col-0 como en el mutante inmunodeprimido *cyp79b2b3*, descendiendo su expresión a 24 hpi en Col-0 a los mismos niveles que *in vitro*. Al carecer de datos de expresión a 24h para *cyp79b2b3* no podemos saber si ocurre lo mismo en estas plantas. Pc2127_AIM009282, homólogo a PcBMM_AIM001598, comienza a expresarse a 10 hpi y alcanza un máximo a las 16 hpi para volver a descender a 24 hpi, con un patrón muy similar en Col-0 y *cyp79b2b3*, aunque su variación temporal no es muy grande, ni tampoco varía mucho respecto a la expresión *in vitro*. La expresión más remarcable es la de Pc2127_AIM010202, homólogo al péptido SCOOP6#2 que, alcanza a 10 hpi, tanto en Col-0 como en *cyp79b2b3*, más de 3 veces su expresión *in vitro*, manteniendo su expresión en el mutante a 16 hpi y desapareciendo en Col-0. Los genes de P0831 no se han representado ya que no tienen expresión *in planta*.

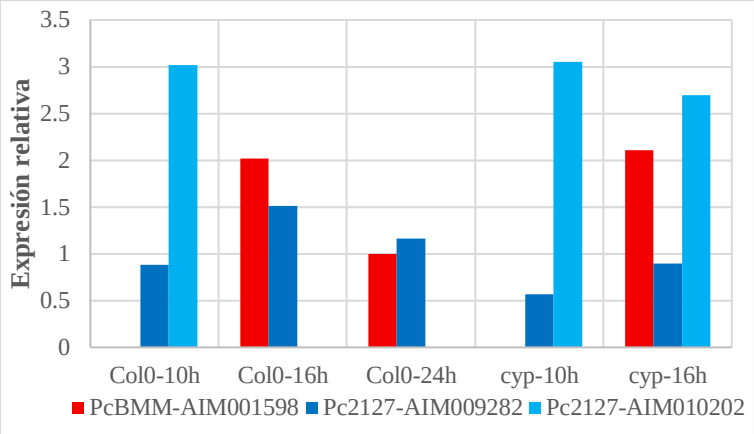


Figura 6: Expresión relativa *in planta* respecto a *in vitro* de los genes de los aislados de *Plectosphaerella* PcBMM y Pc2127 con homología a motivos SCOOP. Los genes de P0831 no han sido representados al carecer de expresión *in planta.*

3.2.2. Homólogos a putativos SSPs

La Tabla 7 muestra el número de motivos homólogos a cada uno de los putativos SSPs de Arabidopsis en cada uno de los tres aislados de *Plectosphaerella*. La isoforma SSP8.3 del péptido SSP8 es la secuencia que mostró un mayor número de motivos conservados en los tres aislados, con un total de 27, de los que la mayor parte corresponden a una zona de la secuencia rica en prolinas. Los siguientes péptidos con más motivos homólogos encontrados en *Plectosphaerella* son la isoforma SSP8.1 del péptido SSP8 y el péptido SSP4, con 22 y 21 motivos respectivamente. Cabe destacar que no se encontró ningún motivo coincidente con la isoforma SSP8.2 del péptido SSP8. Aunque esta isoforma es idéntica a SSP8.1 con excepción del extremo C-terminal, MEME no reconoció motivos comunes a *Plectosphaerella* hasta que no se eliminó la isoforma 1 del análisis. Lo mismo ocurre con SSP1 y SSP14. El aislado con un mayor número de motivos candidatos es Pc2127, seguida de PcBMM y en último lugar, P0831. El número de motivos encontrados en los aislados PcBMM y Pc2127 es muy similar en todos los péptidos. No es así

con P0831, cuyo valor difiere con respecto al resto, siempre encontramos más o menos motivos que en los otros dos aislados, y en algunos casos, como SSP12, ninguna coincidencia.

Tabla 7: Número de motivos homólogos a SSPs de Arabidopsis en los aislados PcBMM, Pc2127 y P0831.

Arabidopsis SSP	PcBMM	Pc2127	P0831	Motivos totales
AT2G25510.1 (SSP1.1)	3	3	4	10
AT2G25510.2 (SSP1.2)	0	0	1	1
AT3G49550 (SSP2)	3	3	3	9
AT4G31130 (SSP3)	1	1	1	3
AT3G48640 (SSP4)	6	9	6	21
AT1G65295 (SSP5)	2	2	2	6
AT5G20790.1 (SSP6.1)	4	4	4	12
AT5G20790.2 (SSP6.2)	3	3	1	7
AT2G37750 (SSP7)	6	6	8	20
AT5G44570.1 (SSP8.1)	7	7	6	22
AT5G44570.2 (SSP8.2)	0	0	0	0
AT5G44570.3 (SSP8.3)	10	11	6	27
AT1G68945 (SSP9)	4	4	1	9
AT5G66052 (SSP10)	3	2	2	7
AT5G24570 (SSP11)	6	4	2	12
AT5G42530 (SSP12)	3	3	0	6
AT4G29735.1 (SSP14.1)	0	1	1	2
AT4G29735.2 (SSP14.2)	3	3	5	11
TOTAL	64	66	53	185

Que un SSP tenga homólogos en dos aislados diferentes no significa que estos contengan el mismo motivo. La Figura 7 muestra el diagrama de Venn de los motivos encontrados en uno, dos, o los tres aislados de *Plectosphaerella*. Existen 23 motivos de aminoácidos que son comunes a los tres aislados. PcBMM y Pc2127, comparten la mayor parte de los motivos, 36, y el número de motivos exclusivos para cada aislado es mucho mayor para P0831 que para PcBMM y Pc2127. P0831, además, no posee ningún motivo común con alguno de los otros dos aislados por separado. En la búsqueda de motivos con la herramienta MEME [31], se estableció una longitud de motivo de hasta 15 aminoácidos. Dentro de esos 15 aminoácidos, el número de residuos conservados con Arabidopsis en cada motivo encontrado osciló entre 1 y 7, siendo más frecuentes los motivos con 3 o 4 residuos conservados (Figura 8).

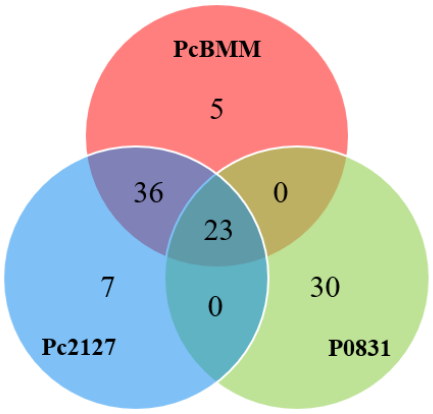


Figura 8: Diagrama de Venn de la distribución de motivos comunes y exclusivos en PcBMM, Pc2127 y P0831.

Aunque se aprecien unos valores más altos en PcBMM y Pc2127, esto se debe a que se ha encontrado un mayor número de motivos (Tabla 7), por lo que se podría concluir que siguen la misma tendencia. Pc2127 es el único aislado con algún motivo con 7 aminoácidos conservados.

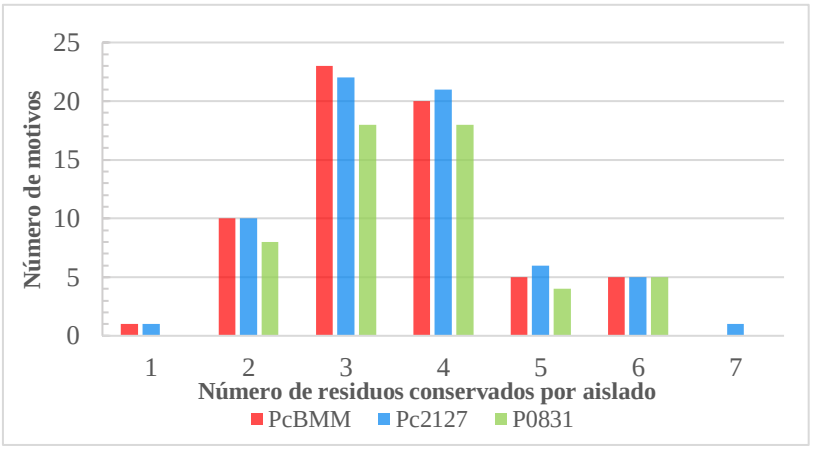


Figura 7: Número de residuos conservados por motivo homólogo a los SSPs de Arabidopsis en cada aislado de *Plectosphaerella*.

Para tener más información sobre la relevancia de estos motivos se estudiaron los perfiles de expresión de los genes que los contienen. Los candidatos para este estudio se seleccionaron en función del número de residuos conservados, y se representaron aquellos cuya expresión superaba al menos dos veces la expresión del gen *in vitro* (Tablas Suplementarias S1, S2 y S3). Los datos de FPKM se tomaron de experimentos anteriores del grupo [1]. Los motivos más interesantes son los que están en genes sobreexpresados *in planta* con relación a *in vitro*, ya que pueden estar implicados en la interacción con el huésped. La Figura 9 muestra la expresión de genes de cada aislado para cada uno de los tratamientos (*in vitro* y en las distintas condiciones *in planta*). Los análisis de mapas de calor agruparon los genes según su expresión en cuatro clústeres distintos.

En PcBMM (Figura 9a), el clúster 1 corresponde a los genes que se expresan mucho más *in vitro* que en las otras condiciones, por lo que se espera que los efectos sobre la interacción hongo-planta serán menos importantes que los de los genes en otros clústeres. El clúster 2 corresponde a genes más expresados a 10 hpi en el mutante *cyp79b2b3*. Este set de genes también incluye algunos

expresados más en Col-0 a 10 hpi. El clúster 3 agrupa genes que se expresan en mayor medida en Col-0 a 24 hpi y en *cyp79b2b3* a 16 hpi. El clúster 4 corresponde un grupo de genes expresados en mayor medida en Col-0 a 16 hpi. En el caso de Pc2127 (Figura 9b), el clúster 1 también corresponde a genes más expresados *in vitro* que en las otras condiciones. La principal diferencia con PcBMM reside en la distribución de los otros tres clústeres. En este caso se agrupan los genes expresados a 16 hpi, tanto en Col-0 como en *cyp79b2b3* en el clúster 2, y quedan bien diferenciados otros dos clústeres, uno con los genes expresados en *cyp79b2b3* a 10 hpi y otro con Col-0 a 24 hpi. Los genes expresados en Col-0 a 10 hpi no se agrupan en ningún clúster concreto, sino que se distribuyen con los genes de los clústeres 2 o 3. En el caso de P0831 (Figura 9c), la mayor parte de los genes solamente se expresan *in vitro* formando un gran grupo (clúster 2). Los otros 3 clústeres están formados únicamente por 1 o 2 genes, y corresponden a genes cuya expresión *in planta* es superior a la *in vitro* en distintas condiciones.

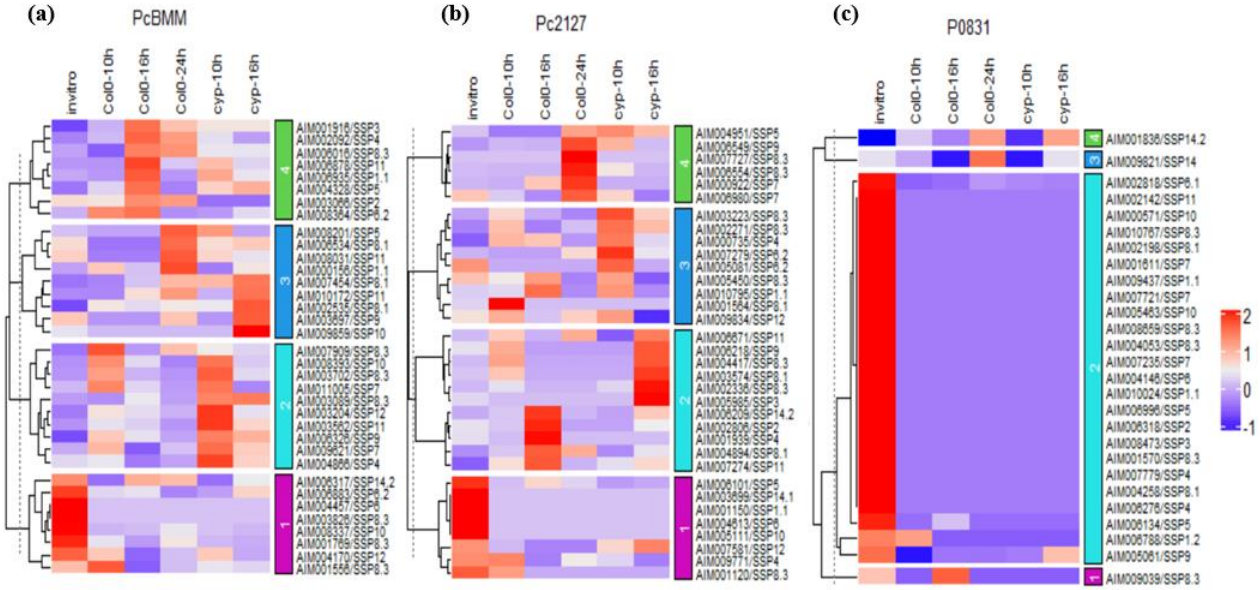


Figura 9: Mapas de expresión de genes de *Plectosphaerella* con motivos homólogos a los SSP de Arabidopsis en cada una de los aislados PcBMM **(a),** Pc2127 **(b)** y P0831 **(c).** Los mapas se realizaron utilizando datos de expresión en FPKM y representados con el paquete *ComplexHeatMap* de R.

En el caso de los 23 motivos compartidos entre proteínas de los tres aislados de *Plectosphaerella* (Figura 10) fue posible comparar entre los tres aislados la expresión de los genes que los contienen. Para ello, se elaboró un mapa de calor donde se representó la expresión relativa a *in vitro* de los 12 genes homólogos de los tres aislados con motivos iguales y expresión *in planta*. Se siguió el mismo protocolo que con los mapas de calor anteriores y, en este caso, se definieron 5 clústeres.

El primer clúster está formado por dos genes con sobreexpresión en Pc2127 a distintas condiciones, principalmente en Col-0 a 24 hpi. El clúster 2 contiene el único gen que muestra alguna expresión en P0831, otro gen que toma valores positivos de expresión en Col-0 a 16 y 24 hpi y tres genes que se sobreexpresan en diferentes condiciones en Pc2127, principalmente en

Col-0 a 16 hpi. Los clústeres 3 y 4 agrupan genes sobreexpresados principalmente en PcBMM en distintas condiciones. Por último, el clúster 5 agrupa genes altamente expresados en Pc2127 *cyp79b2b3* a 16 hpi, aunque también se percibe sobreexpresión en PcBMM

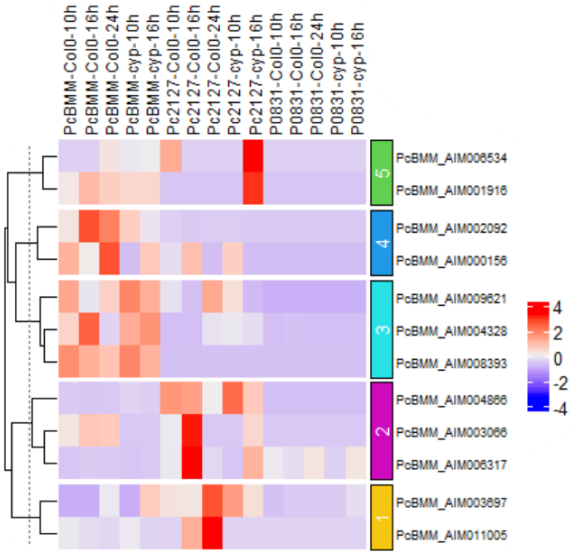


Figura 10: Mapas de expresión de genes de *Plectosphaerella* con motivos homólogos a los SSP de Arabidopsis comunes a los tres aislados. El nombre de las filas se ha fijado con el identificador del gen de PcBMM, pero todos ellos tienen su homólogo en Pc2127 y P0831. Los mapas se realizaron utilizando datos de expresión relativa in planta vs in vitro y representados con el paquete *ComplexHeatMap* de R.

3.3. Análisis de las CAZymas presentes en los tres aislados de *Plectosphaerella*

La Tabla 8 muestra las CAZymas totales y secretadas detectadas por las herramientas CUPP [19] y dbCAN2 [36]. Se realizó un BLAST-P para contrastar ambos grupos de datos y se comprobó que todas las CAZymas detectadas por dbCAN2 estaban presentes en la salida de CUPP, por lo que se decidió trabajar únicamente con las primeras, ya que han sido detectadas por las dos herramientas y por lo tanto los resultados eran más robustos. La proporción de CAZymas secretadas fue muy similar para ambos métodos, siendo algo mayor para el aislado patogénico, PcBMM y algo menor para el aislado P0831, pero las diferencias no fueron estadísticamente significativas (X^2> 0,70, P>0,05).

Tabla 8: CAZymas detectadas por las herramientas CUPP y dbCAN2

	PcBMM	**Pc2127**	**P0831**
Proteoma CUPP			
CAZymas totales	635	641	641
CAZymas secretadas	155	148	144
Proporción	24,4%	23,1%	22,5%
Proteoma dbCAN2			
CAZymas totales	590	591	575
CAZymas secretadas	145	137	129
Proporción[1]	24,6%	23,2%	22,4%

[1] Proporción de CAZymas secretadas respecto al total, detectadas mediante la herramienta SECRETOOL [37].

La Tabla Suplementaria S4 muestra la clasificación de las CAZymas encontradas en cada aislado según los grupos glucosiltransferasas [GT], glucósido hidrolasas [GH], polisacáridos liasas [PL], esterasas de carbohidratos [CE] y enzimas para las actividades auxiliares [AA], y la Tabla Suplementaria S5 la distribución de módulos CBM en cada uno de esos grupos.

Se realizó un análisis filogenético de las CAZymas dentro de cada grupo etiquetando las que pertenecen a cada aislado y si son secretadas o no (Figura 11). Además, se realizó un árbol para las CAZymas de cada uno de los cinco grupos que además tenían módulos CBM. Los resultados mostraron que la distribución de CAZymas es homogénea en relación con el tipo de aislado, lo que indica que cada gen que codifica para CAZymas posee, en la mayor parte de los casos, un homólogo en los tres aislados, o, al menos, en PcBMM y Pc2127. La topología de los árboles agrupó juntos los homólogos de estos dos aislados y, en un nodo diferente, el aislado P0831, lo que coincide con la posición filogenética de los dos primeros aislados [1].

Existe una distribución desigual del número de CAZymas y de la proporción de CAZymas secretadas dentro de cada uno de los grupos. El árbol con un mayor número de miembros es GH, donde casi la mitad de las CAZymas contienen el péptido señal, y están distribuidas por todo el árbol (Figura 11a). El grupo PL (Figura 11b) es el que menos instancias posee, y la distribución de CAZymas secretadas forma una serie de bloques que agrupan instancias contiguas dentro del árbol, seguido del grupo CE (Figura 11c). El grupo GT (Figura 11d) posee un número intermedio de instancias, pero la cantidad de CAZymas secretadas es muy inferior al resto, 8, y estas son cercanas filogenéticamente. El grupo AA (Figura 11e) asocia las CAZymas secretadas en pequeños grupos normalmente pertenecientes a genes homólogos de los tres aislados. En el árbol que representa aquellos genes que codifican CAZymas con módulos CBM (Figura 11f) se ha añadido un tercer anillo donde se hace referencia al grupo al que pertenece cada CAZyma. La mayor parte pertenece a el grupo GH, aunque se pueden apreciar dos grandes bloques de genes de CAZymas del grupo AA, otros dos de menos tamaño del grupo CE y un pequeño grupo PL. El grupo GT solo cuenta con 3 genes, uno de cada aislado, homólogos entre sí, con dichos módulos (Tabla Suplementaria S5).

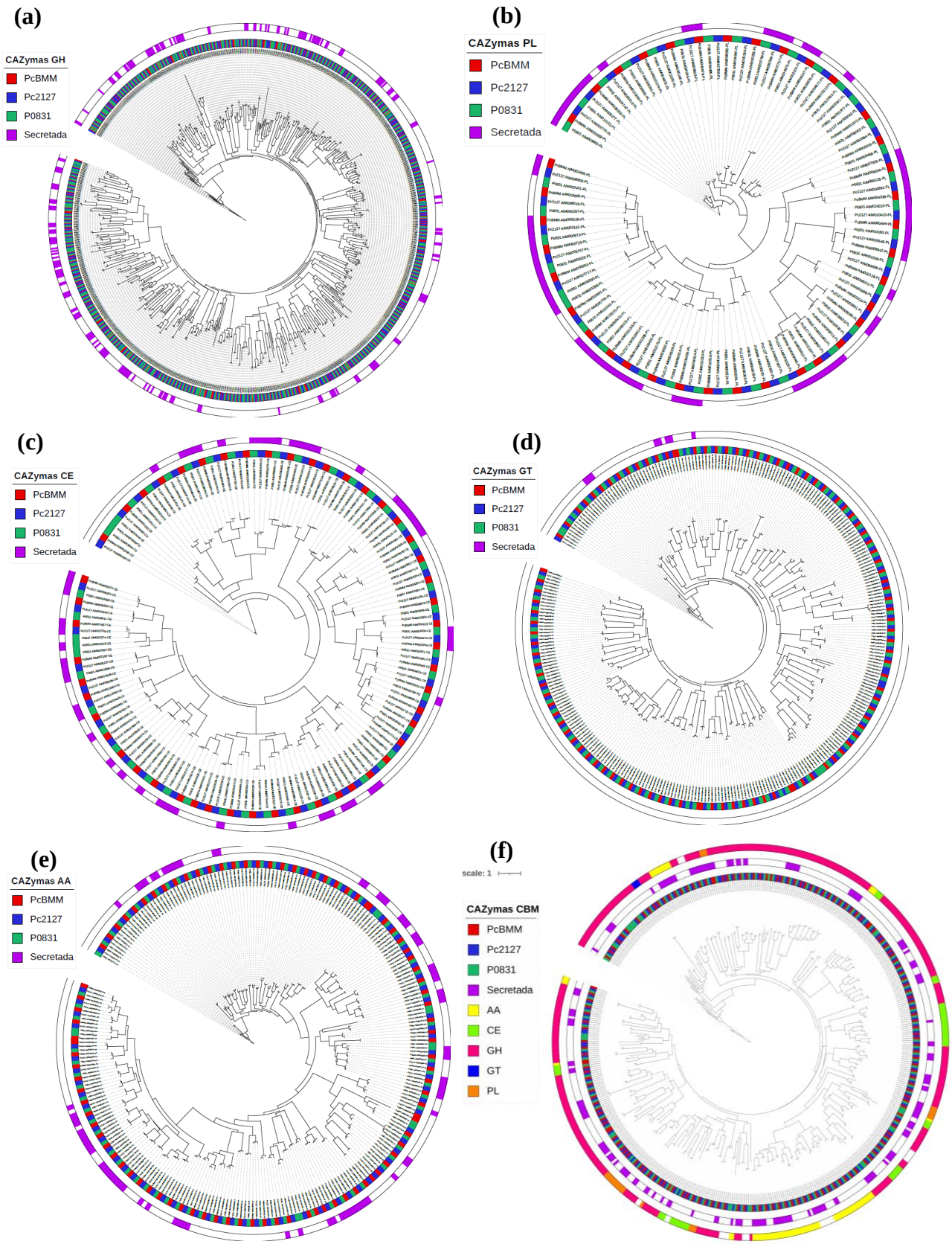


Figura 11: Arboles filogenéticos de los grupos de CAZymas **(a)** glucósido hidrolasas [GH], **(b)** polisacáridos liasas [PL], **(c)** esterasas de carbohidratos [CE], **(d)** glucosiltransferasas [GT], **(e)** enzimas para las actividades auxiliares [AA] y **(f)** enzimas con módulos CBM, etiquetadas el aislado a la que pertenecen y si son secretadas o no. La agrupación se realizó utilizando el criterio de máxima verosimilitud (ML) eligiendo el modelo de sustitución aminoacídica óptimo para cada árbol con la herramienta *-m TEST* de IQ-TREE y un bootstrap *-b 1000*.

4. CAPÍTULO 4: DISCUSIÓN Y CONCLUSIONES

4.1 Búsqueda de micovirus en datos de RNAseq

La búsqueda de micovirus en los datos de RNAseq de los tres aislados de *Plectosphaerella* de este trabajo no ha arrojado resultados positivos. En realidad, la presencia de micovirus en los aislados individuales de hongos utilizados en los laboratorios no es algo frecuente. En el trabajo de Gilbert, K., *et al* (2019) [9], que utiliza el protocolo de análisis de RNAseq similar al utilizado en nuestro trabajo, identificaron 1067 contigs en 284 BioProjects, que correspondían únicamente con 59 virus de 44 hongos diferentes [9]. En el mismo estudio incluyeron datos de RNAseq de un aislado de *Plectosphaerella*, donde tampoco encontraron micovirus, lo que indica que en esta especie no aparecen micovirus con frecuencia. En cualquier caso, se deben probar nuevas estrategias o buscar en aislados distintos, ya que aislados de campo de una especie pueden contener en algunos casos micoviromas distintos. Por ejemplo, Ruiz-Padilla, A, *et al*, (2021) [10], utilizando el método en el que se ha basado nuestro estudio, encontraron distintos micovirus en diferentes aislados de *Botrytis* obtenidos de plantas vid.

En el artículo de Gilbert, K., *et al*, (2019) [9] encontraron una secuencia de un virus de planta que afecta a *Brassicaceae*, el *Turning Ringsot Virus* (TRV), en datos de RNAseq de *C. tofieldiae*. Aunque hemos podido reproducir el resultado realizando el procesado de una manera similar al utilizado por Gilbert, K., *et al.* (2019) [9], no ha sido posible encontrar este virus usando el método adaptado de Ruiz-Padilla, A., *et al* (2021) [10]. No haberlo detectado con este método hace plantearse si las diferencias en la ejecución con otros métodos afectan al resultado. Para comprobar en qué paso se habían podido eliminar las secuencias homólogas a TRV, realizamos un BLAST-X de los ensamblados de *Trinity* de *C. tofieldiae* obtenidos con el método de Ruiz-Padilla, A., *et al* (2021) [10] contra las poliproteínas del TRV y contra la secuencia del virus TRV encontrada con el método de Gilbert, K., *et al*, (2019) [9]. En este caso sí identificamos una secuencia viral con un 50% de identidad a la secuencia del TRV del NCBI y un 60% con la homóloga encontrada con el método de Gilbert, K., *et al*, (2019) [9]. Esto indica que los procesados iniciales no eliminaron la secuencia homóloga a TRV. Sin embargo, no lo encontramos al realizar el mismo BLAST-X utilizando los archivos obtenidos tras el primer filtrado con la base de datos de virus personalizada, por lo que el paso limitante en el método propuesto por Ruiz-Padilla, A., *et al*, (2021) [10] está en dicha base de datos, que podría estar dejando fuera a virus realmente presentes en las muestras. Los resultados obtenidos indican que la base de datos debería ser ampliada o actualizada constantemente. Por otro lado, el método de Gilbert, K., *et al*, (2019) [9] es menos restrictivo y da un mayor número de resultados, pero también puede generar más falsos positivos. En conclusión, los métodos son complementarios, y este trabajo nos plantea la necesidad de buscar un tercer método que combine ambos y mejore en sensibilidad y fiabilidad la búsqueda de micovirus utilizando datos de RNAseq. En todo caso,

sería recomendable realizar la búsqueda de micovirus de *Plectosphaerella* utilizando el método de Gilbert, K, *et al.*, (2019) [9], que no hemos podido realizar por falta de tiempo.

4.2. Presencia de posibles motivos miméticos de péptidos fitorreguladores

En este trabajo, hemos encontrado 5 genes con suficiente homología con los péptidos SCOOP como para considerarlos buenos candidatos SCOOP-LIKE. En general, los motivos encontrados fueron más similares a los SCOOP-LIKE de *Fusarium* que a los SCOOP de *Arabidopsis*. Dentro de los motivos SCOOP-LIKE, se sabe que los residuos conservados SXS son necesarios para que presenten actividad asociada al reconocimiento por MIK2 [16 - 17] y, los residuos AGG en el extremo C-terminal están altamente conservados, por lo que también podrían tener algún significado biológico. En este trabajo hemos encontrado estos residuos conservados en la traducción de los genes de *Plectosphaerella* PcBMM_AIM001598 y su homólogo Pc2127_AIM009282, y los residuos conservados GGR en el gen P0831_AIM002183. Estos tres genes parecen ser los candidatos más convincentes como precursores de péptidos miméticos SCOOP-LIKE de *Plectosphaerella*, aunque habría que sintetizarlos y comprobar su actividad en el laboratorio. El gen Pc2127_AIM010202 presenta una identidad de secuencia muy alta con SCOOP6#2, que desencadena la producción de ROS. Sin embargo, no se encontraron los motivos conservados que activan la respuesta inmune, por lo que se desconoce su efecto en la interacción hongo-planta, o si habrá alguno. El gen P0831_AIM006279 se seleccionó por su alta homología con el SCOOP-LIKE de *Fusarium oxysporum 47*, pero no contiene ninguno de los motivos destacables, por lo que no se espera que desencadene una respuesta de la planta.

Una característica importante a tener en cuenta para predecir la relevancia biológica de los candidatos encontrados es su expresión en planta. Hemos analizado la expresión de estos genes a partir de los datos obtenidos de experimentos anteriores en los que se estudió la expresión de los genes de los tres aislados en plantas de Arabidopsis a tiempos tempranos post inoculación [1]. Los tres genes presentes en PcBMM y Pc2127 se sobreexpresan *in planta* respecto a *in vitro*, especialmente el gen de Pc2127 homólogo a SCOOP6#2, que se expresa 3 veces más que *in vitro* a 10 hpi tanto en las plantas *wild-type* (Col-0), como en los mutantes inmunodeprimidos *cyp79b2b3*. Es remarcable que deje de expresarse en Col-0, donde no es patogénico, a 16 hpi, mientras que en *cyp79b2b3*, donde produce enfermedad, la expresión se mantiene. Ninguno de los dos genes encontrados en P0831 mostraron expresión *in planta*. Cabe destacar que solo unos pocos genes de P0831 presentaron algún tipo de expresión *in planta* superior a *in vitro* [1] por lo que es muy difícil encontrar un gen P0831 cuya expresión relativa supere 0,5.

Además de la familia SCOOP, que está ya caracterizada [12, 16, 17], nos propusimos investigar la presencia de otros posibles SSPs de *Arabidopsis*. En concreto, investigamos la presencia de un conjunto de péptidos que, en trabajos anteriores, han evidenciado una posible relación con la

defensa de *Arabidopsis* frente a PcBMM y otros patógenos [13]. Sin embargo, estos péptidos de *Arabidopsis* no han sido caracterizados aún, y desconocíamos la existencia de motivos que puedan ser relevantes. Por ello, hemos buscado *de novo* cualquier motivo en *Plectosphaerella* homólogo a los putativos SSPs de *Arabidopsis.* Para obtener más información sobre la posible relevancia de los motivos conservados, hicimos una búsqueda previa de homólogos en *Arabidopsis*. Sin embargo, la inclusión de los homólogos de *Arabidopsis* encontrados no mejoró la detección motivos conservados, por lo que, en este caso, no parece que se trate de familias génicas como ocurre con SCOOP. De entre los 185 motivos encontrados, hemos propuesto como principal candidato el gen PcBMM_AIM008393, que coincide con SSP10 en 6 aminoácidos de 10 que constituyen el motivo. Este gen además se sobreexpresa en Col-0 10 hpi y *cyp79b2b3* 10 hpi, siendo superior la expresión en el mutante. Cabe destacar que el motivo está presente en genes homólogos en Pc2127 y P0831 (Pc2127_AIM005111 y P0831_AIM005463), pero en estos casos no hay expresión en planta. En los tres casos, el motivo se encuentra situado en el extremo C-terminal del gen (Figura Suplementaria 1). Lo más interesante de este gen es que funciona como regulador de la iniciación de la traducción, y tiene función ATPasa, por lo que sería interesante sintetizarlo y ver qué efecto tiene su aplicación a la planta.

4.3. Análisis de las CAZymas presentes en tres aislados de *Plectosphaerella*

El número de CAZymas detectadas fue muy similar al encontrado en otros hongos asociados a plantas con diferentes estilos de vida [20]. Hay que tener en cuenta que en nuestro caso hemos sido muy restrictivos, ya que analizado los genomas con dos herramientas diferentes [19, 36] y nos hemos quedado con los que han dado positivo con las dos herramientas.

En general, no se observan grandes diferencias en la cantidad de CAZymas entre los tres aislados, ni en el porcentaje de estas que son secretadas. Sin embargo, hay algunas instancias que sí difieren entre aislados que pueden ser candidatas a análisis más detallados en el futuro. La subclasificación de los genes de CAZymas según fuesen o no secretadas realizada con SECRETOOL [37] mostró que la distribución de CAZymas secretadas es diferente según el grupo al que pertenece. Los grupos con un mayor número de genes presentaron una distribución de CAZymas secretadas más dispersa, mientras que, a menor número de genes, las CAZymas secretadas tendieron a agruparse en zonas adyacentes del árbol. Así, el porcentaje de CAZymas secretadas con respecto del total osciló alrededor del 24% en el caso de GH y entre el 45-55% en el caso de PL, mientras que en la familia GT solo entre un 2-3% de las CAZymas son secretadas.

Los módulos CBM pueden tener importancia en la interacción de los hongos con las plantas, ya que pueden secuestrar quitina y otras moléculas derivadas que de otra manera serían reconocidos como PAMPs por el sistema inmune de la planta [20]. Los árboles filogenéticos también mostraron que la distribución de los módulos CBM es homogénea en los tres aislados, pero son

más abundantes en aquellas CAZymas que son secretadas. Esto se aprecia sobre todo en la débil presencia de la familia GT en el árbol de CBM, contando solo con tres hojas del árbol, correspondiente a un gen homólogo en los tres aislados que codifica para CAZymas secretadas. Por lo tanto, se puede afirmar que, a mayor número de CAZymas secretadas, mayor número de módulos CBM.

Durante la realización de estos análisis nos percatamos de que había grandes diferencias en la proporción de CAZymas secretadas con los resultados obtenidos por Muñoz-Barrios, A., *et al* (2020) [1]. Estas diferencias no se explicaban por la utilización de una versión más avanzada de dbCAN [40] en nuestro análisis. Tras contrastar ambos resultados hemos concluido que las discrepancias se deben a un error en la clasificación de las CAZymas como secretadas o no en la base de datos generada en Muñoz-Barrios, A., *et al.* (2020) [1], ya que el análisis del secretoma utilizando la herramienta SECRETOOLS fue idéntica en ambos trabajos. Gracias a este trabajo hemos podido descubrir este error y actualmente estamos trabajando en la publicación de una Fe de Erratas de la publicación para corregirlo.

4.4 Conclusiones

Las conclusiones a las que llega este trabajo son las siguientes:

1. No se encontraron micovirus con el método adaptado de Ruiz-Padilla, A., *et al,* (2021) en los transcriptomas de *Plectosphaerella spp.* y *Colletotrichum tofieldiae*, pero sí en *C.tofieldiae* con el método complementario de Gilbert, K., *et al* (2019), por lo que ambos métodos son complementarios entre sí y deberían utilizarse ambos o una fusión de estos para aumentar la eficiencia y precisión en la búsqueda de micovirus.
2. Se han encontrado cinco genes en los genomas de *Plectosphaerella* con motivos homólogos a los péptidos SCOOP. De estos, los genes homólogos de PcBMM y Pc2127 tienen expresión en planta a tiempos cortos, por lo que podrían tener actividad en la planta, aunque es necesario realizar los ensayos correspondientes para comprobarlo.
3. Se han detectado 185 motivos homólogos a putativos SSPs de *A.thaliana* en los genomas de los tres aislados de *Plectosphaerella* analizados. La expresión diferencial, el número de residuos conservados y la función del gen donde se encuentra el motivo hacen del gen PcBMM_AIM008393 el principal candidato como posible mimético de la actividad de estos péptidos.
4. La cantidad de CAZymas totales y secretadas es homogénea en los tres aislados de *Plectosphaerella* analizados, aunque se ha observado distribución desigual según el grupo funcional.
5. Existe una relación entre la presencia de módulos CBM en los genes de *Plectosphaerella* y si la CAZymas para la que codifica es o no secretada.

5. CAPÍTULO 5: BIBLIOGRAFÍA

[1] Muñoz-Barrios A, Fernández V, San Felipe L, Díaz S, Sopeña S, González-Melendi P, Molina A, Sacristán S (2020). Differential Expression of Fungal Genes Determines the Lifestyle of Plectosphaerella Strains During Arabidopsis thaliana Colonization. Molecular Plant-Microbe Interactions. https://apsjournals.apsnet.org/doi/full/10.1094/MPMI-03-20-0057-R

[2] Ramos, B., González-Melendi, P., Sánchez-Vallet, A., Sánchez-Rodríguez, C., López, G., & Molina, A. (2012). Functional genomics tools to decipher the pathogenicity mechanisms of the necrotrophic fungus Plectosphaerella cucumerina in Arabidopsis thaliana. *Molecular Plant Pathology*, 14(1), 44–57. https://doi.org/10.1111/j.1364-3703.2012.00826.x

[3] Sanchez-Vallet, A., Ramos, B., Bednarek, P., López, G., Piślewska-Bednarek, M., Schulze-Lefert, P., & Molina, A. (2010). Tryptophan-derived secondary metabolites in Arabidopsis thaliana confer non-host resistance to necrotrophic Plectosphaerella cucumerina fungi. The Plant Journal, no. https://doi.org/10.1111/j.1365-313x.2010.04224.x

[4] Abbas, A. (2016). A Review Paper on Mycoviruses. Journal of Plant Pathology & Microbiology, 7(12). https://doi.org/10.4172/2157-7471.1000390

[5] Ghabrial, S. A., & Suzuki, N. (2009). Viruses of Plant Pathogenic Fungi. Annual Review of Phytopathology, 47(1), 353–384. https://doi.org/10.1146/annurev-phyto-080508-081932

[6] Nuss, D. L. (2011). Mycoviruses, RNA Silencing, and Viral RNA Recombination. Advances in Virus Research, 25–48. https://doi.org/10.1016/b978-0-12-385987-7.00002-6

[7] Zhang, H., Xie, J., Fu, Y., Cheng, J., Qu, Z., Zhao, Z., Cheng, S., Chen, T., Li, B., Wang, Q., Liu, X., Tian, B., Collinge, D. B., & Jiang, D. (2020). A 2-kb Mycovirus Converts a Pathogenic Fungus into a Beneficial Endophyte for Brassica Protection and Yield Enhancement. *Molecular Plant*, *13*(10), 1420–1433. https://doi.org/10.1016/j.molp.2020.08.016

[8] Marquez, L. M., Redman, R. S., Rodriguez, R. J., & Roossinck, M. J. (2007). A Virus in a Fungus in a Plant: Three-Way Symbiosis Required for Thermal Tolerance. *Science*, *315*(5811), 513–515. https://doi.org/10.1126/science.1136237

[9] Gilbert, K. B., Holcomb, E. E., Allscheid, R. L., & Carrington, J. C. (2019). Hiding in plain sight: New virus genomes discovered via a systematic analysis of fungal public transcriptomes. *PLOS ONE*, *14*(7), e0219207. https://doi.org/10.1371/journal.pone.0219207

[10] Ruiz-Padilla, A., Rodríguez-Romero, J., Gómez-Cid, I., Pacifico, D., & Ayllón, M. A. (2021). Novel Mycoviruses Discovered in the Mycovirome of a Necrotrophic Fungus. mBio. Published. https://doi.org/10.1128/mbio.03705-20

[11] Vanyushin, B. F., Ashapkin, V. V., & Aleksandrushkina, N. I. (2017). Regulatory peptides in plants. Biochemistry (Moscow), 82(2), 89–94. https://doi.org/10.1134/s0006297917020018

[12] Gully, K., Pelletier, S., Guillou, M. C., Ferrand, M., Aligon, S., Pokotylo, I., Perrin, A., Vergne, E., Fagard, M., Ruelland, E., Grappin, P., Bucher, E., Renou, J. P., & Aubourg, S. (2019). The SCOOP12

peptide regulates defense response and root elongation in Arabidopsis thaliana. Journal of Experimental Botany, 70(4), 1349–1365. https://doi.org/10.1093/jxb/ery454

[13] Sopeña-Torres, S., Jordá, L., Sánchez-Rodríguez, C., Miedes, E., Escudero, V., Swami, S., López, G., Piślewska-Bednarek, M., Lassowskat, I., Lee, J., Gu, Y., Haigis, S., Alexander, D., Pattathil, S., Muñoz-Barrios, A., Bednarek, P., Somerville, S., Schulze-Lefert, P., Hahn, M. G., . . . Molina, A. (2018). YODA MAP3K kinase regulates plant immune responses conferring broad-spectrum disease resistance. New Phytologist, 218(2), 661–680. https://doi.org/10.1111/nph.15007

[14] Ojito-Ramos, K., Portal, O. (2010). Introducción al sistema inmune vegetal. Biotecnología Vegetal Vol. 10, No. 1: 3 – 19. https://revista.ibp.co.cu/index.php/BV/article/view/266

[15] Yuan, N., Furumizu, C., Zhang, B., & Sawa, S. (2021). Database mining of plant peptide homologues. Plant Biotechnology, 38(1), 137–143. https://doi.org/10.5511/plantbiotechnology.20.0720a

[16] Hou, S., Liu, D., Huang, S., Luo, D., Liu, Z., Wang, P., Mu, R., Han, Z., Chai, J., Shan, L., & He, P. (2021). Immune elicitation by sensing the conserved signature from phytocytokines and microbes via the Arabidopsis MIK2 receptor. PLOS ONE. Published. https://doi.org/10.1101/2021.01.28.428652

[17] Rhodes, J., Yang, H., Moussu, S., Boutrot, F., Santiago, J., & Zipfel, C. (2021). Perception of a divergent family of phytocytokines by the Arabidopsis receptor kinase MIK2. Nature Communications, 12(1). https://doi.org/10.1038/s41467-021-20932-y

[18] Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Research, 37(Database), D233-D238. https://doi.org/10.1093/nar/gkn663

[19] Barrett, K., Hunt, C. J., Lange, L., & Meyer, A. S. (2020). Conserved unique peptide patterns (CUPP) online platform: peptide-based functional annotation of carbohydrate active enzymes. Nucleic Acids Research, 48(W1), W110-W115. https://doi.org/10.1093/nar/gkaa375

[20] Hacquard, S., Kracher, B., Hiruma, K., Münch, P. C., Garrido-Oter, R., Thon, M. R., Weimann, A., Damm, U., Dallery, J. F., Hainaut, M., Henrissat, B., Lespinet, O., Sacristán, S., Ver Loren van Themaat, E., Kemen, E., McHardy, A. C., Schulze-Lefert, P., & O'Connell, R. J. (2016). Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. *Nature Communications*, *7*(1). https://doi.org/10.1038/ncomms11362

[21] Chiapello, M., Rodríguez-Romero, J., Ayllón, M. A., Turina, M. (2020). Analysis of the virome associated to grapevine downy mildew lesions reveals new mycovirus lineages. Virus Evolution, 6(2). https://doi.org/10.1093/ve/veaa058

[22] Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLOS ONE*, *12*(10), e0185056. https://doi.org/10.1371/journal.pone.0185056

[23] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., . . . Regev, A. (2011). Full-length transcriptome assembly

from RNA-Seq data without a reference genome. Nature Biotechnology, 29(7), 644–652. https://doi.org/10.1038/nbt.1883

[24] Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

[25] Huang, X. (1999). CAP3: A DNA Sequence Assembly Program. Genome Research, 9(9), 868–877. https://doi.org/10.1101/gr.9.9.868

[26] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), 10. https://doi.org/10.14806/ej.17.1.200

[27] Langmead, B., Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. https://doi.org/10.1038/nmeth.1923

[28] Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics, 11(1). https://doi.org/10.1186/1471-2105-11-431

[29] Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412-D419. https://doi.org/10.1093/nar/gkaa913

[30] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. BMC Bioinformatics, 10(1), 421. https://doi.org/10.1186/1471-2105-10-421

[31] Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009b). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research, 37(Web Server), W202-W208. https://doi.org/10.1093/nar/gkp335

[32] Sievers, F., & Higgins, D. G. (2017). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, *27*(1), 135–145. https://doi.org/10.1002/pro.3290

[33] Robert, X., & Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. Nucleic Acids Research, 42(W1), W320-W324. https://doi.org/10.1093/nar/gku316

[34] Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

[35] Rodrigo, J. A. (2021). Clustering y heatmaps: aprendizaje no supervisado. Clustering y heatmaps: aprendizaje no supervisado. https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps#Bibliograf%C3%ADa

[36] Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P. K., Xu, Y., & Yin, Y. (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Research, 46(W1), W95-W101. https://doi.org/10.1093/nar/gky418

[37] Cortázar, A. R., Aransay, A. M., Alfaro, M., Oguiza, J. A., & Lavín, J. L. (2013). SECRETOOL: integrated secretome analysis tool for fungi. Amino Acids, 46(2), 471–473. https://doi.org/10.1007/s00726-013-1649-z

[38] Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution, 37(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

[39] Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research. Published. https://doi.org/10.1093/nar/gkab301

[40] Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., & Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Research, 40(W1), W445-W451. https://doi.org/10.1093/nar/gks479

ANEXOS:

Tabla Suplementaria S1: Genes de PcBMM con homología a putativos SSPs de Arabidopsis utilizados para el estudio del análisis de expresión de la Figura 9 y 10

Gen PcBMM	SSP *Arabidopsis*	Residuos conservados	Anotación gen PcBMM
AIM001916	SSP3	M--SMK---	alpha beta fold partial
AIM002092	SSP4	--E---KT--	hypothetical protein VDAG_06439
AIM006016	SSP8.3	P-PPPP----	hypothetical protein UCDDA912_g08162
AIM004328	SSP5	N--FG-C	polysaccharide deacetylase
AIM008364	SSP6.2	--KGK-----	aconitate hydratase 1
AIM008201	SSP5	-Q-C--C--C	hypothetical protein BN1708_007015
AIM000156	SSP1.1	P-G--KA--- --C----FLH	aldehyde dehydrogenase
AIM007454	SSP8.1	---L-I--Q	ribosome biogenesis MAK21
AIM010172	SSP11	-SH-R P---N-S	---NA---
AIM002535	SSP8.1	---VC-I---	EAP30 Vps36 family
AIM009859	SSP10	-H-AAPH	Cellulase
AIM007909	SSP8.3	PP-PPPL--- -V--LQ--M	alpha beta-glucosidase agdC
AIM008393	SSP10	LR-GFFA---	translation initiation regulator
AIM003702	SSP8.3	P-P--PLP--	aspartic ase
AIM011005	SSP7	M--L---V-- ----Y-RL-G	ABC transporter
AIM003089	SSP8.3	-PPPP-----	hnrnp arginine n-methyltransferase
AIM003204	SSP12	-PNLP-PP-- -SPP-H	hypothetical protein BN1708_015069
AIM009621	SSP7	-------GNE	glycoside hydrolase family 75
AIM004866	SSP4	--K-L-Q-V	hypothetical protein BN1708_006063

Tabla Suplementaria S2: Genes de P0831 con homología a putativos SSPs de Arabidopsis utilizados para el estudio del análisis de expresión de la Figura 9 y 10

Gen P0831	Arabidopsis SSP	Residuos conservados	Anotación gen P0831
AIM009039	SSP8.3	---PPPL---	hypothetical protein CH063_05422
AIM009821	SSP14.1	PIP----PTL	Transcription factor jumonji [Metarhizium robertsii ARSEF 23]
	SSP14.2	MA--P--SP IP-----PTLS -------LFL	
AIM001836	SSP14.2	I-----P-LS ----F---FL	C6 zinc finger

Tabla Suplementaria S3: Genes de Pc2127 con homología a putativos SSPs de Arabidopsis utilizados para el estudio del análisis de expresión de la Figura 9 y 10

Gen Pc2127	Arabidopsis SSP	Residuos conservados	Función gen Pc2127
AIM004951	SSP5	N--FG-C	polysaccharide deacetylase
AIM006549	SSP9	MA---V----	dehydratase
AIM007727	SSP8.3	P-PPPP---M	stage V sporulation K
AIM006554	SSP8.3	P-P--PLP--	aspartic ase
AIM000922	SSP7	M--L---V--	ABC transporter
		----Y-RL-G	
AIM003223	SSP8.3	PPPPP-L---	bZIP transcription factor
		HV---Q-QM	
		I--A-H-Q-	
AIM002271	SSP8.3	P--PP-L---	hypothetical protein BN1723_002555
AIM000735	SSP4	--K-L-Q-V	hypothetical protein BN1708_006063
		Y-E--T----	
		------T-EE	
AIM007279	SSP6.2	--KGK-----	myosin class ii heavy chain
AIM010795	SSP1.1	P-G--KA---	aldehyde dehydrogenase
		--C----FLH	
AIM001564	SSP8.1	---VC-I---	EAP30 Vps36 family
AIM006218	SSP9	M-HHH	glutamate synthase
		-FIKN	
AIM004417	SSP8.3	PP-PPPL---	alpha beta-glucosidase agdC
		-V--LQ--M	
AIM003574	SSP8.1	F-A-C----V	fungal specific transcription factor domain-containing
AIM002336	SSP8.3	P-P-P-L---	hypothetical protein BN1723_004292, partial
AIM005985	SSP3	M--SMK---	alpha beta fold partial
AIM006209	SSP14.2	RSV--E-A-	C6 zinc finger
		I-----P-LS	
AIM002806	SSP2	----GNM--C	glycosyl hydrolase
AIM001939	SSP4	--E---KT--	acetyl- carboxylase
		-Y---V--L	
AIM004894	SSP8.1	---L-I--Q	ribosome biogenesis MAK21

Tabla Suplementaria S4: Número de CAZymas totales y secretadas (SEC) por grupo y aislado de *Plectosphaerella*. Entre paréntesis se indica el porcentaje de CAZymas secretadas sobre el total.

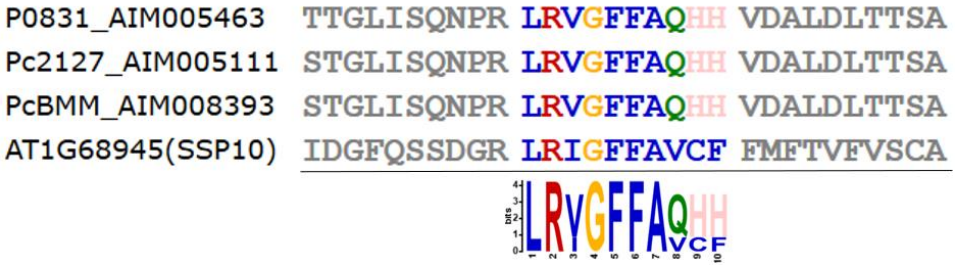
Familia	PcBMM	Pc2127	P0831
AA	91	89	88
AA SEC	29 (31.87%)	26 (29.21%)	24 (27.27%)
CE	50	50	48
CE SEC	17 (34%)	13 (26%)	20 (41.67%)
GH	305	303	294
GH SEC	74 (24.26%)	73 (24.09%)	69 (23.50%)
GT	101	105	103
GT SEC	3 (2.97%)	3 (2.86%)	2 (1.94%)
PL	37	38	37
PL SEC	20 (54.05%)	21 (55.26%)	17 (45.94%)
CBM	156	157	151
CBM SEC	57 (36.53%)	57 (36.30%)	49 (32.45%)

Tabla Suplementaria S5: Distribución de módulos CBM según familia de CAZymas y aislado de *Plectosphaerella*. Entre paréntesis se indica el porcentaje de CAZymas con módulos CBM sobre el total (Tabla Suplementaria S4).

Familia	PcBMM	Pc2127	P0831
AA	21 (23.07%)	21 (23.6%)	23 (26.17%)
CE	19 (38%)	18 (36%)	17 (35.42%)
GH	102 (33.44%)	103 (33.99%)	98 (33.33%)
GT	1 (0.99%)	1 (0.95%)	1 (0.97%)
PL	7 (18.91%)	8 (21.05%)	7 (18.92%)
	150	151	146

Figura Suplementaria 1: Alineamiento del motivo común a SSP10 de Arabidopsis en los genes homólogos de los tres aislados de *Plectosphaerella.* y logo obtenido con la herramienta MEME (ref) **(a)** y posición del motivo en cada uno de los genes **(b)**

(a)



(b)