Facultad de Ciencias Experimentales

Grado en Biología

UNIVERSIDAD DE JAÉN
Facultad de Ciencias Experimentales

Análisis transcriptómico mediante RNAseq en el olivo

Almudena García Consuegra Ruiz de la Hermosa

Septiembre, 2014

Facultad de Ciencias Experimentales

Grado en Biología

UNIVERSIDAD DE JAÉN
Facultad de Ciencias Experimentales

Análisis transcriptómico mediante RNAseq en el olivo

Almudena García Consuegra Ruiz de la Hermosa

Septiembre, 2014

1. RESUMEN

El olivo es considerado uno de los árboles frutales más importantes a nivel mundial, por su alto valor económico y nutricional. Los cultivos y sus productos son dañados por numerosas enfermedades y plagas, siendo la patología más importante la Verticilosis causada por el hongo del suelo *Verticillium dahliae*. Determinadas variedades de esta planta responden de forma diferencial al parásito, siendo algunas de ellas más resistentes, como es el caso de la variedad “Frantoio”, y otras más susceptibles como la variedad “Picual”. El estudio de la interacción entre el parásito y la planta a través de nuevas herramientas transcriptómicas como el RNA-seq, constituyen el principal objetivo de la Fitopatología. Ésta permite analizar los niveles de expresión de miles de transcritos en una gran diversidad de situaciones. En el presente trabajo se ha ensamblado un transcriptoma completo que permitirá un amplio estudio de la interacción planta-patógeno en una variedad tolerante y en una susceptible.

ABSTRACT

Olive tree is considered one of the most important fruit trees worldwide, mainly for its high economic and nutritional value. Crops and its products are damaged by many diseases and pest, being the most important disease Verticillium wilt, caused by the soil fungus *Verticillium dahliae*. Certain varieties of this plant respond differentially to the parasite, some of which are more resistant, such as the cultivar “Frantoio”, and more sensitive as the cultivar “Picual”. The study of the interaction between the parasite and the plant through new transcriptomic tools, such as RNA-seq, is the main objective of Phytopathology. This allows to analyze the expression levels of thousands of transcripts in a wide variety of situations. In the present work a complete transcriptome was assembled, which will allow a large study of the plant-pathogen interaction in both cultivars, tolerant and sensitive.

2. INTRODUCCIÓN.

2.1. El olivo (*Olea europaea L.*)

El olivo (*O. europaea L*.), es uno de los árboles frutales más antiguos cultivados en la Cuenca Mediterránea por su alto valor económico (Dominguez-Garcia, Belaj et al. 2012). Pertenece a la familia botánica *Oleaceae*, la cual comprende especies de plantas distribuidas por las regiones tropicales y templadas del mundo (Barranco 2008). Se clasifica dentro del género *Olea* dónde se incluyen todas las variedades de olivo cultivado (*O. europaea* ssp*. europaea* var*. europaea*), junto con los acebuches u olivos silvestres (*O. europaea* ssp*. europaea* var*. sylvestris*).

En cuanto a su estructura vegetativa, el olivo es un árbol perennifolio con un tamaño mediano comprendido entre los 4 u 8 metros de altura dependiendo de la variedad, llegando en ocasiones a alcanzar hasta los 15 m. El tronco es grueso, a menudo corto y retorcido, presentando la corteza un color gris o verde grisáceo. La ramificación natural tiende a producir una copa bastante densa, pero las diversas prácticas de poda sirven para aclararla y permitir el paso de la luz, por lo que en general, es redondeada, más o menos lobulada. Además, la forma del árbol está influida en cierta medida por las condiciones agronómicas y ambientales de su crecimiento y en particular, por el tipo de poda, presentando, una gran plasticidad morfogenética. (Fig. 1.)

Figura 1. Olivo centenario.

Las hojas del olivo son opuestas de forma lanceolada con bordes enteros y una longitud comprendida entre 3 y 9 cm con una anchura entre 1 y 1,8 cm. El peciolo es muy corto, llegando apenas a medio centímetro de longitud, apareciendo a menudo opuestas. Por el haz, la superficie superior, las hojas son de color verde-oscuro y brillan debido a la presencia de una cutícula gruesa. El envés, la superficie inferior, tiene un color blanco-plateado ya que se encuentra cubierto por pelos aparasolados.
(Fig. 2)

Figura 2: 2a) Arranque de dos hojas opuestas. Figura 2b) Detalle de hoja de olivo de la variedad "Picual".

Las flores son pequeñas y actinomorfas, con simetría regular y se localizan en las ramificaciones de las inflorescencias de forma aislada o formando grupos de tres o cinco (Barranco 2008). La corola es blanca cuyo período de floración está comprendido entre mayo y julio.

El olivo es la única especie de la familia *Oleaceae* con fruto comestible, la aceituna, el cual es pequeño de forma elipsoidal a globosa. Mide de 1 a 4 cm de longitud y de 0,6 a 2 cm de diámetro. En su madurez, la aceituna presenta un color negro-violáceo, pero en muchos casos se cosecha antes, cuando todavía está verde (Fig. 3). En términos botánicos, la aceituna es una drupa con una sola semilla compuesta por tres tejidos principales: endocarpo, que es el más interior o hueso; el mesocarpo, es la pulpa o carne y el exocarpo, piel o capa exterior. El conjunto de estos tejidos se denomina pericarpo, y tiene su origen en la pared del ovario. Los tejidos del fruto se desarrollan del ovario por los procesos de división, expansión y diferenciación celular, a partir de la fecundación y del cuajado inicial (Barranco 2008).

Figura 3. Aceituna verde, aún sin madurar (izquierda). Aceituna de "Picual" madurando en el olivo presentando el color negro característico a las faldas del Alcázar de Santa Catalina en Jaén (derecha).

El olivo es una de las plantas cultivadas por el hombre más antiguas, cuyos orígenes se datan hace 3.000-4.000 años a. C. en la zona de Palestina, aunque debido a su difusión el 95% del área mundial cultivada en la actualidad se encuentra en la Cuenca Mediterránea (Barranco 2008) y se halla en expansión en áreas de Australia, América (Argentina, Chile, Estados Unidos) y Sudáfrica (Rugini and Fedeli 1990). Su cultivo se concentra entre las latitudes 30º y 45º, tanto en el hemisferio norte como en el sur, en regiones climáticas del tipo Mediterráneo, caracterizadas por un verano seco y caluroso y un invierno húmedo y templado. El patrimonio oleícola a nivel mundial se estima aproximadamente en 1.000 millones de olivos, ocupando una superficie de aproximadamente 10 millones de hectáreas. Los países de la Cuenca Mediterránea comprenden el 98% del total (Fig. 4), 1,2 % se sitúa en el continente americano, 0,4% en Asia Oriental y 0,4% en Oceanía.(Barranco 2008).

La mayor parte de los cultivos provienen de países del sur de Europa: como Italia, que cuenta con 538 variedades diferentes; España (183), Francia (88) y Grecia (52), representando esta biodiversidad una rica fuente de variabilidad para la mejora genética de esta planta (Baldoni and Belaj 2010).

Figura 4. Distribución del olivar en los diferentes países de la Cuenca Mediterránea.

Dada la importancia del cultivo del olivo en toda la cuenca del Mediterráneo, el aceite de oliva constituye un componente básico en la dieta de los países que lo bordean.

PRODUCCIÓN DE ACEITE DE OLIVA POR ZONAS GEOGRÁFICAS (Miles T.)			
ZONA	Camp. 11/12	Camp. 12/13	Camp. 13/14 (est.)
MUNDO	3.321	2.425	3.098
U.E.	2.395	1.459	2.308
ESPAÑA	1.615	616,3	1.607
ANDALUCÍA	1.363	514	1.312
JAÉN	682	170	715
COMPARATIVA POR ZONAS GEOGRÁFICAS			
U.E./MUNDO	72%	60%	75%
ESPAÑA/U.E.	67%	42%	70%
ANDALUCÍA/ESPAÑA	84%	83%	82%
JAÉN/ANDALUCÍA	50%	33%	55%

Tabla 1. Producción mundial de aceite de oliva por campañas 2011/12, 12/13 y 13/14 (datos estimados), a nivel mundial, en Europa, España,Andalucía y Jaén (C.O.I.).

Existe una marcada concentración de la producción y el consumo de frutos y aceite de oliva en la cuenca Mediterránea, siendo de esta zona 18 de los 30 países con cultivos de olivar. Estos 18 países abarcan el 86% del consumo y el 98% de la producción a nivel mundial a finales de los noventa (Grigg 2001). España es uno de los primeros productores con un 51% de la producción mundial (Tabla 1), Andalucía abarca el 82-84% de la producción española y Jaén el 50-55% de la producción andaluza (denominada "Capital Mundial del Aceite de oliva") (C.O.I.).

Nos hemos centrado en la variedad "Picual" por ser la más valorada agronómicamente por su producción, su elevada calidad del aceite, un alto contenido en ácido oleico, la gran estabilidad del aceite y por ser la más importante y representada en el panorama olivarero peninsular, ocupando en Jaén el 97% de la superficie cultivada.

La larga tradición del cultivo del olivo en la región mediterránea ha quedado reflejada en la referencia histórica de sus enfermedades (Barranco 2008). El olivo y sus productos pueden ser dañados por numerosas enfermedades y plagas, pudiendo estar causadas por microorganismos como la bacteria *Pseudomonas savastanoi* (Fig. 6.1), la cual produce tuberculosis en ramas y tallos del árbol; hongos, como *Cycloconium oleaginum* (Fig. 6.2) que daña hojas y frutos, o *Verticillium dahliae,* afectando este último al crecimiento del árbol. Entre los fitófagos más perjudiciales, destacan *Bactrocera oleae,* la mosca del olivo (Fig. 6.3); y *Prays oleae*, la polilla del olivo, los cuales pueden llegar a causar graves daños económicos debidos a la pérdida de producción del olivo. Además, es susceptible a varios virus, estando más del 70% de los olivos cultivados afectados por virus latentes (Rugini, Mencuccini et al. 2005).

2.2. Infección por *Verticillium dahliae*.

Una de las enfermedades más importantes de dicho cultivo es la Verticilosis del olivo (Rodríguez Jurado, Blanco López et al. 1994). Actualmente es la patología que más preocupación causa a los olivicultores en España, especialmente en Andalucía, por la severidad de sus ataques y expansión en los últimos años (Rodriguez Jurado, Lopez et al. 1993, Blanco-López and Jiménez-Díaz 1995). Se diagnosticó por primera vez en Italia en los años 40 y

Figura 6. 1) Tuberculosis en la hoja de olivo producida como consecuencia de la bacteria *P. savastanoi*. 2) Daños producidos por el hongo *C. oleaginum*. 3) *B. oleae* (mosca del olivo)

posteriormente ha sido descrita en todos los países en los que el olivo es un cultivo de relevada importancia o incluso en aquellas áreas geográficas en las que su difusión está en auge. La enfermedad refleja una amplia distribución por países de la Cuenca Mediterránea como Grecia, Chipre, Turquía, Francia, España, etc. Dicha patología no solamente se encuentra muy extendida (Rodríguez Jurado, Blanco López et al. 1994), sino que también supone importantes pérdidas económicas que se ven aumentadas en nuevas plantaciones con sistemas de cultivo intensivo (Martos Moreno, Raya Ortega et al. 2005).

La Verticilosis se diagnosticó por primera vez España en el año 1975 en campos experimentales del CIDA en Córdoba (Blanco-López, Jiménez-Díaz et al. 1984). Posteriormente, se diagnosticó en el resto de Andalucía y otras regiones olivareras. En 2009, las estimaciones realizadas por la Red de Alerta e Información Fitosanitaria de Andalucía indicaban una incidencia media de 0,4% de olivos afectados, aunque existe una gran diferencia entre comarcas, llegándose a alcanzar en algunas más del 50% de campos afectados y una incidencia media del 9%. Aunque esta enfermedad se ha detectado en olivares

adultos, los ataques más graves se producen comúnmente en olivos jóvenes de 4 a 10 años localizados en zonas de regadío (Trapero, Escudero et al. 2011).

Figura 7. Síndrome de apoplejía en olivo.

Los primeros síntomas suelen aparecer a partir de los dos años de plantación, aunque pueden aparecer antes en función de la susceptibilidad del cultivar, de la cantidad y virulencia del patógeno existente en el suelo y de las condiciones ambientales (Rodríguez Jurado, Blanco López et al. 1994). La verticilosis del olivo se manifiesta como dos síndromes que suelen expresarse en diferentes épocas del año, la apoplejía y el decaimiento lento.

La **apoplejía** comprende la muerte agresiva, rápida y extensa de brotes y ramas de la planta (Fig. 7) (Trapero, Escudero et al. 2011). Es de desarrollo rápido y se manifiesta inicialmente por una pérdida del color verde intenso de las hojas (Rodríguez Jurado, Blanco López et al. 1994) que tienden a un color pajizo y al mismo tiempo se abarquillan. Este síndrome es más común a finales del invierno y principios de la primavera (Trapero, Escudero et al. 2011).

El **decaimiento lento** se caracteriza por la necrosis de las inflorescencias, las flores quedan momificadas y las hojas caen antes de secarse (Fig. 8). (Rodríguez Jurado, Blanco López et al. 1994). Estre grupo de síntomas aparece durante la primavera y en ciertas ocasiones tras el verano, con otoños cálidos siendo en este caso los frutos los que quedan secos.

Ambos síndromes aparecen a menudo en el mismo árbol aunque los síntomas frecuentemente se manifiestan de forma parcial en la planta. Salvo que los árboles mueran, las plantas enfermas suelen recuperarse en los años siguientes, es decir, se produce una recuperación natural de las infecciones (Rodríguez Jurado, Blanco López et al. 1994).

Figura 8. Desecación de ramas y momificación de flores (izquierda). Hojas de olivo afectadas que se caen antes de secarse (derecha).

El agente causal de esta enfermedad es el hongo hifomiceto *V. dahliae kleb.* Este patógeno del suelo se multiplica por esporas asexuales (conidios) producidas en conidióforos verticilados formando unas estructuras multicelulares de resistencia, denominadas microesclerocios (Fig. 9). Éstos son de color negro y de forma y tamaño variables que pueden permanecer en reposo en el suelo durante largos períodos de tiempo. El hongo es inespecífico de huésped y presenta una extensa gama de plantas susceptibles, tanto cultivadas como silvestres. Entre ellas destacan plantas hortícolas (alcachofa, berenjena, lechuga, patata, etc.), leguminosas (garbanzo, guisante, entre otras), industriales (algodón, girasol, tabaco, remolacha, etc.), omamentales (boj, clavel, rosal, crisantemo), frutales (aguacate, almendro, etc.), así como especies forestales (arce, castaño, fresno, olmo) y numerosas dicotiledóneas silvestres, muchas de las cuales forman parte de la flora arvense del olivar. Este hecho facilita el mantenimiento y la distribución del hongo en el suelo, resultando muy díficil la eliminación del mismo. En el caso del olivo, tanto las variedades cultivadas como el olivo silvestre o acebuche son susceptibles al patógeno (Trapero, Escudero et al. 2011).

El ciclo vital de *V. dahliae* comprende varias etapas: superviviencia y dispersión del inóculo, germinación de las estructuras de superviviencia, penetración del patógeno en la planta, colonización del sistema vascular, desarrollo de síntomas y producción de nuevo inóculo.

Es un patógeno monocícilico, es decir, solamente produce una generación de inóculo en una estación de crecimiento del huésped. La superviviencia del hongo se debe a los microesclerocios, ya que éstos son capaces de soportar condiciones ambientales y biológicas ad versas durante muchos años, hasta el momento de su germinación, estimulada por las sustancias emitidas por las raíces de las plantas huéspedes (Trapero, Escudero et al. 2011).

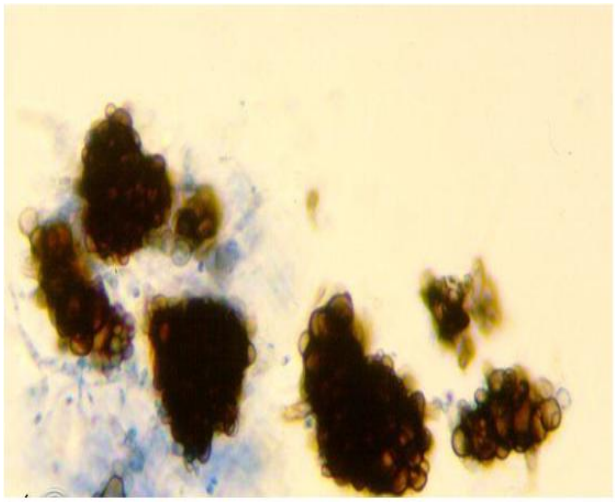

Figura 9. Estructuras de resistencia del hongo *V.dahliae*

La infección de la planta por parte de los microesclerocios del suelo se inicia a través de las raíces. Una vez penetrado en el córtex, el hongo se establece en el xilema dónde produce micelio y conidios colonizando de esta manera la planta de forma sistémica. Debido al pequeño tamaño de los conidios, el desplazamiento de éstos por el flujo xilemático se ve favorecido. Cuando la colonización alcanza un determinado nivel aparecen a desarrollarse los síntomas en la planta y el hongo sale del xilema y coloniza otros tejidos del huésped. Finalmente, el resultado es la formación de microesclerocios en todos los tejidos infectados de la planta. Cuando ha transcurrido el tiempo necesario para la muerte y posterior descomposición de los tejidos infectados, los microesclerocios libres o embebidos en los restos vegetales se incorporan al suelo, cerrándose el ciclo y quedando los microesclerocios preparados para un nuevo ciclo de infección (Fig. 10). El olivo no es el único huésped de *V. dahliae* en un olivar, por lo que en este ciclo hay que considerar la aportación de inóculo procedente de las plantas arvenses huéspedes del patógeno (Trapero, Escudero et al. 2011).

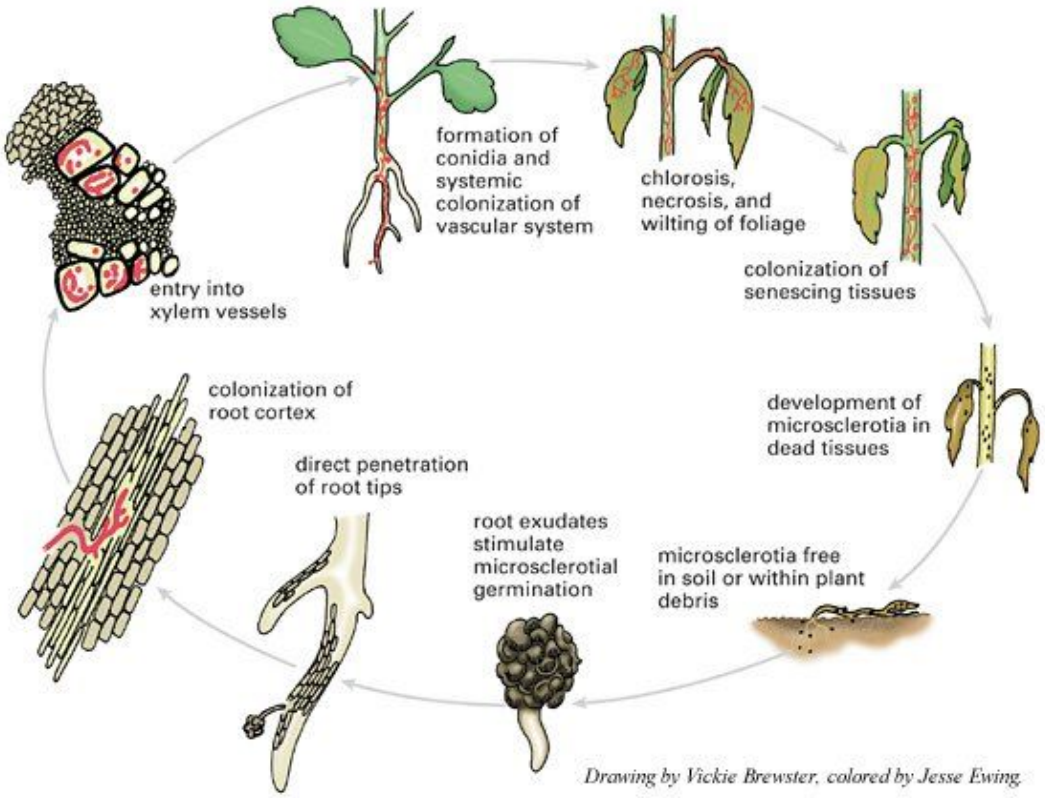

Figura 10. Ciclo de vida de *Verticillium* spp durante la infección de la planta. (http://www.apsnet.org/Education/lessonsPlantPath/Verticillium/discycle.htm).

La tasa de infección está determinada por varios factores dependientes del huésped (nivel de susceptibilidad, edad, nutrición, etc.), del patógeno (virulencia) y del ambiente (temperatura del aire, humedad, tipo de suelo, etc.). Aunque entre los factores ambientales favorables para la Verticilosis, destacan la humedad constante del suelo y la temperatura del aire proxima al óptimo térmico de crecimiento del patógeno (20-25ºC). Estas condiciones coinciden con las que favorecen el crecimiento vegetativo del olivo, por lo que la máxima actividad del patógeno ocurre en primavera y otoño coincidiendo con el crecimiento activo del árbol (Trapero, Escudero et al. 2011).

A pesar de ser estrictamente asexual, la especie *V. dahliae* posee una amplia diversidad genética en diferentes poblaciones que difieren entre sí en características morfológicas, fisiológicas y patogénicas. Al presentar una gran variabilidad genética tiene la posibilidad de encontrar fuentes de resistencia a la Verticilosis dentro de las variedades existentes. La resistencia de una planta de olivo consiste en la restricción de la infección y/o colonización del patógeno en

la planta y es una característica que puede heredarse al estar controlada genéticamente (Martos Moreno, Raya Ortega et al. 2005). Se han descrito aislados más virulentos en ciertos huéspedes, e incluso razas patogénicas que se diferencian por su virulencia sobre cultivares del huésped (ej. Lechuga y tomate) (Trapero, Escudero et al. 2011).

Entre los aislados que afectan al olivo se han descrito dos patotipos que se diferencian por su virulencia en algodón y olivo. Los denominados **defoliantes (D)** resultan ser muy virulentos, produciendo síndromes más graves y en menor tiempo (Trapero, Escudero et al. 2011). Se incluyen síntomas como el marchitamiento, clorosis, defoliación, reducción drástica de peso y altura, e incluso, finalmente, la muerte del árbol (Blanco-López, Jiménez-Díaz et al. 1984, Rodrıguez-Jurado 1993, López-Escudero and Blanco-López 2001, Birem, Alcántara et al. 2009). El patotipo **no defoliante (ND)**, sin embargo, causa en el olivo los mismos síntomas pero los ataques son más leves o moderados (López-Escudero and Blanco-López 2001, López-Escudero, Del Rio et al. 2004, Martos-Moreno, López-Escudero et al. 2006, López-Escudero, Blanco-López et al. 2007).

Por lo tanto, los aislamientos del patotipo D se caracterizan por ser más virulentos que los aislamientos del patotipo ND (Schnathorst and Mathre 1966, Bejarano-Alcázar, Blanco-López et al. 1996, Bejarano-Alcázar, Blanco-López et al. 1997, López-Escudero and Blanco-López 2001, López-Escudero, Blanco-López et al. 2007). Como consecuencia, cultivos de olivo y algodón tolerantes a ND se ven muy afectados por el patotipo D, apareciendo los síntomas rápidamente y con mayor severidad siendo altamente susceptibles. Por ello, la compresión de la diversidad genética y patogénica de las poblaciones de *V. dahliae* es de relevancia para la aplicación efectiva de la enfermedad. Es decir, que la utilización óptima de genotipos resistentes y/o la aplicación eficiente de medidas adicionales de control de la enfermedad, necesitan de un conocimiento adecuado de la estructura, historia y potencial evolutivo de las poblaciones de patógenos (McDonald and Linde 2002).

Estos dos patotipos de *V. dahliae* están presentes en variedades de olivo españolas como “Cornicabra”, “Hojiblanca”, “Manzanilla” y “Picual”. La propagación del patotipo D de este patógeno en España (Bejarano-Alcázar, Blanco-López et al. 1996) y su presencia en aceite de oliva comercial (López-

Escudero and Blanco-López 2001) hacen necesario determinar que variedades de olivo tienen una mayor resistencia a *V. dahliae*.

Por ello se han llevado a cabo estudios cuyo objetivo principal fue evaluar dicha resistencia en la planta y poder así utilizarlas para la replantación, como portainjertos o como fuentes para la resistencia en los futuros programas de mejora. Uno de ellos utilizó plantas inoculadas con los aislados de ambos patotipos de diferente virulencia (López-Escudero, Del Rio et al. 2004). Los resultados del estudio indicaban que la variedad “Picual” es extremadamente susceptible al patotipo D y susceptible a los aislados ND (Hartmann, Schnathorst et al. 1971, Rodrıguez-Jurado 1993), mientras que algunas de las plantas de la variedad “Frantoio” inoculadas con el aislado del patotipo ND mostraron una recuperación de la enfermedad a partir de la séptima semana, lo cual fue asociado a un cierto nivel de resistencia. Por lo tanto la variedad “Frantoio” es considerada resistente a *V. dahliae* y podría ser utilizada para la replantación o como portainjertos para otras variedades susceptibles.

2.3. El transcriptoma y su ensamblaje.

El conocimiento del transcriptoma y su regulación es fundamental para la interpretación articulada de los diversos constituyentes moleculares que integran la red de respueta génica ante un determinado evento inductor, como los que se presentan en interacciones planta-patógeno. En plantas, la respuesta frente a estados de estrés biótico y abiótico está controlada por la actividad transcripcional de activación o represión de genes. La transcripción es el proceso nuclear cuya activación depende de estímulos intra o extracelulares que activan cascadas de señalización para determinar cuáles genes deben expresarse o reprimirse de acuerdo con el tipo de estímulo inicial. Todos los transcritos derivados de genes que se producen en una célula en un momento y bajo una condición fisiológica determinada se denomina transcriptoma, cuyo estudio y análisis es esencial para el entendimiento de la función de los genes. De manera general, se puede establecer que si un gen determinado se expresa

en una condición o célula concreta es porque cumple allí una función (Sedano and Carrascal 2012).

Actualmente, y gracias a los avances en las técnicas de secuenciación del ADN a través de tecnologías de nueva generación, NGS, se han revolucionado campos como los de la genómica y la transcriptómica. Estas tecnologías han permitido no solo generar información con altos rendimientos y a bajo costo, sino también abrir nuevos horizontes para el entendimiento detallado y global de procesos de expresión génica (Mochida and Shinozaki 2011, Schneeberger and Weigel 2011, Ward, Ponnala et al. 2012). La caracterización completa y el análisis global de la expresión génica en una célula o tejido, aún sin ninguna información genómica previa, es ahora posible a través de la implementación de la secuenciación de ADNc, o más recientemente de la secuenciación directa de ARN, tecnología conocida como **RNA-seq** (Wang, Gerstein et al. 2009, Garber, Grabherr et al. 2011, Egan, Schlueter et al. 2012, Ward, Ponnala et al. 2012). Esta herramienta transcriptómica cambia la manera de analizar y comprender los transcriptomas (Wang, Gerstein et al. 2009).

El RNA-seq se caracteriza por una captura el ARN total o ARNm, el cual se fragmenta y convierte en una librería de ADNc, la secuenciación masiva y el análisis de los resultados nos permiten tener acceso a infinidad de información relacionada con la expresión génica, posibilitando además comparar dichos niveles de expresión. Uno de los pasos fundamentales es la obtención de un ARN de buena calidad que represente todos los transcritos que se producen en la condición y el tejido de estudio (Sedano and Carrascal 2012).

El modo en que los patógenos son reconocidos por sus hospedadores y cómo establecen las interacciones de resistencia y susceptibilidad son uno de los mayores retos de la fitopatología. Por ello, la biología molecular y la bioinformática así como el estudio de la expresión génica en eventos de patogenicidad han contribuido de manera importante a la comprensión de las relaciones que se establecen entre planta y patógeno (Verhage, van Wees et al. 2010, Lodha and Basak 2012, Schenk, Carvalhais et al. 2012). La herramienta RNA-seq ha mostrado ser altamente sensible y prometedora para el análisis de transcriptomas en plantas. En el caso del algodón, el conocimiento existente de la respuesta de defensa por parte de la planta era

muy limitado. Con la aplicación de la tecnología de RNA-seq y la plataforma Illumina (compañía que se basa en el principio de amplificación en puente y el uso de marcaje por fluorescencia de nucleótidos modificados como terminadores reversibles), se obtuvo finalmente el primer análisis global de transcriptoma de defensa en algodón. En dicho estudio se pudo monitorear los perfiles de expresión en raíces a 4, 12, 24 y 48 horas post-inoculación, detectándose expresión diferencial en más de 3000 genes, lo que permitió comenzar a comprender, por ejemplo, cómo los genes involucrados en actividad enzimática, especialmente en la ruta fenilpropanoide están implicados en eventos de respuesta (Xu, Zhu et al. 2011).

En un estudio de este tipo se han de tener en cuenta determinados factores que pueden aportar una información más completa de las muestras que vamos a utilizar. Son significativos los factores genómicos, dónde hay que tener en cuenta el número de especies comprendidas en las muestras, así como la poliploidía/heterocigosidad presente en las especies a estudiar. En referencia a los factores biológicos, es importante conocer el órgano, tejido o célula en la que se está trabajando, así como también la etapa de desarrollo en la cual se encuentra y los tratamientos que se llevan a cabo. Además, hay que tener en cuenta también el hardware disponible y las habilidades técnicas y las consideraciones económicas.

En resumen, los resultados obtenidos en diferentes estudios ponen de manifiesto la existencia de variedades de olivo con diferentes niveles de resistencia y susceptibilidad al patotipo D de *V. dahliae,* además de cómo la herramienta RNA-seq ha resultado ser muy útil para el estudio de la expresión génica a nivel del transcriptoma. Por lo que en el presente Trabajo Fin de Grado se lleva a cabo el ensamblaje de un transcriptoma con dos variedades de olivo, “Picual”, muy susceptible a la infección y “Frantoio”, que presenta cierta resistencia, ambas infectadas con el patotipo D del hongo. Estableciendo de esta forma, el primer paso para un estudio de expresión génica diferencial a nivel transcriptómico mediante la herramienta RNA-seq.

3. OBJETIVO

El objetivo general de este Trabajo Fin de Grado (TFG), es ensamblar un transcriptoma a partir de datos de secuenciación masiva procedentes de la variedad de olivo tolerante “Frantoio” y la susceptible a la infección, “Picual”, por parte del hongo del suelo *V. dahliae*, estableciendo una herramienta para el estudio de la interacción entre la planta y el patógeno hacia estudios posteriores.

4. MATERIALES Y MÉTODOS

4.1. Material Vegetal.

Las muestras vegetales proceden de distintos tejidos de olivo (raíces y hojas), las cuales fueron sometidas al agente infeccioso, utilizándose plantas control sin infectar, plantas con daños en las raíces como controles de sistema de infección para así aislar la respuesta a la infección (Tabla 3.)

Tabla 3. Muestras utilizadas en el estudio sometidas a diferentes tratamientos en raíces y hojas

	Tejido	**Tiempo**	**Código "Picual"**	**Código "Frantoio"**
Control	Raíces Control	0 horas	PRC	FRC
Infección por *V. dahliae*	Raíces de plantas infectadas	48 horas	PRV48H	FRV48H
		7 días	PRV7D	FRV7D
		15 días	PRV15D	FRV15D
	Hojas de plantas infectadas	15 días	PHV15D	FHV15D
Daño mecánico	Raíces heridas	48 horas	PRH48H	FRH48H
		7 días	PRH7D	FRH7D
	Hojas de plantas con heridas en las raíces	15 días	PHH15D	FHH15D

Cada una de estas muestras de raíces y hojas de plantas sometidas a infección y daño mecánico, tomadas a diferentes tiempos, procedían en cada caso de tres plantas diferentes de la variedad "Picual" y tres plantas de la variedad "Frantoio". Se agruparon las muestras procedentes de la extracción de ARN de las tres plantas en cada una de las situaciones. Posteriormente se realizó la retrotranscripción a ADNc y se secuenciaron dos réplicas técnicas de cada muestra de ADNc en líneas diferentes (L001 y L002). Como resultado de la secuenciación masiva se obtuvieron una gran cantidad de lecturas pareadas procedentes de dichas muestras. Las lecturas pareadas o *paired-end* (Fig. 11) son fragmentos de ADN, anotados como R1 y R2, de aproximadamente 100 pb

cada una, que van emparejadas y que además poseen un sentido opuesto como muestra la Figura 11. Entre ambas lecturas *paired-end* se encuentra un fragmento de secuencia desconocida de unos 100-200 pb y que facilita, junto con la información de cada pareja, el ensamblaje. Cada una de las réplicas obtenidas de las muestras tiene un R1 y un R2. Estas lecturas han resultado ser la mejor elección para poder realizar un ensamblaje sin ningún genoma de referencia.

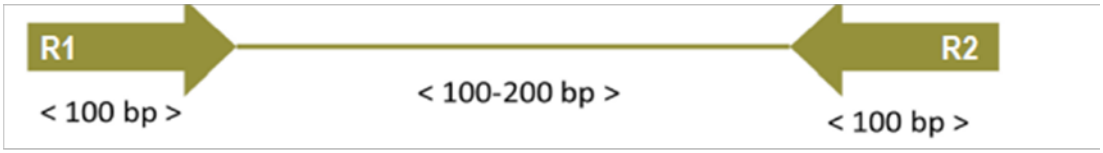

Figura 11. Esquema gráfico de una pareja de lecturas *paired-end* R1/R2.

4.2. Herramientas informáticas

Este TFG se ha llevado a cabo en un entorno operativo Linux, que nos ha permitido trabajar conectados en red al supercomputador Picasso, situado en el Centro de Supercomputación y Bioinformática de la Universidad de Málaga (SCBI). El supercomputador Picasso (scbi.uma.es) ofrece un procesamiento rápido de los datos debido a sus infraestructuras de alto rendimiento. Esta conexión se ha realizado a través de la máquina virtual Oracle VM Virtualbox (Version 4.3.6 r91406) (Fig. 12), instalada sobre un sistema Windows XP (AMD Turion ™ Dual-Core, 779 MHz, 1,74 GB de RAM). El entorno local de Linux (en nuestro caso Debian 6.0.7), se trata de un sistema multiusuario que trabaja en 32 bits, procesando la información y administrando determinadas cantidades de memoria de acceso aleatorio (RAM).

Linux permite una total y completa administración y manejo del sistema, aportando una gran estabilidad. Para llevar a cabo los análisis y la manipulación de los ficheros se necesita el conocimiento previo de una serie de comandos básicos, como serían herramientas para comprimir y descomprimir archivos, traslado de archivos, así como comandos de procesamiento de texto básico (Anexo I).

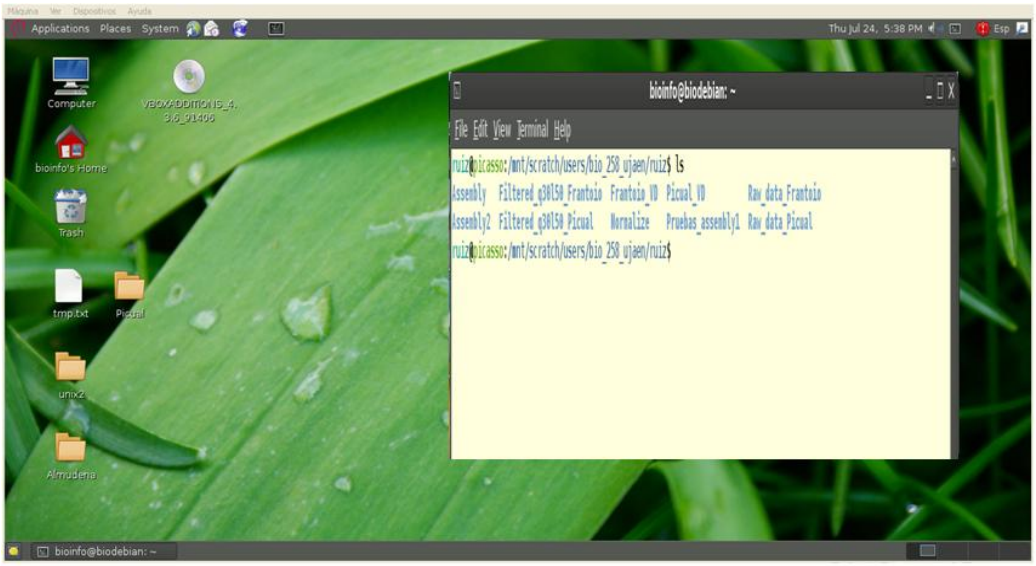

Figura 12. Máquina Virtual

4.3. Análisis de datos de partida o *raw data*

Las secuencias estaban en formato *Fastq*, un formato utilizado para el almacenamiento de secuencias de nucleótidos o péptidos que contiene además de la secuencia, información de la calidad de la misma, en dos líneas independientes. Antes de comenzar el ensamblaje, se realizó un análisis de calidad de las lecturas procedentes de la secuenciación masiva mediante la herramienta *FastQC* (bioinformatics.babraham.ac.uk/projects/fastqc). Este análisis tiene como objetivo proporcionar algunas comprobaciones de control de calidad de datos, en este caso de los datos de partida o *raw data*. Comprende un conjunto de análisis como serían estadísticas básicas de las lecturas, análisis de calidad por secuencia, contenido GC por base y por secuencia, número de indeterminaciones, análisis de la distribución de la longitud, estudio de duplicaciones, etc. (Fig. 13), que se utilizan para comprobar las secuencias y en su caso tener en cuenta estos resultados a la hora de realizar los pasos sucesivos.

El análisis se realizó en cada una de las muestras de ambas variedades. La línea de comando para ejecutarlo quedaría de la siguiente manera:

```
Fastqc -f <file_1> <file_2> <file_3> …
```

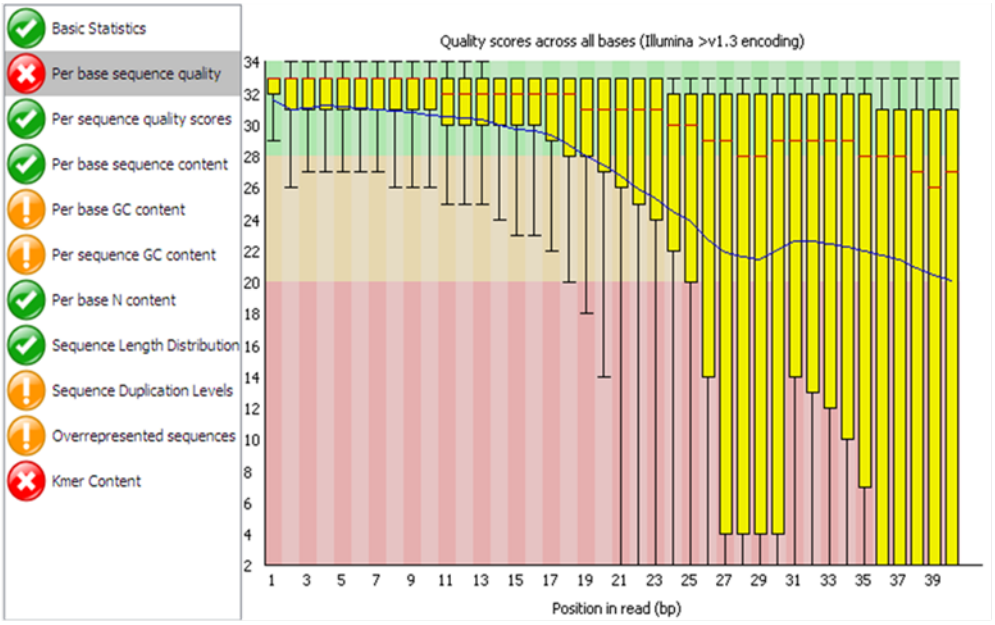

Figura 13. Análisis estadístico mediante la herramienta *FastQC*. En este caso se muestra un ejemplo de secuencias de baja calidad.

A través del comando "grep" se contaron el número de secuencias de cada una de las variedades para tener un control del número de secuencias en todo momento. La línea de comando que se utilizó es la que se muestra a continuación, repitiéndose para realizar el conteo de cada una de las muestras tanto de "Picual" como de "Frantoio".

```
grep –c '^+$' <file>
```

4.4. Filtrado.

El filtrado de las lecturas *paired-end* se realizó mediante la herramienta *fastq-mcf*, (code.google.com/p/ea-utils) utilizando para ello determinados parámetros de calidad y de longitud. De esta forma se eliminan aquellas secuencias que posean una longitud mínima de 50 pb (L50) y una calidad de 30, (Q30), además de retirar los adaptadores utilizados para la secuenciación. Este último parámetro Q o Qphred, nos indica la calidad de cada uno de los nucleótidos en la secuencia (Q= -10log10(e)), por lo que Q30 se corresponde

con un valor e=0,001 por nucleótido, eliminando aquellos con una calidad menor.

```
Fastq-mcf –q 30 –l 50 -o <output_R1> –o <output_R2>
<adaptors> <input_R1> <input_R2>
```

4.5. Análisis de datos filtrados

Tras el proceso de filtrado, se realizó de nuevo un análisis de la calidad de las secuencias mediante la misma herramienta, *FastQC,* observándose una mejora considerable de la calidad de las secuencias (Fig. 14). Del mismo modo se llevo a cabo un recuento del número de secuencias para así calcular cuantas se habían eliminado tras el proceso de filtrado. Para ello se usó de nuevo el comando “grep”.

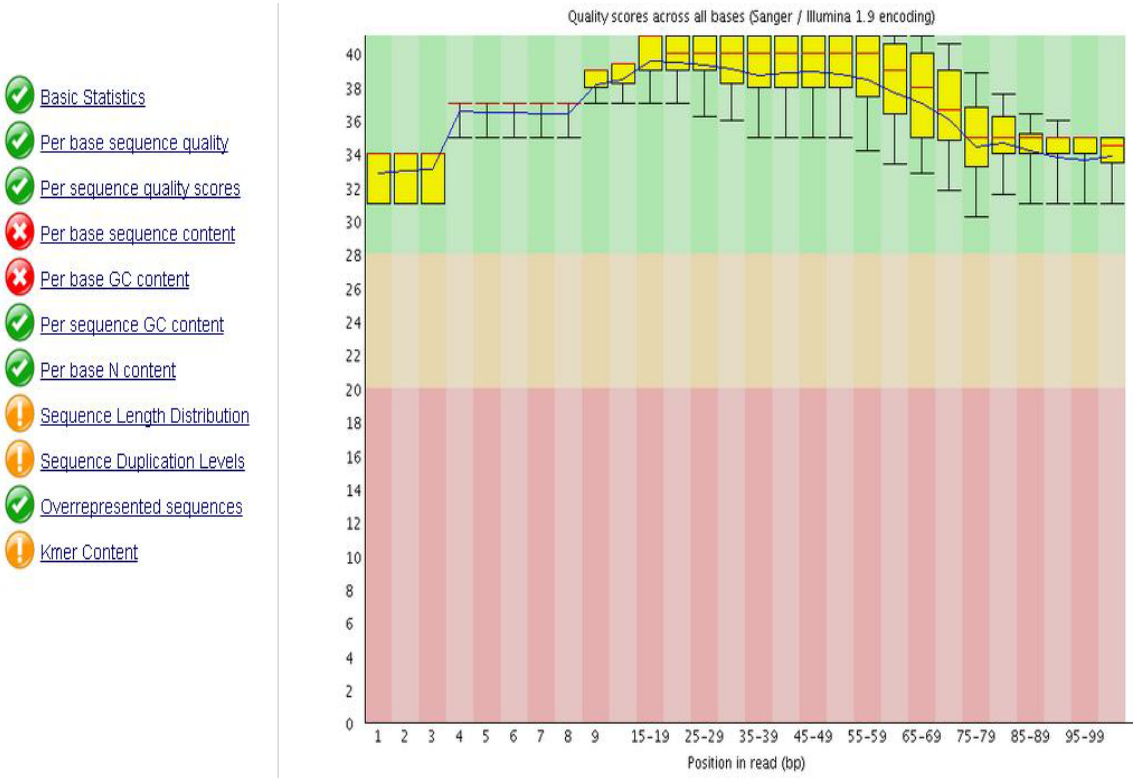

Figura 14. Análisis estadísticos mediantes la herramienta FastQC tras el filtrado de las muestras

4.6. Ensamblaje.

Figura 15. Plataforma utilizada para llevar a cabo el ensamblaje del transcriptoma.

Para realizar el ensamblaje del transcriptoma *de novo* se utilizó la plataforma **Trinity** (Grabherr, Haas et al. 2011) (Fig. 15). Trinity puede utilizar como punto de partida gran cantidad de lecturas cortas, sin usar un genoma de referencia y basadas en la resolución de gráficos complejos (gráficos de Brujin) con diferentes algoritmos. Se utilizó como indican los desarrolladores, modificando únicamente los valores de memoria requerida. Una vez filtradas las secuencias y antes de proceder al ensamblaje, los desarrolladores aconsejan que para gran cantidad de datos, más de 300 millones de lecturas o *reads*, se normalicen las muestras de forma virtual o *in silico* (trinityrnaseq.sourceforge.net/trinity_insilico_normalization.html), de manera que se minimicen los requerimientos computacionales y el tiempo de ensamblaje, sin perder información.

Trinity permite la reconstrucción eficiente y robusta de transcriptomas *de novo* a partir de grandes volúmenes de datos de RNA-seq. Antes de llevar a cabo la normalización se necesitan dos archivos que contengan todas las lecturas R1 de “Picual” y “Frantoio” y todas las R2 respectivamente, por lo que se unieron de forma concatenada ejecutando el comando “cat”. Tras comprobar que ambas tengan el mismo número de secuencias R1 y R2, usando el comando “grep”, se realizó una modificación de los identificadores de las secuencias para que todas ellas acabaran de forma determinada, ya que es necesario para el ensamblaje. De esta forma, usando el comando “sed”, todos las R1 acaban en /1 y todas las R2 acaban en /2. Para correr la normalización con Trinity:

```
Insilico_read_normalization.pl
```

Hay una serie de opciones o *flags* requeridos para su ejecución, como son: *-seqType*; para indicar el tipo de secuencia o *input* que puede ser "-fq" o "-fa", en el caso del formato fastq o fasta respectivamente, *–JM*; para marcar el número de Gigabytes a utilizar, la máxima cobertura para las lecturas también se requiere, *-max_cov <NºGB>.* Cuando se trabaja con *paired-end*, se tienen lecturas izquierda y derecha, es decir, R1 y R2, se utiliza *-left* para referirse a los R1 y *-right* para los R2. Otro *flag* sería -pairs_together; el cual obliga a la herramienta a trabajar con las lecturas pareadas manteniendo su información, es decir, si se elimina una de las dos, retira la pareja entera. *-Parallel_stats*; genera lecturas en paralelo para las *paired-end*, disminuyendo los requerimientos del ordenador. Con *-Jelly_CPU* se puede definir el número de CPUs con los que es posible trabajar. Por último, *-min_kmer_cov* se refiere al número mínimo de *kmer* (longitud de las subsecuencias en las que se va a dividir cada lectura antes de proceder a su ensamblaje) para construir el catálogo, La línea de comando, por consiguiente, quedaría de la siguiente manera:

```
Insilico_read_normalization.pl –seqType fq – JM100G – max_cov 100 –
left All_R1.fq – right All_R2.fq – pairs_together – PARALLEL_STATS –
output Normalize
```

Una vez finalizada la normalización se ejecuta el ensamblador Trinity con las opciones o *flags* que se indican a continuación, utilizando en este caso los archivos de secuencias normalizadas obtenidas en la normalización "All_R1.fq_normalized_K25_C100_pctsD200.fq" y "All_R2.fq_normalized_K25_C100_pctsD200.fq"

4.7. Manipulación del transcriptoma

Se utilizó el paquete GenoToolBox (github.com/aubombarely/GenoToolBox) el cual es una amplia colección de scripts que se utilizan para manipular datos genómicos de secuencias Fasta y Fastq, como por ejemplo

para crear modelos génicos basados en un genoma de referencia o para realizar búsquedas rápidas y extracciones de secuencias de un archivo dado.

Dentro de este paquete se utilizaron principalmente dos herramientas de análisis de secuencias: *FastaSeqStats* y *FastaExtract. FastaSeqStats* es una herramienta utilizada para analizar la longitud de las secuencias que ofrece además algunos datos estadísticos, como el número total de secuencias, la longitud total de las mismas, la secuencia más corta o más larga y varios estadísticos; N95, N50, N25, etc.

```
FastaSeqStats.pl –i <input_fasta_file> [-o <output>]
```

Siendo el *flag -i* o *input* el archivo Fasta y *–o* u *outoput* la carpeta dónde se desean guardar los análisis. Los resultados quedarían como se muestra en la (Fig. 16).

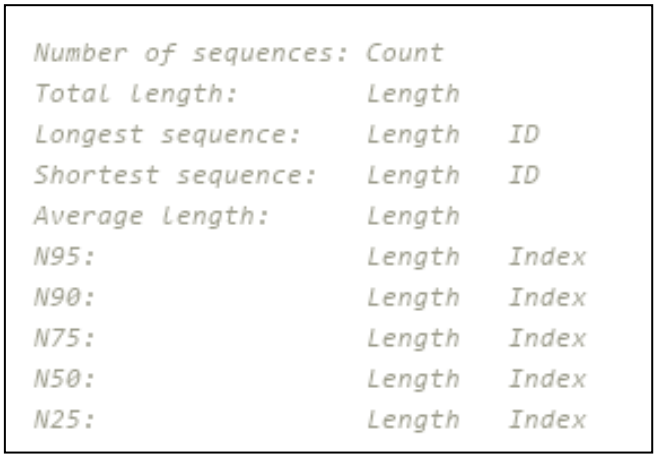

Figura 16. Resultado del análisis de las secuencias mediante la herramienta FastaSeqStats.

Por su parte, la herramienta *FastaExtract* permite extraer un subconjunto de secuencias de un archivo Fasta que comprende un conjunto de secuencias mayor. La extracción se puede realizar respondiendo a varios criterios; en función de su longitud, por su contenido en nucleótidos (%GC, poliA, etc), a partir de una lista de identificadores definida, etc. Un ejemplo para extraer las secuencias en función de la longitud quedaría:

```
fasta_extract.pl –f <input_fasta_file> -l <extract_by_length> [-o
output_basename]
```

5. RESULTADOS

5.1. Datos de partida o r*aw data*

El resultado de la secuenciación fueron aproximadamente 445 millones de lecturas *paired-end* para la secuenciación de “Picual” y 390 millones para la de “Frantoio” (Tabla 4) que abarcan ambas réplicas de la secuenciación de cada una de las muestras. Poseen un tamaño medio de 28 millones de secuencias por réplica para la primera variedad y 24,5 millones para la segunda.

5.2. Filtrado

Estas secuencias fueron sometidas a un pre-procesamiento o filtrado mediante la herramienta *fastq-mcf* utilizando determinados parámetros de calidad y de longitud, Q30 y L50, respectivamente. Con esto se pretenden retirar aquellas lecturas que resultan ser de baja calidad o que se encuentran desapareadas y que pueden intervenir negativamente en procesos posteriores como el ensamblaje. Después de realizarse el filtrado, el número de lecturas fue de 443 millones para “Picual”, por lo que fueron retiradas aproximadamente 4 millones de lecturas. Con respecto a “Frantoio”, el número de secuencias que se obtuvo tras el filtrado fue de 388 millones eliminándose por tanto alrededor de 3 millones de lecturas de baja calidad (Ver tabla 4).

5.3. Ensamblaje

El elevado número de lecturas superior a los 300 millones de secuencias y con objeto de realizar el ensamblaje correctamente se llevó a cabo previamente una normalización *in silico* recomendada por los desarrolladores del programa Trinity. De esta forma se consigue reducir la necesidad de memoria y los requerimientos computacionales para el posterior ensamblaje. El proceso normalizado no produce una pérdida de información, pues el objetivo es eliminar lecturas repetidas que no son necesarias para el proceso de ensamblado.

Tabla 4. Resultados del análisis de los datos de partida y del filtrado de las muestras.

	PICUAL			FRANTOIO		
	Datos de partida	**Filtrado**	**Eliminados**	**Datos de partida**	**Filtrado**	**Eliminados**
Raíz Control	18974562	18803949	170613	31156629	30878901	277728
	19096074	18924764	171310	31507480	31211785	295695
Raíz Herida 48h	22547755	22342392	205363	28364301	28134195	230106
	22620697	22413949	206748	28675204	28430657	244547
Raíz Herida 7 días	30930898	30730174	200724	21557747	21372869	184878
	31030239	30827818	202421	21795740	21599276	196464
Raíz VD 48h	27214837	26952334	262503	22255518	22070657	184861
	27143904	26886101	257803	22510658	22314365	196293
Raíz VD 7 días	27681432	27437498	243934	26464120	26231820	232300
	27390675	27136149	254526	26748365	26500816	247549
Raíz VD 15 días	22286016	22019596	266420	17900103	17795887	104216
	22030842	21755642	275200	17114393	17004163	110230
Hoja VD 15 días	37391858	37162192	229666	21075313	20982380	92933
	36899984	36663230	236754	20275244	20177659	97585
Hoja Herida 15 días	47498270	47203490	294780	26715547	26553047	162500
	26325267	26174566	150701	27552400	27368177	184223
Total	**447063310**	**443433844**	**3629466**	**391668762**	**388626654**	**3042108**

El ensamblador Trinity nos permite diferenciar las secuencias ensambladas a diferentes niveles jerárquicos como son los genes o unigenes y por debajo de éste el nivel de transcritos o isoformas, debidos a fenómenos de "*splicing*" alternativo. De esta forma si analizamos la distribución de las secuencias según la longitud de los fragmentos obtenemos una representación como la que se muestra en la figura 16, en la que se observa que la mayor parte de las secuencias son de pequeño tamaño, aunque existe una pequeña cantidad de secuencias ensambladas por encima de 2.500 pb.

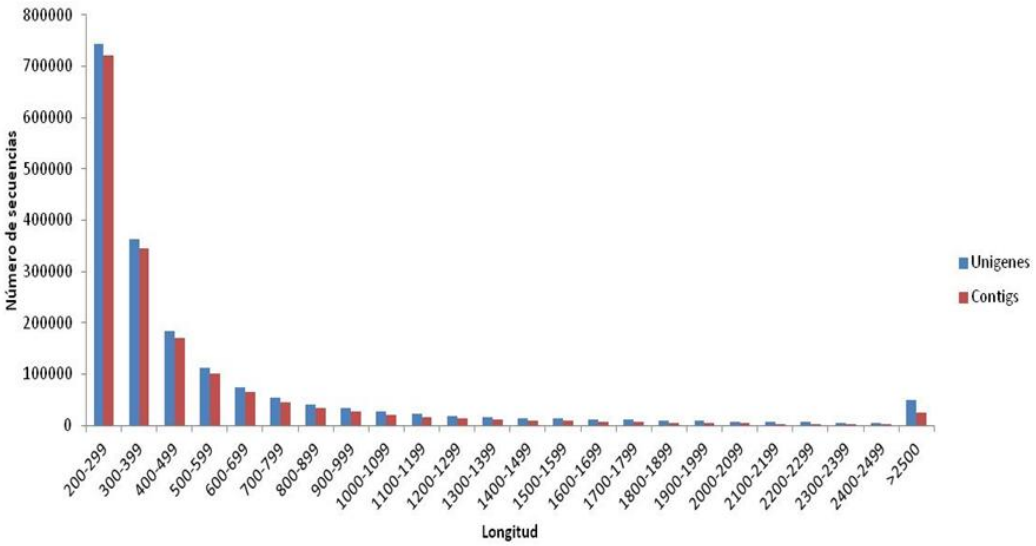

Figura 16. Distribución de las secuencias según la longitud representado en azul los unigenes y en rojo los contigs.

A la hora de evaluar distintos ensamblajes, una opción sería compararlos cuantitativa y cualitativamente con datos de otros transcriptomas que han sido publicados. Para ello se realizó una búsqueda de genomas publicados de plantas filogenéticamente cercanas al olivo. Se eligieron los transcriptomas de la flor-mono (*Mimulus guttatus*) que compartía el orden las Lamiales con el olivo y por lo tanto es la más cercana desde un punto de vista filogenético y el del tomate (*Solanum lycopersicum L*.) representante de la subclase Asterales (Fig. 17).

Se llevó a cabo la normalización de los mismos para poder compararlos entre sí, que consistió en filtrar los ensamblajes (*Solanum*, Mimulus y Trinity) para secuencias menores de 200 pb, eliminando aquellas de menor tamaño. Además, el ensamblaje generado se filtró para secuencias menores de 300 pb (Trinity_300), ya que valores menores no se consideran representativos, pues formarían parte de secuencias con ensamblajes insuficientes como para cubrir la longitud de las secuencias *paired-end* de partida. Los datos de los análisis cuantitativos de los transcriptomas realizados para los cuatro transcriptomas (Tabla 5) muestran como Trinity ensambla un número muy elevado de *contigs* con respecto a los dos transcriptomas de plantas de referencia, *S. lycopersicum* y *M.guttatus*, mientras que en el caso de Trinity_300 el número de secuencias totales disminuye considerablemente al eliminar las secuencias ensambladas de tamaño menor a 300 pb.

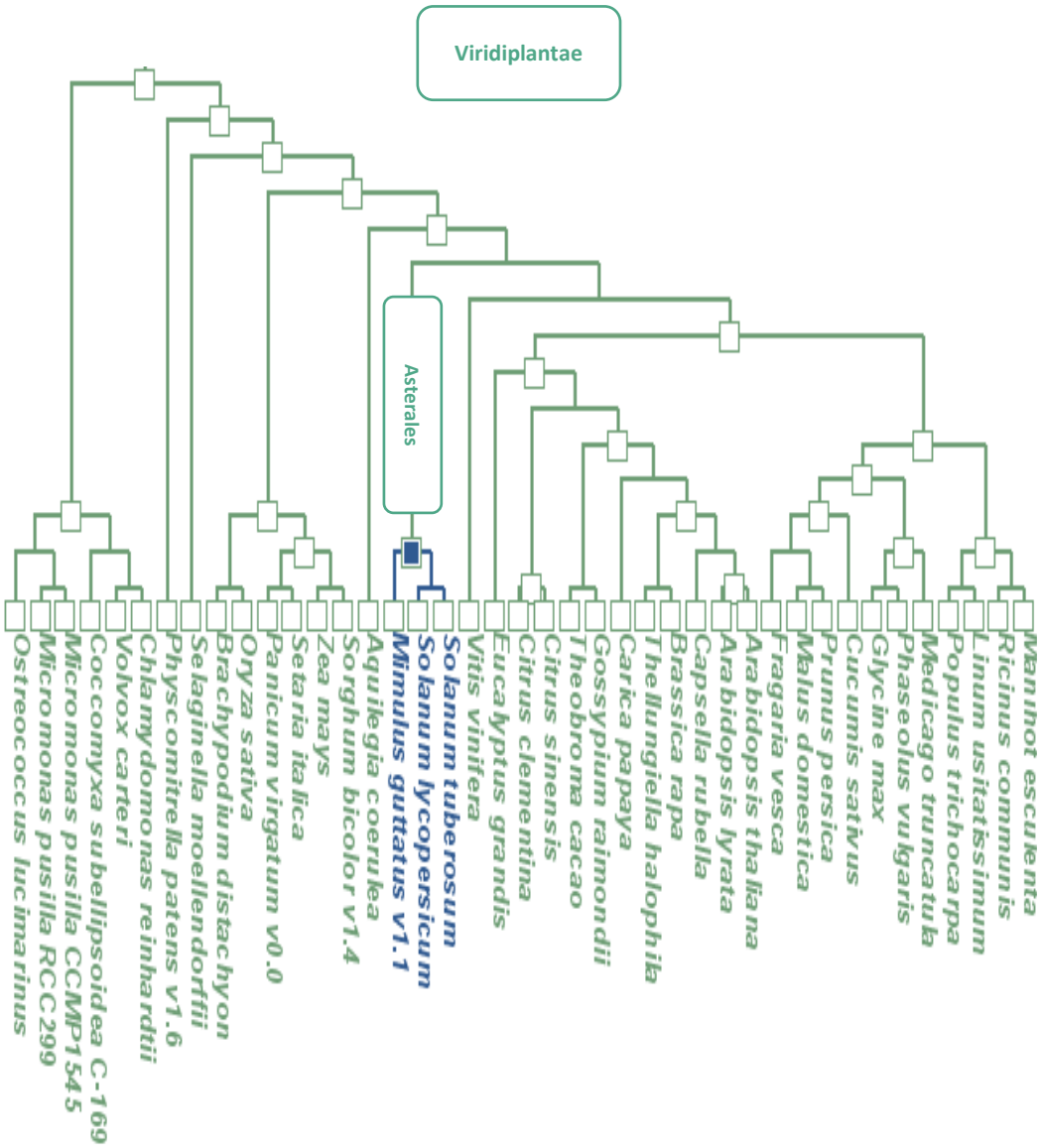

Figura 17. Árbol filogenético de los genomas de plantas publicados y disponibles en las bases de datos (modificado de phytozome.net), en azul se marcan las plantas de la subclase Asterales.

Tabla 5. Datos significativos del análisis de los transcriptomas; Se indica el número de contigs o secuencias finales, secuencia de mayor longitud, valor de N50, porcentaje y número de secuencias mayores de 1 kb de Trinity de olivo, *S. lycopersicum* y *M. guttatus.*

	Nº contings	> longitud	N50	% > 1kb	Nº sec. > 1kb
Trinity	1.837.796	19.287	785	12.79%	235.094
Trinity_300	1.094.769	19.287	1066	21,47%	235.094
S. lycopersicum	32.518	23.220	1693	53,30%	17.345
M. guttatus	28.262	15.339	1658	62,30%	17.621

Otro de los parámetros que se utilizó para evaluar la calidad del ensamblaje del transcriptoma de olivo fue la comparación de la distribución de la longitud de los transcritos ensamblados, con aquella procedente del transcriptoma de las especies seleccionadas (Fig. 18).

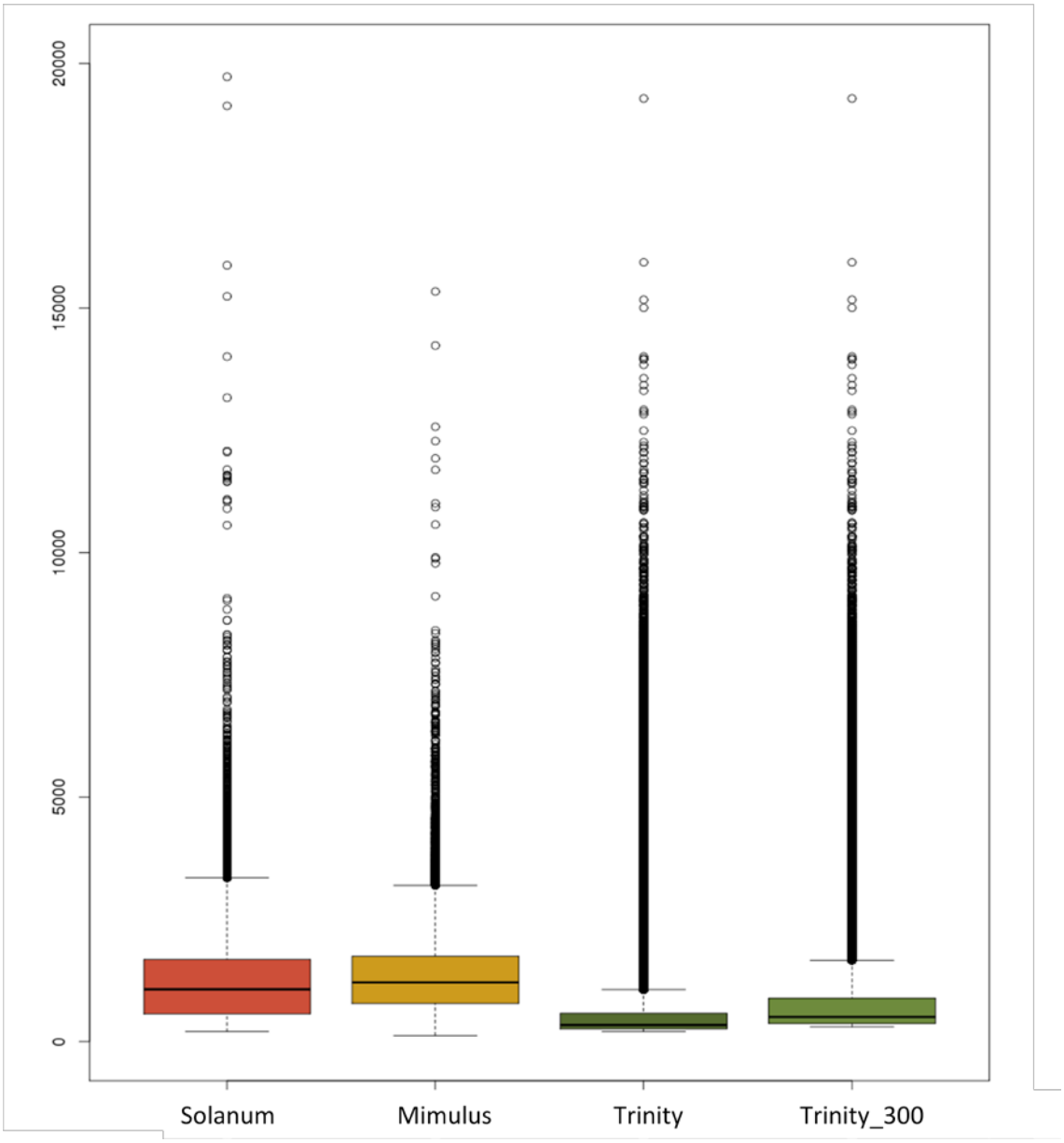

Figura 18. Diagramas de cajas que representan la distribución de longitudes de los distintos transcriptomas ensamblados (verde oliva) y del transcriptoma de *S. lycopersicum* (rojo) y de *M. guttatus* (dorado).

6. DISCUSIÓN

Las variedades de olivo cultivado “Picual” y “Frantoio” representan extremos de sensibilidad a V. dahliae por lo que comparación entre ambas variedades pueden resultar de gran interés de cara a conocer la base de la susceptibilidad/tolerancia a la Verticilosis en el olivo cultivado. Por esto se han analizado muestras pareadas de RNA-seq de plantas sometidas a la infección por *V. dahliae* y sus controles y se ha ensamblado un transcriptoma que puede ser usado de referencia para estudiar la respuesta del olivo a esta enfermedad.

Los resultados obtenidos de la secuenciación masiva generaron aproximadamente 447 millones de lecturas *paired-end* en la variedad “Picual” y 391 millones de lecturas en la variedad “Frantoio”. Con el objetivo de obtener un ensamblaje de calidad se aumentó la calidad de dichas secuencias mediante un filtrado en condiciones estrictas. Todas las muestras se filtraron para una calidad mínima por base de Q30 y para una longitud mínima, L50, eliminando de esta forma alrededor de 7 millones de secuencias en ambas variedades de olivo. El ensamblaje se realizó utilizando la plataforma Trinity con las 830 millones de lecturas restantes, mediante un paso previo de normalización, debido al elevado número de secuencias, a la capacidad computacional y al tiempo disponible en el superordenador *Picasso*, todos ellos parámetros limitantes a la hora de ensamblar los datos.

El ensamblaje bruto o primario (Trinity) mostró un elevado número de secuencias ensambladas o transcritos, casi dos millones de secuencias, además de una distribución de la longitud de las secuencias sesgada hacía las de pequeño tamaño (ver Fig. 16). Este hecho puede estar debido a múltiples factores como serían el elevado número de muestras, la utilización de diferentes tejidos y en mayor medida a la presencia de muestras procedentes de la raíz (Ver tabla 3), pues durante la manipulación de la misma no es posible aislarla completamente de los diferentes componentes de la rizosfera, tanto a nivel intra como extracelular, además de contener un elevado número de muestras infectadas con *V. dahliae*.

Debido a la naturaleza de las lecturas *paired-end* se eliminó del ensamblaje aquellas secuencias de menos de 300 pb (Trinity_300), ya que se

consideró que muestras de menor tamaño eran de longitud insuficiente para que hubiesen llegado a conectarse ambos miembros de cada pareja de lecturas. Reduciendo de forma considerable el número final de contigs que pasó a ser de aproximadamente un millón de secuencias.

Para conocer la calidad del ensamblaje se analizó cuantitativamente utilizando parámetros comparables entre el ensamblaje generado (nº de *contigs*, secuencia de mayor longitud, N50, porcentaje y número de *contigs* mayores de 1 kb) y ensamblajes de otros transcriptomas publicados de plantas cuyo genoma está disponible (phytozome.net). Se utilizaron los transcriptomas de aquellas especies más cercanas al olivo desde un punto de vista filogenético, *S. lycopersicum* y *M. guttatus*. El análisis comparativo de los datos cuantitativos de ambos transcriptomas con los ensamblajes generados, mostró que Trinity_300 posee mejores resultados que el ensamblaje bruto, aun siendo el número de *contigs* superior al de los transcriptomas de referencia. Además, la plataforma utilizada permite diferenciar unigenes de isoformas, por lo que el número total de secuencias a nivel de genes que queda en Trinity_300 sería de 934744.

En pasos sucesivos del análisis del RNA-seq el número total de secuencias sigue un complejo método de filtrado, como sería la limpieza de contaminantes, retirada de las secuencias procedentes del patógeno y posibles secuencias ensambladas “quimera”.

De este modo se ha establecido la base para un estudio RNA-seq, se ha realizado el ensamblaje de un transcriptoma *de novo* que permitirá el mapeo de las lecturas y los análisis de expresión posteriores, además de un amplio abanico de posibilidades como es la identificación de SNPs, la detección de variantes de *“splicing”* alternativo, así como una herramienta fundamental para ayudar en la anotación de genes.

7. CONCLUSIONES

1. Se ha llevado a cabo el ensamblaje de un transcriptoma completo a partir de dos variedades de olivo que responden de diferente manera al patógeno *V. dahliae.*
2. El transcriptoma de olivo ensamblado tiene un tamaño N50 de 1066 pb y 235.094 secuencias con tamaño superior a 1 kb, lo que indica una buena calidad del mismo.
3. La comparación con transcriptomas de S. lycopersicum y M. guttatus, mostró también una buena calidad del ensamblaje del transcriptoma del olivo que se ha realizo en este trabajo.
4. Este transcriptoma es una herramienta que puede ser de gran utilidad parael estudio de la base genética de la susceptibilidad/tolerancia a la Verticilosis del olivo.

8. BIBLIOGRAFÍA

Baldoni, L. and A. Belaj (2010). Olive. Oil Crops. J. Vollmann and I. Rajcan, Springer New York. **4:** 397-421.

Barranco, D. (2008). El cultivo del olivo, Mundi-Prensa Libros.

Bejarano-Alcázar, J., et al. (1996). "Etiology, importance, and distribution of Verticillium wilt of cotton in southern Spain." Plant Disease **80**(11): 1233-1238.

Bejarano-Alcázar, J., et al. (1997). "The influence of Verticillium wilt epidemics on cotton yield in southern Spain." Plant Pathology **46**(2): 168-178.

Birem, F., et al. (2009). Physiologycal differences expressed by susceptible and resistant olive cultivars inoculated with Verticillium dahliae. 10th International Verticillium Symposium, Book of Abstracts, Corfu Island, Hellas.

Blanco-López, M. and R. Jiménez-Díaz (1995). "Una propuesta de lucha integrada contra la Verticilosis del olivo." Fruticultura Profesional **70**: 52-58.

Blanco-López, M., et al. (1984). "Symptomatology, incidence and distribution of Verticillium wilt of olive trees in Andalucta." Phytopathologia Mediterranea **23**(1): 1-8.

C.O.I. "OLIVÆ." Consejo Oleícola Internacional." Revista Oficial del Consejo Oleícola Internacional **117**.

Dominguez-Garcia, M. d. C., et al. (2012). "Development of DArT markers in olive (< i> Olea europaea</i> L.) and usefulness in variability studies and genome mapping." Scientia horticulturae **136**: 50-60.

Egan, A. N., et al. (2012). "Applications of next-generation sequencing in plant biology." American journal of botany **99**(2): 175-185.

Garber, M., et al. (2011). "Computational methods for transcriptome annotation and quantification using RNA-seq." Nature methods **8**(6): 469-477.

Grabherr, M. G., et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nature biotechnology **29**(7): 644-652.

Grigg, D. (2001). "Olive oil, the Mediterranean and the world." GeoJournal **53**(2): 163-172.

Hartmann, H., et al. (1971). "Oblonga… a clonal olive rootstock resistant to verticillium wilt." California Agriculture **25**(6): 12-15.

Lodha, T. and J. Basak (2012). "Plant–Pathogen Interactions: What Microarray Tells About It?" Molecular biotechnology **50**(1): 87-97.

López-Escudero, F. and M. Blanco-López (2001). "Effect of a single or double soil solarization to control Verticillium wilt in established olive orchards in Spain." Plant Disease **85**(5): 489-496.

López-Escudero, F., et al. (2004). "Evaluation of olive cultivars for resistance to Verticillium dahliae." European Journal of Plant Pathology **110**(1): 79-85.

López-Escudero, F. J., et al. (2007). "Response of olive cultivars to stem puncture inoculation with a defoliating pathotype of Verticillium dahliae." HortScience **42**(2): 294-298.

Martos-Moreno, C., et al. (2006). "Resistance of olive cultivars to the defoliating pathotype of Verticillium dahliae." HortScience **41**(5): 1313-1316.

Martos Moreno, C., et al. (2005). "Resistencia del olivo a la Verticilosis causada por" Verticillium dahliae"."

McDonald, B. A. and C. Linde (2002). "Pathogen population genetics, evolutionary potential, and durable resistance." Annual Review of Phytopathology **40**(1): 349-379.

Mochida, K. and K. Shinozaki (2011). "Advances in omics and bioinformatics tools for systems analyses of plant functions." Plant and Cell Physiology **52**(12): 2017-2038.

Rodrıguez-Jurado, D. (1993). Interacciones huésped-parásito en la Verticilosis del olivo (Olea europaea L.) inducida por Verticillium dahliae Kleb, Ph. D. Thesis, University of C órdoba, Spain.

Rodríguez Jurado, D., et al. (1994). "La verticilosis del olivo."

Rodriguez Jurado, D., et al. (1993). "Present status of Verticillium wilt of olive in Andalucía (southern Spain)." Bulletin OEPP (EPPO).

Rugini, E. and E. Fedeli (1990). Olive (Olea europaea L.) as an oilseed crop. Legumes and Oilseed Crops I, Springer**:** 593-641.

Rugini, E., et al. (2005). Olive (Olea europaea L.). Protocol for Somatic Embryogenesis in Woody Plants, Springer**:** 345-360.

Schenk, P. M., et al. (2012). "Unraveling plant–microbe interactions: can multi-species transcriptomics help?" Trends in biotechnology **30**(3): 177-184.

Schnathorst, W. and D. Mathre (1966). "Host range and differentiation of a severe form of Verticillium albo-atrum in cotton." Phytopathology **56**(10): 1155-1161.

Schneeberger, K. and D. Weigel (2011). "Fast-forward genetics enabled by new sequencing technologies." Trends in plant science **16**(5): 282-288.

Sedano, J. C. S. and C. E. L. Carrascal (2012). "RNA-seq: herramienta transcriptómica útil para el estudio de interacciones planta-patógeno." Árbitros externos **16**(2): 101-113.

Trapero, C., et al. (2011). "La verticilosis, un grave problema de la olivicultura actual." Agricultura: Revista agropecuaria(937): 106-110.

Verhage, A., et al. (2010). "Plant immunity: it's the hormones talking, but what do they say?" Plant Physiology **154**(2): 536-540.

Wang, Z., et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews Genetics **10**(1): 57-63.

Ward, J. A., et al. (2012). "Strategies for transcriptome analysis in nonmodel plants." American journal of botany **99**(2): 267-276.

Xu, L., et al. (2011). "Lignin metabolism has a central role in the resistance of cotton to the wilt fungus Verticillium dahliae as revealed by RNA-Seq-dependent transcriptional analysis and histochemistry." Journal of experimental botany **62**(15): 5607-5621.

9. ANEXOS

9.1 Anexo I

Linux Command line cheat sheet

File system Commands	
ls	lists directories and files
ls -a	lists all files including hidden files
ls -lh	formatted list including more data
ls -t	lists sorted by date
pwd	return current working dir
cd *dir*	changes directory
cd ..	goes to parent directory
cd /	goes to root directory
cd	goes to home directory
touch *file_name*	creates en empty file
cp *file file_copy*	copy a file
cp -r	copy files contained in directories
rm *file*	deletes a file
mv *file1 file2*	moves or renames a file
rename *regexp files*	renames multiple files
mkdir *dir_name*	creates a directory
rmdir *dir_name*	deletes a directory
locate *file_name*	searches a file
man *command*	shows commands manual
top	shows process activity
df -h	shows disk space info
free -m	shows available RAM in megabytes
kill *process_id*	stops a process

Compression commands	
gzip	compress a file (faster)
gunzip	decompress a file
bzip2	compress a file (better compression)
bunzip2	decompress a file
pbzip2	compress a file (better compression with several cores)
pbzip2 -d	decompress a file
tar -cvf	groups files
tar -xvf	ungroups files

Text handling commands	
> *file*	saves STDOUT in a file
>> *file*	appends STDOUT in a file
cat *file*	concatenate and print files
cat *file1 file2* > *file3*	merges files 1 and 2 into *file3*
cat *fasta > all.fasta	concatenates all fasta files in the current directory
head *file*	prints first lines from a file
head -n 5 *file*	prints first five lines from a file
tail *file*	prints last lines from a file
tail -n 5 *file*	prints last five lines from a file
less *file*	view a file
less -N *file*	includes line numbers
less -S *file*	chops long lines
grep *'pattern' file*	Prints lines matching a pattern
grep -c *'pattern' file*	counts lines matching a pattern
cut -f 1,3 *file*	retrieves data from selected column in a tab-delimited file
sort *file*	sorts lines from a *file*
sort -u *file*	sorts and return unique lines
uniq *file*	filters adjacent repeated lines
wc *file*	counts lines, words and bytes
paste *file1 file2*	concatenates the lines of input files
paste -d ","	concatenates the lines of input files by commas
join *file1 file2*	joins input files by a common id
sed	transforms text
awk	pattern scanning and processing language

Networking Commands	
wget *URL*	download a file from an URL
ssh *user@server*	connects to a server
scp	copy files between computers
ping	check connection to a host
apt-get install	installs applications in linux