UNIVERSIDAD DE MÁLAGA
DEPARTAMENTO DE BIOLOGÍA CELULAR, GENÉTICA Y FISIOLOGÍA
ÁREA DE GENÉTICA

TESIS DOCTORAL

BIOINFORMATICS TOOLS FOR THE ANALYSIS OF PLANT-ASSOCIATED BACTERIAL GENOMES

PEDRO MANUEL MARTÍNEZ GARCÍA

JUNIO 2015

Memoria presentada por

Pedro Manuel Martínez García

para optar al grado de Doctor por la Universidad de Málaga

Bioinformatics tools for the analysis of plant-associated bacterial genomes

Directores:

Dr. Cayo J. Ramos Rodríguez

Catedrático. Área de Genética
Departamento de Biología Celular,
Genética y Fisiología
Universidad de Málaga
Instituto de Hortofruticultura Subtropical
y Mediterránea "La Mayora" (IHSM)

Dr. Pablo Rodríguez Palenzuela

Catedrático. Área de Bioquímica
y Biología Molecular
Departamento de Biotecnología
Universidad Politécnica de Madrid
Centro de Biotecnología
y Genómica de Plantas (CBGP)

Universidad de Málaga

Málaga, 2015

COMITÉ EVALUADOR

Presidente

Dr. José Manuel Palacios Alberti

Departamento de Biotecnología

Centro de Biotecnología y Genómica de Plantas (CBGP)

Universidad Politécnica de Madrid

Secretario

Dr. Javier Ruíz Albert

Departamento de Biología Celular, Genética y Fisiología

Instituto de Hortofruticultura Subtropical y Mediterránea (IHSM)

Universidad de Málaga

Vocales

Dr. Miguel Redondo Nieto

Departamento de Biología

Universidad Autónoma de Madrid

Dr. Antonio Jesús Pérez Pulido

Departamento de Biología Molecular e Ingeniería Bioquímica

Universidad Pablo de Olavide

Dr. José María Vinardell González

Departamento de Microbiología

Universidad de Sevilla

Suplentes

Dra. Carmen Beuzón López

Departamento de Biología Celular, Genética y Fisiología

Instituto de Hortofruticultura Subtropical y Mediterránea (IHSM)

Universidad de Málaga

Dr. Francisco Javier López Baena

Departamento de Microbiología

Universidad de Sevilla

Área de Genética.

Departamento de Biología Celular, Genética y Fisiología.

Instituto de Hortofruticultura Subtropical y Mediterránea (IHSM)

Universidad de Málaga-Consejo Superior de Investigaciones Científicas

Dr. **CAYO J. RAMOS RODRÍGUEZ**, Catedrático del Área de Genética del Departamento de Biología Celular, Genética y Fisiología de la Universidad de Málaga, y
Dr. **PABLO RODRÍGUEZ PALENZUELA**, Catedrático del Departamento de Biotecnología de la Universidad Politécnica de Madrid y Centro de Biotecnología y Genómica de Plantas (CBGP),

INFORMAN: Que, **PEDRO MANUEL MARTÍNEZ GARCÍA** ha realizado en este Departamento y bajo su dirección el trabajo titulado "*Bioinformatics tools for the analysis of plant-associated bacterial genomes*", que constituye su memoria de Tesis Doctoral para aspirar al grado de Doctor por la Universidad de Málaga.

Y para que así conste, y tenga los efectos que correspondan, en cumplimiento de la legislación vigente, extienden el presente informe.

En Málaga, a 10 de Abril de 2015.

Fdo. Cayo J. Ramos Rodríguez

Fdo. Pablo Rodríguez Palenzuela

AGRADECIMIENTOS

Durante estos dos años y medio son muchas las personas que han intervenido directa o indirectamente en el presente trabajo, con lo que no puedo más que dedicarles a ellos esta Tesis Doctoral.

En primer lugar, quiero agradecer a mis dos directores, Cayo y Pablo, por confiar en mi para desarrollar este proyecto. Gracias por todo lo que me habéis enseñado y por todas y cada una de las sugerencias que han hecho que esta Tesis llegue a buen puerto. Gracias a los dos por valorar mi trabajo y considerar siempre mis opiniones. Cayo, gracias por elegir el papelito sin foto que era mi currículum cuando los azares de la administración andaluza lo hicieron llegar a tu despacho. Te la jugaste conmigo, y no sabes cuánto te lo agradezco. Pablo, gracias por tu respaldo constante en todo lo que hago y por las conversaciones y elucubraciones que hemos compartido en este tiempo.

Quiero también dar las gracias a Emilia, mi jefa *number three*. Gracias por contar conmigo en tus proyectos, por tus consejos científicos y por dedicarme tiempo siempre que lo he necesitado.

No puedo olvidar al resto de compis del 285. A nuestras alumnas Bea y Silvia, por refrescar el labo con esa juventud que tanto envidiamos. A Mariela, por venir todos los días con un sonrisón latino que quita el sentío. A Chechu, por tantas conversaciones, trascendentales o cotidianas, pero siempre interesantes. A Saray, la burgalesa menos siesa que haya conocido un andaluz. Ha sido un gustazo compartir contigo estos meses, me llevo una compi de las buenas. A mi Isabella, que hasta mi llegada fue la más guapísima del CBGP. Gracias por tu compañía estos años, no sabes cómo te echamos de menos. Y sobre todo a mi Piluca de mi corazón! Mi super secre, mi confidente y mi reina mora...gracias por todo lo que has hecho por mi y por ser taaaaaaaaan buenísima gente.

A todos mis compañeros del CBGP, en especial a Bea, Marco, Adri y los Alex, los "otros" bioinformáticos. Gracias por formar parte del rato de ocio diario que han sido nuestros almuerzos. Gracias a Bea, mi gurú del aprendizaje automático. A Alex Junior, por tu complicidad salmantina y tantas frikadas que he aprendido contigo. Y gracias también a Alex Senior, por tus consejos técnicos, procedimentales, y cómo no, por alguna que otra noche loca que nos hemos pegao por Lega y Madrid.

A los compañeros y jefazos de las áreas de Genética y Microbiología de la Universidad de Málaga, en especial a Antonio de Vicente. Gracias por tu eficiencia en la gestión de todo

lo referente al genoma de la UMAF0158 y por tus expertas indicaciones necrosis-apicalianas. Y por supuesto gracias a Conchita, mi única compi de despacho en estos años. Por efímera que fuera tu estancia, me encantó tenerte conmigo esos meses. Los cromosomas circulares van por ti!

No querría pasar sin dar las gracias a Jesús Mercado por confiar en nosotros para la anotación y análisis del genoma de la PICF7. Ha sido un placer colaborar contigo.

Gracias a todos los amig@s que he ido dejando en Tomares, Sevilla, Salamanca, Florencia, Madrid, Copenhage y Barcelona, sin excepción.

Gracias a Luisito, Mateo, Juan, Jorge, Enrique y Pedro por el pasado y el presente juntos.

Gracias a Arbe, Julito, Chico, Rafa, Javi, María, Mariwa, Pedro Ángel, Esther, Gloria, Juanma, Grego, Magneto, Antuan, Quinito, Carlete, Carmela, Recu, Gema y Joselito, mis tomareños del alma. Por todos los momentos juntos. Los fines de año locos en la sierra, las infinitas ferias y carnavales. Por los Al Rumbos habidos y por haber. Siempre todos. Siempre juntos. Todos sois parte de mi y como tal todos y cada uno sois parte de este trabajo. Sin vosotros no sería el piltrafa en el que me he convertido.

Gracias a Benavente, Luisma, Moro y Elvi, mis europeos perpetuos. A mis primos, en especial a Chio y Alberto, sois más que familia. A Rodri, Agus, Morta y Antonio, mi familia salmantina. A Fernan y Laura, mi familia fiorentina. A Peter, Thomas y Steven, mi familia danesa. A Ro y Arlette, mi familia catalana. A Mari, Patas y Aza, mi familia madrileña (gracias por la portada Patorras!!). A Fidel, Tito, Sulas y Madaleno, por amenizar los inamenizables años de facultad.

A mis padres y hermana. Por vuestro eterno apoyo y cariño, tan incondicionales que sólo la genética puede explicarlos.

A Julia. Por todo. Por ser el mejor público para mis chistes, la mejor comensal cuando cocino y la mejor receptora involuntaria de cuantas parrafadas y comeduras de tarro me han surgido en el desarrollo de esta Tesis. Porque tu amor incondicional no lo explica la genética. Te quiero.

Gracias a todos los políticos de esta nuestra madre patria por vuestra inconmensurable labor, tan necesaria en estos momentos difíciles. No hay dinero en el mundo que recompense el esfuerzo y dedicación que prestáis al servicio público. Vuestro constante compromiso para con la sociedad os eleva a la condición de mártires del bien común,

tributos con que las deidades del presente han tenido a bien agasajarnos para compensar con vuestro sacrificio los privilegios y desplantes propios de nosotros los contribuyentes. Quiero dar las gracias a José Luis Rodríguez Zapatero, en especial a la curvatura de sus cejas y su inspiradora pusilanimidad. Gracias a Mariano Rajoy por el afán con que ha impulsado la educación pública, la sanidad y, sobre todo, la ciencia. No puedo olvidar a Josep Antoni Duran i Lleida y sus siempre constructivas críticas a un pueblo, el andaluz, que sin duda tiene en su más alta estima y lleva grabado en el corazón. Por último, no querría pasar sin enarbolar el inestimable empeño de Doña Esperanza Aguirre y Gil de Biedma, condesa consorte de Bornos y grande de España, por hacer de Madrid una ciudad limpia de las chabacanerías propias de la plebe, cuyos prosaicos quehaceres no tienen cabida en la capital del otrora imperio.

"*Hace falta imaginar, experimentar cosas y cambiar algo. Hace falta arriesgarse. Yo ya sabía de antemano lo que iba a pasar, claro. Es que los puristas no experimentan nada de nada. Si se queda uno sólo con los puristas nos quedaríamos siempre en el mismo sitio. Están metidos en un círculo del que no se salen, y yo creo que hay que salirse un poco, ¿no? Experimentar.*"

José Monge Cruz

Este trabajo ha sido financiado por los proyectos P10-AGR-5797 de la Junta de Andalucía y AGL2011-30343-C02-01 del Plan Nacional I+D del Ministerio de Economía y Competitividad (ambos cofinanciados por FEDER), así como por el Campus de Excelencia Internacional Andalucía TECH.

ÍNDICE

Índice

PREFACIO

Esta Tesis Doctoral se ha dirigido al estudio bioinformático de los genomas de dos cepas bacterianas asociadas a cultivos de mango y olivo, ambos de relevancia en la agricultura andaluza, así como al desarrollo de herramientas computacionales para analizar genomas bacterianos en el contexto de las interacciones planta-bacteria. Durante el desarrollo de la misma se contribuyó en varios trabajos como consecuencia de la colaboración con el Grupo de Bacterias Fitopatógenas del Centro de Biotecnología y Genómica de Plantas de la Universidad Politécnica de Madrid (CBGP-UPM-INIA), coordinado por los Dres. Emilia López Solanilla y Pablo Rodríguez Palenzuela, este último director de la presente Tesis Doctoral junto el Dr. Cayo Ramos de la Universidad de Málaga. En uno de estos trabajos se estudió el papel de la percepción de la luz en la virulencia de *Pseudomonas syringae* pv. tomato DC3000, agente causal de la mancha bacteriana del tomate (Río-Álvarez *et al.*, 2013). Asimismo, se colaboró en el estudio funcional de los quimiorreceptores implicados en el proceso de entrada en la planta a través de heridas de *Dickeya dadantii* 3937, agente causal de la podredumbre blanda de la patata (Río-Álvarez *et al.*, 2014). Los resultados obtenidos en ambos trabajos no se han incorporado en el cuerpo de esta Tesis Doctoral, si bien las publicaciones científicas derivadas de los mismos se han incluido en el Anexo. Un tercer trabajo se centró en analizar la resistencia de *P. syringae* pv. tomato DC3000 frente a compuestos antimicrobianos de tomate mediada por bombas de extrusión multidroga, cuyos resultados se incluyeron en un manuscrito que actualmente se encuentra en proceso de revisión. Por otro lado, y en consecuencia de la estrecha colaboración con el resto de integrantes de la línea de investigación "Biología y control de enfermedades de plantas" del Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora" – Universidad de Málaga (IHSM-UMA-CSIC), se analizaron los genomas de dos cepas de *Bacillus amyloliquefaciens*, potenciales agentes de biocontrol frente al oídio de cucurbitáceas (Romero *et al.*, 2007). El manuscrito que describe los resultados de este estudio ha sido recientemente aceptado para su publicación en una revista de relevancia en el área y se encuentra en proceso de producción.

En lo que respecta al cuerpo de esta Tesis Doctoral, éste se divide en cuatro capítulos que cubren sus dos objetivos principales. Los capítulos 1 y 2 abordan el análisis bioinformático de los genomas de las cepas *P. syringae* pv. syringae UMAF0158 y *Pseudomonas fluorescens* PICF7, respectivamente. *P. syringae* pv. syringae UMAF0158 es el agente causal de la necrosis apical del mango, enfermedad que ha sido objeto de intenso estudio por parte de nuestro grupo de investigación desde su descripción en 1998 (Cazorla *et al.*, 1998). Los resultados obtenidos tras el análisis de su genoma están siendo revisados en una revista de impacto para su publicación. Por su parte, *P. fluorescens* PICF7 es una cepa probada

experimentalmente como agente efectivo contra la verticilosis del olivo, y fue aislada en 2003 por el grupo de Etiología y Control de Enfermedades de Cultivos del Instituto de Agricultura Sostenible (IAS-CSIC) (Mercado-Blanco *et al.*, 2004), dirigido por el Dr. Jesús Mercado-Blanco, con quien nuestro grupo ha establecido diversas colaboraciones. El estudio bioinformático del genoma de esta cepa se publicó en la revista Standards in Genomic Sciences (SIGS) con título ***Complete genome sequence of* Pseudomonas fluorescens *strain PICF7, an indigenous root endophyte from olive (*Olea europaea *L.) and effective biocontrol agent against* Verticillium dahliae** (Martínez-García *et al.*, 2015a). Por su parte, los capítulos 3 y 4 se focalizan en herramientas bioinformáticas de anotación y análisis de genomas bacterianos, con especial énfasis en las interacciones planta-bacteria. El capítulo 3 describe una herramienta *online* para la identificación genómica de sistemas de secreción bacterianos, publicada en la revista PLOS ONE con título ***T346Hunter: A novel web-based tool for the prediction of type III, type IV and type VI secretion systems in bacterial genomes*** (Martínez-García *et al.*, 2015b). En el capítulo 4 se aborda el problema de la predicción de estilos de vida bacterianos usando información genómica. Para ello se implementó un *pipeline* que combina búsquedas de secuencias por homología con técnicas de aprendizaje automático supervisado para clasificar genomas bacterianos asociados a plantas. El correspondiente manuscrito que describe dicha herramienta se encuentra actualmente en proceso de revisión para su publicación en una revista relevante en patogénesis.

RESUMEN

El desarrollo de las plataformas de secuenciación de alto rendimiento han supuesto una extraordinaria disminución en tiempo y costes, lo que ha permitido a los expertos disponer fácilmente de las secuencias genómicas de los organismos con que trabajan. El campo de la fitobacteriología no es una excepción, y en la actualidad existe una colección prominente de genomas de bacterias asociadas a plantas en repositorios públicos. En lo que respecta al complejo *Pseudomonas syringae*, se dispone de más de 50 genomas correspondientes a sendas cepas, aunque sólo de 3 de ellos se conoce la secuencia completa: *P. syringae* pv. tomato DC3000, patógena de tomate, *P. syringae* pv. syringae B728a y *P. syringae* pv. phaseolicola 1448A, patógenas de judía. Recientemente se ha determinado el genoma completo de la cepa *P. syringae* pv. syringae UMAF0158, agente causal de la necrosis apical del mango (Cazorla *et al.*, 1998). Dado que distintas cepas de *P. syringae* provocan distintas enfermedades en un amplio rango de huéspedes, el primer objetivo de esta Tesis Doctoral ha sido el análisis bioinformático del genoma de UMAF0158, haciendo énfasis en factores bacterianos potencialmente implicados en su capacidad para infectar plantas de mango. El genoma consta de un cromosoma de 5.8 Mb y un plásmido de 63 Kb, y presenta una alta similitud (91% de las secuencias codificantes predichas) respecto al genoma de B728a. De entre los determinantes genéticos diferenciales que podrían explicar su interacción con mango, destacan el operon *mbo* de síntesis de mangotoxina (Carrión *et al.*, 2012), un clúster potencialmente implicado en la producción de celulosa, dos sistemas de secreción diferentes de tipos III y VI (T3SS y T6SS, respectivamente) y un repertorio particular de efectores del T3SS. El genoma de UMAF0158 constituye el primero completo que se obtiene de una *P. syringae* que afecta a plantas leñosas, y este estudio sienta las bases para posteriores análisis experimentales, que permitirán esclarecer los mecanismos bacterianos que explican la capacidad de este patógeno para infectar mango.

Por otro lado, recientemente se ha obtenido el genoma completo de la cepa endofita de olivo *Pseudomonas fluorescens* PICF7, probada como agente de biocontrol contra la verticilosis, una enfermedad vegetal provocada por el hongo *Verticillium dahliae* Kleb. (Mercado-Blanco *et al.*, 2004; Prieto, Mercado-Blanco, 2008). El genoma de PICF7 está formado por un cromosoma circular de 6.1 Mb, cuyo análisis bioinformático reveló genes potencialmente implicados en la asociación de dicha cepa con plantas de olivo, como los codificantes de un T3SS y dos T6SS, sideróforos, enzimas detoxificadoras y compuestos volátiles. La identificación de estos factores ayudará a direccionar posteriores estudios funcionales que permitan describir los mecanismos moleculares implicados en el estilo de vida endofítico de PICF7, así como su capacidad de biocontrol.

En paralelo, se han desarrollado dos herramientas bioinformáticas para el análisis

de genomas bacterianos en el contexto de las interacciones planta-bacteria. La primera de ellas, llamada T346Hunter, es una aplicación web que permite, dado un genoma bacteriano, identificar genes implicados en la síntesis de los componentes estructurales de los sistemas de secreción T3SS, T4SS y T6SS. La herramienta presenta los resultados mediante un documento HTML intuitivo y fácilmente interpretable, mediante el que el usuario puede "navegar" por las diferentes regiones detectadas y visualizar la localización genómica de los distintos componentes que éstas contienen. Se implementó a su vez una segunda aplicación web, llamada PIFAR (*Plant-bacteria Interaction FActors Resource*), consistente en un repositorio público de determinantes genéticos bacterianos implicados en interacciones planta-bacteria. Para ello, se realizó una búsqueda pormenorizada en la bibliografía científica con objeto de identificar aquellos productos génicos descritos experimentalmente como implicados en asociaciones bacterianas con plantas huésped. A través de la interfaz de PIFAR, el usuario puede consultar la base de datos, así como descargarla en diferentes formatos. También se incluye un formulario que permite subir nuevos factores para su eventual inclusión en la base de datos, y dispone de una herramienta de anotación para genomas bacterianos de entrada. Dicha herremienta se ejecutó sobre el conjunto de genomas bacterianos completos disponible en el NCBI en Diciembre de 2014, correspondiente a 3.042 cromosomas y 2.200 plásmidos. Los resultados obtenidos son también accesibles desde la interfaz de PIFAR.

Por último, se abarcó el problema de la clasificación de genoma bacterianos. Combinando T346Hunter y PIFAR con el método de aprendizaje automático Random forests (Breiman, 2001), se generó un modelo probabilístico basado en las anotaciones de ambas herramientas sobre una selección de 420 genomas bacterianos. Dicho modelo permite asignar probabilidades de asociación con planta a secuencias de entrada. Su aplicación sobre el conjunto de aproximadamente 9.500 genomas bacterianos almacenados en el NCBI en Diciembre de 2014 ha revelado potenciales asociaciones con plantas de una serie de patógenos bacterianos comúnmente asociados a mamíferos. De entre ellos destacan un conjunto amplio de enterobacterias cuya capacidad para inducir enfermedades en humanos ha sido probada en los últimos años.

INTRODUCCIÓN GENERAL

1. Bioinformática y análisis genómico

En la actualidad, la cantidad ingente de datos generados por las nuevas tecnologías es sencillamente inabordable sin la ayuda de herramientas computacionales, que permiten automatizar la extracción de información relevante para generar nuevo conocimiento. Cuando la naturaleza de dichos datos es de índole biológica, el conjunto de herramientas que los tratan se enmarcan en el ámbito de la bioinformática. Según Luscombe *et al.* (2001), *la bioinformática conceptualiza la biología en términos de moléculas (en un sentido físico-químico) y aplica técnicas informáticas para entender y organizar a gran escala la información asociada a dichas moléculas*. Es, pues, en el contexto de la biología molecular en el que la bioinformática toma sentido y se presenta como una disciplina indispensable. En la era de la biotecnología (Ganguly *et al.*, 2014), en que las técnicas de secuenciación masiva producen a diario inmensas cantidades de información genómica, la necesidad de implementar herramientas bioinformáticas para procesar y analizar dicha información es, cuanto menos, crucial.

En genómica, la secuenciación consiste en la aplicación de métodos bioquímicos con objeto de determinar el orden de los nucleótidos (A, C, G y T) que componen una secuencia de ADN. Desde que Sanger y Gilbert (Sanger *et al.*, 1977; Maxam, Gilbert, 1977), desarrollaron las primeras aproximaciones en los años 70, las tecnologías de secuenciación han experimentado una evolución notable. Como consecuencia, en 1995 se publicó por vez primera el genoma secuenciado de un organismo de vida libre, la bacteria *Haemophilus influenzae* (Fleischmann *et al.*, 1995). Seis años después vio la luz el primer borrador del genoma humano (Lander *et al.*, 2001), gracias a un proyecto coordinado de investigación sin precedentes. Este hito, aunque excepcional, supuso un gasto de 3.000 millones de dólares (un dólar por nucleótido leído), lo que implicaba un coste prohibitivo para los laboratorios. La necesidad de abaratar la secuenciación llevó al desarrollo de lo que hoy conocemos como secuenciación de nueva generación (*next generation sequencing*), que permite secuenciar un genoma humano por 1.800 dólares y un genoma bacteriano por 500. No es de extrañar que sólo en la base de datos del NCBI (*http://www.ncbi.nlm.nih.gov*) haya depositados a día de hoy (Marzo de 2015) cerca de 200 millones de secuencias, de las cuales alrededor de 10.000 corresponden a genomas de cepas bacterianas. Esta desmesurada producción de datos genómicos traslada el problema de costes de secuenciación al de ensamblar, procesar y manejar dichos datos de manera que se pueda extraer información útil de ellos. Es así como los métodos computacionales adquieren una importancia sustancial en la biología moderna, adaptando sus paradigmas a problemas específicos a nivel molecular.

La etapa actual de progreso en la generación de datos biológicos coincide, no por

casualidad, con el momento de más auge en lo que a tecnologías de la información se refiere. Irremediablemente, el efecto de la sociedad de la información también ha sido notable en el desarrollo de la electrónica, la ingeniería del software y las telecomunicaciones, lo que ha supuesto un estímulo en el avance de las técnicas de procesamiento y análisis inteligente de datos (Peek, Swift, 2012; Pea, Vityaev, 2010). Así pues, si la generación masiva de datos biológicos deriva en el problema evidente de procesar dicha información, ésto, por fortuna, sucede en un momento en que la informática está preparada para dar soluciones efectivas. Por su relevancia en esta Tesis Doctoral, a continuación se revisan algunas de las técnicas computacionales más relevantes para el tratamiento y análisis de datos biológicos.

1. 1. Análisis de secuencias biológicas

El análisis de secuencias biológicas consiste en la aplicación de métodos analíticos a secuencias de ADN, ARN o aminoácidos con objeto de inferir propiedades biológicas de las mismas, como función, estructura o evolución (Durbin *et al.*, 1998). En este contexto, los métodos de comparación de secuencias son de especial importancia. Por un mero proceso evolutivo, secuencias nuevas son necesariamente resultado de adaptaciones de otras ya existentes. Si una secuencia tiene un alto grado de similitud con otra conocida, es razonable inferir que su función será a su vez semejante. Cuando esto ocurre, se dice que ambas secuencias son homólogas, y se asume que tienen un origen evolutivo común. Aquí toma sentido el concepto de alineamiento, que permite comparar dos o más secuencias resaltando zonas de alta similitud, indicando posibles relaciones funcionales entre los genes o proteínas comparados.

Desde el punto de vista de la programación, una secuencia (por biológica que sea) no es más que una sucesión de caracteres, o formalmente, cadena de caracteres. Típicamente, los algoritmos computacionales de alineamiento implementan métodos de optimización que buscan el mejor alineamiento entre dos cadenas en función de un sistema de puntuación determinado. Las cadenas pueden alinearse local o globalmente. En el primer caso, se buscan subsegmentos de una secuencia A en otra secuencia B, lo que resulta especialmente útil para búsquedas en bases de datos de gran tamaño. Un ejemplo de programa basado en este tipo de alineamiento es BLAST (Altschul, 2005). Los alineamientos locales permiten identificar secuencias con cierto grado de similitud, pudiendo ser o no homólogas. Por su parte, los algoritmos de alineamiento global comparan las secuencias enteras, forzando al alineamiento a ocupar la longitud total de las mismas. Este tipo de alineamiento es computacionalmente más caro, siendo útil cuando se comparan un número no demasiado grande de secuencias similares, como ocurre con los alineamientos

múltiples. Una aplicación directa de estos alineamientos es el estudio filogenético de un conjunto de secuencias. Dado que las secuencias que forman el conjunto a alinear tienen una supuesta relación evolutiva, mediante el alineamiento de secuencias sucesivamente menos emparentadas se infiere el árbol filogenético del conjunto. Así es como operan los llamados métodos jerárquicos o de árbol, como ClustalW (Larkin *et al.*, 2007) o T-Coffee (Notredame *et al.*, 2000). Otro uso común de los alineamientos múltiples es el de, dado el alineamiento de un conjunto de secuencias pertenecientes a una familia determinada, comprobar si una secuencia nueva pertenece a dicha familia en función de lo bien que se alinee con el resto. Si queremos afirmar que esa secuencia pertenece a la familia en cuestión, sería deseable que al menos los patrones más conservados en el alineamiento múltiple estén presentes. Para capturar esta información se suele hacer uso de modelos probabilísticos, en especial los llamados modelos ocultos de Markov o HMM (*hidden Markov models*) (Zhang, Wood, 2003; Eddy, 2004).

Así, los distintos algoritmos de alineamiento a aplicar dependen en gran medida del conocimiento biológico que se quiera extraer, haciendo de la comparación de secuencias una parte esencial de la bioinformática. De hecho, una vez obtenida la secuencia del genoma de un organismo concreto, el siguiente paso no es otro que el de compararla con secuencias ya existentes, para lo que el investigador tiene a su disposición una suerte de repositorios públicos con información sobre genes, proteínas, y otros componentes genéticos conocidos.

1.2. Bases de datos biológicas

Las bases de datos biológicas son a día de hoy herramientas indispensables para los científicos a la hora de abordar cualquier tipo de fenómeno biológico, ya sea la estructura de las biomoléculas y sus interacciones o la propia evolución de los organismos. El conocimiento biológico moderno está en gran medida almacenado en infinidad de repositorios generalistas o especializados que crecen en proporción a la capacidad de generación de información, haciendo especialmente laboriosa la tarea de asegurar la coherencia de los datos. Por ello, iniciativas como la de la revista *Nucleic Acids Research* (NAR), que publica anualmente una exhaustiva clasificación con las bases de datos biológicas y bioinformáticas más relevantes (*http://www.oxfordjournals.org/our_journals/nar/database/c/*), son particularmente útiles.

El DDBJ (*DNA Data Bank of Japan*), el EMBL-EBI (*European Molecular Biology Laboratory*) y el GenBank del NCBI (*The National Center for Biotechnology Information*) forman el consorcio INSDC (*International Nucleotide Sequence Database Collaboration*), e intercambian información a diario para formar el actual mayor banco de secuencias de ADN existente (Nakamura *et al.*, 2013). GenBank, en concreto, almacena una colección

anotada de todas las secuencias de ADN disponibles públicamente (Benson *et al.*, 2014), hasta un total de 181.336.445 en Febrero de 2015.

En lo que respecta a repositorios de proteínas, UniProtKB (*Universal Protein Knowledgebase*, Boutet *et al.* 2007) surge con la misión de proveer a la comunidad científica de una base de datos pública, exhaustiva y de alta calidad de secuencias de proteínas e información funcional de las mismas. A día de hoy, UniProtKB es sin duda la base de datos de proteínas más prominente y usada, y consta a su vez de dos repositorios: Swiss-Prot (547.599 secuencias anotadas manualmente) y TrEMBL (90.860.905 secuencias anotadas de forma automática). Otras bases de datos tienen como objeto almacenar familias de proteínas. Tal es el caso de PFAM (Finn *et al.*, 2014), que consiste en una extensa colección de alineamientos múltiples y modelos ocultos de Markov, y que resulta especialmente útil en la identificación de regiones funcionales (dominios) en secuencias de proteínas.

Estas y otras bases de datos biológicas son rastreadas a diario por los investigadores, y algunas de ellas constituyen el núcleo de información en que se basan los programas de anotación automática.

1.3. Herramientas de anotación

Una vez obtenida la secuencia de un genoma pasamos a la fase de anotación del mismo. En genómica se distinguen dos tipos de anotaciones, la estructural y la funcional. La primera se encarga de identificar qué genes están presentes, dónde se ubican en la secuencia y qué proteínas codifican. Ésto puede inferirse *ab initio* directamente de las propiedades específicas de la secuencia de ADN, como la existencia de codones de inicio y terminación donde empiezan y acaban los marcos de lectura. Este tipo de información se integra en detectores de contenido, en base a los cuales se identifican regiones codificantes e intergénicas y se determina la posición de los genes. Un ejemplo de programas de predicción de genes *ab initio* es Glimmer (Delcher *et al.*, 2007), que se basa en los llamados modelos de Markov interpolados para identificar marcos de lectura en bacterias, arqueas y virus. Otro forma de predecir genes es en base a posibles homologías con otros genes conocidos. Haciendo uso de herramientas de comparación de secuencias como las descritas en el apartado 1.1, se puede analizar el genoma en busca de genes razonablemente parecidos a otros ya estudiados. En la práctica, los pipelines de anotación automática de genomas suelen usar modelos híbridos, de manera que en primera instancia predicen las posiciones de los genes y las proteínas que codifican, y a continuación éstos se comparan con otros conocidos mediante búsquedas en bases de datos biológicas. Una de las herramientas de anotación de genomas bacterianos más usadas es PGAP (*Prokaryotic Genome Annotation Pipeline*),

diseñada por el NCBI (Angiuoli *et al.*, 2008). En base a una secuencia de ADN bacteriano, PGAP combina predicción de genes con búsquedas por homología para producir una anotación general que almacena en archivos ASN.1 (*Abstract Syntax Notation One*), avalados por ISO. NCBI ToolKit ofrece una serie de programas para, a partir de estos archivos, extraer la información sustancial del genoma anotado, generando a su vez ficheros en formato FASTA y tablas fácilmente procesables. Tanto los ASN.1 como las tablas y archivos FASTA son los formatos más extendidos entre la comunidad científica y suelen ser compatibles con la mayoría de programas de análisis de secuencias, lo que supone una ventaja para su procesamiento. Es por ello que han sido los formatos usados en el desarrollo de esta Tesis Doctoral para anotar y tratar genomas bacterianos.

Otra de las herramientas más usadas de anotación de genomas bacterianos es RAST (*Rapid Annotation using Subsystem Technology*; Overbeek *et al.* 2014). RAST ofrece una serie de ventajas respecto a otros sistemas de anotación, lo que le ha hecho adquirir una popularidad notable en los últimos años. Entre estas ventajas, destaca la rapidez con que el usuario obtiene la anotación del genoma. En cuestión de minutos, RAST produce una anotación pormenorizada tomando como entrada la secuencia de un genoma bacteriano junto con información básica sobre la cepa. RAST se basa en un repositorio propio que integra anotaciones de una gran variedad de fuentes distintas, curadas a mano por expertos microbiólogos. A raíz de esa base se generan anotaciones automáticas de bastante precisión, generando a su vez anotaciones específicas de colecciones de proteínas funcionalmente relacionadas, o subsistemas. Este tipo de anotación aporta un componente de especificidad que no ofrecen otros sistemas como PGAP. Pero esta búsqueda de subsistemas es dependiente del repositorio de genomas sobre el que RAST trabaja, con lo que a efectos de cepas de reciente descripción, la anotación que este servidor ofrece no dista demasiado de la que produce PGAP. Es por ello que en la actualidad, y especialmente cuando se trabaja con bacterias no descritas previamente en la literatura científica, resulta más práctico complementar la anotación general con anotaciones específicas, que producen *outputs* con información exhaustiva de sistemas moleculares concretos.

1.4. Minería de datos y aprendizaje automático

La minería de datos utiliza métodos estadísticos y de inteligencia artificial, entre otros, para interpretar grandes volúmenes de datos y estructurarlos de manera que la información subyacente sea comprensible. Su fin último es explotar y analizar datos para ayudar a la toma de decisiones, así como generar sistemas inteligentes capaces de entenderlos (Hastie *et al.*, 2005; Gorunescu, 2011). En bioinformática, una de las aplicaciones de minería de

datos más usadas es la generación de modelos predictivos, en especial aquellos enfocados a problemas de clasificación (Larrañaga *et al.*, 2006). Para ello se utilizan técnicas de aprendizaje automático, que inducen modelos probabilísticos en base a un conjunto de datos llamados de entrenamiento. Estos datos de entrenamiento constituyen la información previa sobre la que se infiere el modelo de representación, que posteriormente puede aplicarse a nuevos datos y observar si éstos se comportan como los datos de entrenamiento. El aprendizaje de los datos puede ser supervisado o no supervisado (Love, 2002). En el primer caso, el entrenamiento se lleva a cabo con conocimiento de qué subconjuntos de datos son representativos de aquello que se quiere clasificar. En el segundo, este conocimiento es desconocido, por lo que la tarea del programa es encontrar patrones que ayuden a definir y clasificar los datos en distintos grupos, en función de su contenido. Para ilustrar ambos tipos de aprendizaje supongamos que disponemos de una base de datos de secuencias de ADN humano y queremos generar un modelo para discernir si esas secuencias corresponden a promotores. Los datos de entrenamiento contendrían secuencias ubicadas tanto en promotores como fuera de ellos. Para aplicar aprendizaje supervisado, tendríamos que etiquetar las secuencias que corresponden a promotores y las que no, de manera que el programa tenga información previa sobre los tipos de datos a clasificar. Si no dispusiéramos de esa información, al aplicar aprendizaje no supervisado esperaríamos que el modelo infiriera los tipos de secuencias, por lo que en este caso toma especial relevancia la interpretación de la información obtenida. Este tipo de aprendizaje no podría usarse para decidir si una secuencia es un promotor o no, pero sí podría, por ejemplo, clasificar las secuencias entre aquellas que tienen un alto contenido en GC, AT, y el resto. De esta información se pueden inferir conocimiento, si por ejemplo tenemos en cuenta que el 70% de los promotores humanos tienen un alto contenido en islas CpG (Saxonov *et al.*, 2006). Como se ha señalado, aquí la interpretación del investigador es fundamental.

2. Interacciones planta-bacteria

La extraordinaria capacidad de adaptación de las bacterias a diferentes condiciones ambientales les ha permitido colonizar prácticamente cada rincón de la Tierra. Como en cualquier organismo vivo, los mecanismos que dictan el estilo de vida de las bacterias se basan en su material genético y las modificaciones que éste sufre como resultado de la interacción de las mismas con el medio. Así, entender esta capacidad de adaptación y los componentes moleculares subyacentes a la misma pasa en gran medida por el estudio de sus genomas. El genoma de una bacteria está esencialmente formado por un cromosoma, generalmente circular y cerrado por enlace covalente. Muchas bacterias disponen también de una o varias moléculas de ADN extracomosómico, normalmente también cerrado y circular, llamado plásmido. Aunque estos últimos suelen codificar información para funciones no esenciales para la vida de la bacteria, sí albergan genes que les pueden proporcionan propiedades fenotípicas útiles en un contexto de adaptación al crecimiento en determinados medios. Por lo general, el tamaño de un plásmido en una bacteria es significativamente menor que el del cromosoma. El tamaño de los genomas bacterianos conocidos varían entre 139 kilobases (kb) y 14 megabases (Mb) (López-Madrigal *et al.*, 2011; Chang *et al.*, 2011), aunque la mayoría de ellos oscila entre 2 y 6 Mb (Koonin, Wolf, 2008).

Como hemos dicho, la ubiquidad de las bacterias se explica por su capacidad de desarrollar estrategias adaptativas que les permiten sobrevivir en prácticamente cualquier ambiente posible. Parte esencial de esas estrategias son las distintas interacciones que las bacterias establecen con otros seres vivos, ya sean más complejos, como plantas, animales y hongos, como con otros microorganismos. A menudo, las bacterias son capaces de penetrar en otros organismos y proliferar en su interior. Esta capacidad de invasión reside en que cuentan con los componentes genéticos necesarios para colonizar los tejidos del huésped, invadirlos y en ocasiones sintetizar factores de virulencia que provocan enfermedad. En general, las bacterias que habitan en organismos eucariotas se benefician de los nutrientes y del hábitat protegido que éstos les proporcionan. Desde el punto de vista del huésped, esta relación puede ser de mutualismo, comensalismo o parasitismo, en función de si es beneficiosa, neutral o perjudicial, respectivamente. Se cree que la mayoría de las interacciones que los seres humanos tenemos con los microorganismos que habitan en nuestro cuerpo son mutualistas o comensalistas (Fukada, 2014) y, desde un punto de vista evolutivo, dicha afirmación tiene pleno sentido. Dado que las bacterias han colonizado el mundo billones de años antes de la existencia del ser humano, en el proceso de evolución de los organismos complejos, éstas han debido desempeñar un papel fundamental en el desarrollo de sus funciones vitales. No es de extrañar que el número estimado de células

bacterianas presentes en el cuerpo humano sea diez veces superior al de las propias células humanas (Pappas, 2009). De hecho, sólo el microbioma intestinal alberga entre 500 y 1000 especies distintas de bacterias, cuyo número total de genes es cien veces mayor que el que contiene todo el genoma humano (Gill *et al.*, 2006). De entre la batería de genes presentes en el genoma de *Lactobacillus casei*, por ejemplo, aquellos implicados en la utilización de nutrientes como el azúcar, cuya concentración es escasa en el lumen intestinal, parecen jugar un papel fundamental en la colonización y el establecimiento de la bacteria en este tejido (Licandro-Seraut *et al.*, 2014). Asimismo, el hospedador obtiene una serie de ventajas inducidas por el la actividad de las bacterias que lo habitan. Un estudio *in vivo* demostró que *Acetobacter pomorum* dispone de genes que modulan la ruta de señalización de insulina en *Drosophila melanogaster*, cuya disrupción afecta significativamente al crecimiento de dicho organismo (Shin *et al.*, 2011).

Pero el mutualismo entre bacterias y organismos complejos no se reduce, ni mucho menos, al reino animal. De hecho, uno de los ejemplos más paradigmáticos de relación mutualista es la asociación que establecen las bacterias del género *Rhizobium* con las plantas leguminosas, que proporcionan a las bacterias un hábitat seguro y nutritivo mientras se benefician de la capacidad de estas para fijar nitrógeno, y así crecer en suelos con bajas concentraciones de este elemento (Gourion *et al.*, 2015). Las bacterias rizobiales se establecen endosimbióticamente dentro de nódulos que se forman en las raíces de las plantas leguminosas como consecuencia de esta interacción. Estas plantas producen flavonoides, que son reconocidos por receptores bacterianos y que inducen la expresión de genes de nodulación, implicados en la formación de los nódulos radiculares (Haag *et al.*, 2013).

Por su especial relevancia en sanidad, la forma de interacción bacteriana tradicionalmente más estudiada es la que establecen las bacterias patógenas con sus distintos huéspedes. No en vano, las enfermedades infecciosas siguen siendo una de las principales causas de mortalidad en los países en vías de desarrollo, más de la mitad de las cuales están provocadas por bacterias (Molicotti *et al.*, 2014). Tuberculosis, peste, neumonía o diarrea son causa directa de la infección de diversos patógenos bacterianos, que cuentan con un arsenal de genes implicados en la síntesis de distintos factores de virulencia, necesarios para establecerse en nuestro organismo y provocar enfermedad. Bacterias del género *Streptococcus*, como *Streptococcus pneumoniae*, agente causal de la neumonía, albergan genes relacionados con el metabolismo de carbohidratos complejos, que les sirven para la adquisición de nutrientes esenciales, adherirse a los tejidos e interferir con las funciones del sistema inmune (Shelburne *et al.*, 2008). Ciertas especies de *Streptococcus* portan además elementos genéticos móviles, como los transposones del tipo Tn*916*/Tn*1545*, que contienen genes responsables

de la resistencia a antibióticos (Santoro *et al.*, 2014). La relativa facilidad con que este tipo de determinantes génicos se transmiten a otras bacterias por transferencia horizontal contribuye en gran medida a la propagación de resistencia a antibióticos de unas bacterias a otras, ya sean o no del mismo género, con el correspondiente impacto en salud pública que ello conlleva. Otros patógenos como *Bacillus anthracis*, agente causal del ántrax, basan su capacidad infecciosa principalmente en la secreción de compuestos tóxicos. Típicamente, las cepas de *B. anthracis* portan, además del cromosoma, dos megaplásmidos (pXO1 y pXO2) que albergan genes encargados de la síntesis y secreción de toxinas, como las llamadas toxina causante de edema (*edema factor*) y toxina letal (*lethal factor*) (Keim *et al.*, 2009; Brossier, Mock, 2001).

Por otro lado, estudios recientes han mostrado que enteropatógenos comúnmente asociados a animales y humanos, como *Salmonella*, pueden tambier infectar plantas, dando lugar a intoxicaciones alimentarias provocadas por la ingesta de frutas y vegetales (Schikora *et al.*, 2012; Wiedemann *et al.*, 2015). Este género bacteriano es el principal causante de infecciones en mamíferos como la gastroenteritis o la fiebre tifoidea, y se estima que, sólo en los Estados Unidos, provoca alrededor de un millón y medio de casos de enfermedad al año (Mead *et al.*, 1999). Se ha demostrado que algunos genes de *Salmonella typhimurium*, como los encargados de la síntesis de celulosa, juegan un papel fundamental en el proceso de adhesion a los tejidos de las plantas (Lapidot *et al.*, 2006). Asimismo, la translocación de proteínas efectoras (o efectores) al interior del huésped resulta esencial en el proceso de infección. El cromosoma de *Salmonella* suele contener al menos dos operones que codifican las subunidades estructurales de los llamados sistemas de secreción de tipo III o T3SS (*type III secretion system*), cuya implicación en la asociación bacteria-huésped se ha descrito ampliamente en la bibliografía (Gophna *et al.*, 2003; Izoré *et al.*, 2011). Sin embargo, los mecanismos específicos del papel del T3SS en la interacción de *Salmonella* con plantas, como la activación de la expresión del mismo, o la función de las proteínas exportadas, están aún por dilucidarse (Wiedemann *et al.*, 2015). La determinación de dichos mecanismos ayudará a comprender el tipo de asociación que establecen las enterobacterias con las plantas, en especial si atendemos a la importancia que tienen los T3SS como factores de virulencia en ciertas bacterias fitopatógenas. *Pseudomonas syringae*, por ejemplo, transloca efectores a través del T3SS que alteran el correcto desarrollo y suprimen el sistema inmune en plantas de tabaco y tomate, entre otras (Block *et al.*, 2008). Tanto el T3SS como sus efectores asociados y otras herramientas moleculares definen el tipo de asociación bacteriana con las distintas plantas que infectan, y son objeto de intenso estudio actualmente.

Dado su peso específico en este trabajo, a continuación se desarrollan en mayor detalle algunas de las interacciones establecidas entre bacterias fitopatógenas y sus plantas huésped.

2.1. Enfermedades vegetales producidas por bacterias fitopatógenas

Como parte de su mecanismo de adaptación, algunas bacterias son capaces de invadir los tejidos de ciertas plantas y, en ocasiones, producir enfermedades. La pérdida o contaminación de cultivos por enfermedades agrícolas no sólo tiene un impacto agroalimentario evidente, sino también económico. La agricultura es en muchos países el principal motor de la economía, con lo que la propagación de plagas vegetales pueden tener consecuencias devastadoras. Desde que en 1878 T. J. Burril demostrara que el fuego bacteriano del peral era causa directa de la acción de la bacteria *Erwinia amylovora* (Burrill, 1878), se han descrito incontables infecciones causadas por diversas bacterias fitopatógenas, así como muchos de los mecanismos moleculares implicados en los procesos de infección de plantas (Vidaver, Lambrecht, 2004; Mansfield *et al.*, 2012). En los siguientes apartados se revisan algunas enfermedades relevantes de plantas y sus correspondientes agentes causales.

Mancha bacteriana

Es el tipo más frecuente de enfermedad vegetal, caracterizándose por la aparición de manchas de distintos tamaños ya sea en hojas, flores, tallo o frutos. Cuando las manchas se propagan con cierta celeridad la infección da lugar a la llamada quema bacteriana, que en ocasiones puede destruir prácticamente la totalidad de la superficie de la planta, marchitándola y provocando la muerte de muchos tejidos (Agrios, 2005). La mayoría de las manchas y quemas bacterianas tienen su origen en la infección de cepas pertenecientes a los géneros *Pseudomonas* y *Xanthomonas*, especialmente las causadas por patovares (pvs.) de las especies *P. syringae* y *Xanthomonas campestris*.

P. syringae pv. tomato es el agente causal de la mancha bacteriana del tomate, una enfermedad propagada a nivel mundial que genera significativas pérdidas económicas desde la década de los 70, y que parece favorecerse en climas húmedos y frescos (Agrios, 2005). Su sintomatología se caracteriza por manchas pequeñas y oscuras en las hojas, denominadas "manchas pardas", rodeadas por un halo clorótico amarillo, que con frecuencia producen defoliación. Las manchas aparecen también en flores, tallo y frutos. La patogenicidad de *P. syringae* pv. tomato se basa en el T3SS y sus efectores, así como, entre otros factores, en la síntesis y secreción de distintas moléculas fitotóxicas. Tal es el caso de la coronatina, una fitotoxina producida por varios patovares de *P. syringae* que induce la formación de los halos cloróticos en la planta (Brooks *et al.*, 2005). Se cree que la coronatina mimetiza al ácido

jasmónico, una hormona vegetal implicada en la respuesta a estrés y en diversos procesos de desarrollo (Gfeller *et al.*, 2010). Estudios en la planta modelo *Arabidopsis thaliana* han demostrado que la coronatina altera la respuesta inmune mediada por el ácido jasmónico, favoreciendo la proliferación de la cepa *P. syringae* pv. tomato DC3000 en el interior de la planta (Uppalapati *et al.*, 2005).

Otra mancha bacteriana que provoca considerables pérdidas es la quema bacteriana de la judía, que afecta especialmente a cultivos en zonas cálidas y subtrópicales. Su principal agente causal es la bacteria *Xanthomonas axonopodis* pv. phaseoli, que penetra a través de los estomas de las hojas y heridas y produce manchas necróticas en hojas, tallos, vainas y semilla, también rodeadas de un halo amarillento (Büttner, Bonas, 2010). Al igual que en *P. syringae*, la secreción de biomoléculas juega un papel fundamental en la virulencia de esta bacteria. Una de estas biomoléculas es el xantano, un exopolisacárido con diversas aplicaciones industriales que producen muchas *Xanthomonas* y que confiere una apariencia mucoide a los cultivos bacterianos (Vojnov *et al.*, 2001). La secreción de este compuesto, altamente hidratado y de naturaleza aniónica, ayuda a proteger a la bacteria frente compuestos tóxicos, deshidratación, y otros factores de estrés ambiental. Igualmente, se ha demostrado que el xantano está implicado en la supresión de las defensas de las plantas, así como en la formación de biofilms en la superficie de las mismas (Büttner, Bonas, 2010).

Marchitez bacteriana

La marchitez bacteriana afecta a plantaciones de importancia económica a nivel mundial, en especial a miembros de la familia de las solanáceas, como la patata, el pimiento, la berenjena o el tomate. Tiene su origen en la infección provocada por bacterias de diversos géneros, como *Erwinia*, *Clavibacter*, *Curtobacterium*, *Pantoea*, *Ralstonia* o *Xanthomonas*. En el proceso de infección, la bacteria accede en primera instancia a los tejidos vasculares de la planta por medio de hendiduras naturales o raíces con heridas. Una vez en el interior, coloniza los vasos del xilema, donde interfiere con el transporte de agua y nutrientes, causando la marchitez (Agrios, 2005).

Un ejemplo típico de esta enfermedad es el fuego bacteriano de las rosáceas, cuyo principal agente infeccioso es *E. amylovora* (Kamber *et al.*, 2012). Otro tipo de marchitez es la podredumbre parda de la patata, provocada por el patógeno *Ralstonia solanacearum* (Mansfield *et al.*, 2012). Esta enfermedad es común en los trópicos y en climas calurosos de todo el mundo, y se caracteriza por la marchitez de las hojas superiores, que puede llegar a ser permanente. El xilema queda afectado y se producen exudados blanquecinos, que pueden dar lugar a la muerte de la planta. El genoma de *R. Solanacearum* se compone típicamente de un cromosoma y un megaplásmido de alrededor de 2Mb (Salanoubat *et*

al., 2002), que codifica un T3SS. La translocación de efectores al interior del huésped mediante dicho sistema, como en otras bacterias fitopatógenas, interfiere en el metabolismo de la planta y modula su respuesta inmune (Coll, Valls, 2013). Sin embargo, en este caso la estrategia molecular de estos efectores es en ocasiones notablemente sofisticada. Se ha demostrado que cepas de *R. solanacearum* y *Xanthomonas* spp. translocan los llamados efectores TAL (*transcription activator-like*), que acceden al núcleo de las células de la planta y actúan como factores de transcripción, activando la expresión de genes específicos (Scholze, Boch, 2011). Aunque aún queda por esclarecerse la función de muchos de estos genes "diana" (Lange de *et al.*, 2013), estudios en *Xanthomonas oryzae* pv. oryzae han mostrado que el efector TAL PthXo1 induce la expresión de genes de mediación del flujo de salida de azúcar, contribuyendo a la adquisición de nutrientes de dicha bacteria (Deslandes, Rivas, 2012).

Tumores y agallas

La formación de tumores y agallas en raíces y tallos, o en la zona de unión entre ambas estructuras, se debe principalmente a la capacidad infecciosa de bacterias pertenecientes a los géneros *Agrobacterium*, *Pseudomonas*, *Rhizobacter* y *Rhodococcus*. Los tumores se presentan en diversas formas y tamaños, afectando a árboles de frutas pomáceas y de hueso, así como a vides y zarzamoras (Agrios, 2005). El agente bacteriano por excelencia de este tipo de patología vegetal es *Agrobacterium tumefaciens*, cuyo genoma incluye al menos un plásmido de gran tamaño denominado Ti (*tumor inducing*), responsable de la proliferación del patógeno en el huésped y de la formación de tumores en la zona de unión entre la raíz y el tallo, denominados tumores o agallas del cuello. Mediante un sistema de secreción de tipo IV (T4SS), *A. tumefaciens* transfiere a la planta un fragmento del plásmido Ti, que se integra en su ADN cromosómico (Voth *et al.*, 2012). Este fragmento porta, entre otros, genes encargados de la producción de las hormonas vegetales auxina y citoquinina, así como de opinas, que las bacterias usan a modo de nutriente (Lacroix, Citovsky, 2013). Con ello, *A. tumefaciens* reprograma la síntesis de biomoléculas de la planta para generar un entorno afín. Por un lado, la expresión de auxinas y citoquininas provoca el sobrecrecimiento de los tejidos y deriva en la aparición de tumores. Por otro, las opinas sirven a la bacteria de fuente de carbono y nitrógeno, con lo que disponen del sustento necesario para colonizar el interior de la planta. Los síntomas comienzan con la aparición de pequeñas agallas redondas y blanquecinas y, a medida que la infección avanza, el desarrollo de las agallas da lugar a tumores de consistencia dura, color marrón y superficie irregular. El plásmido Ti de *Agrobacterium* ha supuesto un paradigma para el estudio del T4SS, y su capacidad para transferir ADN ha dado lugar a multitud de aplicaciones en biología vegetal, agricultura y biotecnología (Colquhoun, 1994; Pitzschke, 2013).

Otra patología vegetal que presenta formación de tumores es la tuberculosis del olivo, cuyo patógeno bacteriano responsable es *Pseudomonas savastanoi* pv. savastanoi (Young, 2004; Ramos *et al.*, 2012). Aunque su distribución geográfica es mundial, la enfermedad es muy común en países del área mediterránea, como España, Grecia o Italia, donde la producción de olivo es mayor. Los árboles afectados se caracterizan por una reducción en su crecimiento y producción de frutos, así como una vigorosidad menor (Quesada *et al.*, 2010). Los síntomas incluyen formación de tumores redondos de distintos tamaños en troncos, ramas, brotes y, ocasionalmente, en hojas y frutos Kennelly *et al.* (2007). Como *A. tumefaciens*, *P. savastanoi* pv. savastanoi induce la formación de tumores en la planta mediante auxinas y citoquininas. Sin embargo, en este caso no ocurre transferencia de ADN, de modo que la síntesis y liberación al medio de dichas hormonas es llevada a cabo por la bacteria (Macdonald *et al.*, 1986; Matas *et al.*, 2009). Otro de los factores en que *P. savastanoi* pv. savastanoi basa su patogenicidad es un T3SS y sus efectores asociados. Cepas mutantes en genes implicados en la formación de este sistema presentan una virulencia disminuida en plantas de olivo, así como una reducción en la capacidad de colonizar los tejidos vegetales (Pérez-Martínez *et al.*, 2010; Matas *et al.*, 2012).

Costra bacteriana

La costra bacteriana es una enfermedad que afecta principalmente a las partes subterráneas de la planta, y se caracteriza por lesiones roñosas en los tejidos superficiales. Su principal agente infeccioso es *Streptomyces scabies*, que produce la llamada sarna o roña común de la patata, presente a nivel mundial en la mayoría de las zonas de cultivo de dicha planta (Loria *et al.*, 2006). Sus síntomas incluyen primero pequeñas lesiones en los tubérculos, de color marrón y forma circular, que luego engrosan, se unen, y pueden llegar a hundirse hacia el interior entre 3 y 4 milímetros. *S. scabies* basa su virulencia en la secreción de fitotoxinas como la taxtomina, que inhibe la síntesis de célulosa de la planta (Bignell *et al.*, 2014), así como de hormonas vegetales, como auxinas.

Podredumbre blanda

La podredumbre blanda es una enfermedad vegetal presente en cultivos a nivel mundial, ya sea en climas templados, tropicales o subtropicales.La infección afecta a patata, zanahoria, tabaco, cebolla y tomate, entre otros cultivos, y está causada principalmente por cepas de *Erwinia*, *Pseudomonas*, *Bacillus* y *Clostridium* (Agrios, 2005). Dada su capacidad infecciosa en patatas y las cuantiosas pérdidas que provoca, el patógeno bacteriano más representativo de la podredumbre blanda es *Dickeya dadantii* (Mansfield *et al.*, 2012). La sintomatología empieza en forma de lesiones acuosas cuyo diámetro engrosa con rapidez,

dando paso a la maceración del tejido afectado, que adquiere un color pardo, se ablanda y se vuelve viscoso, además de desprender un mal olor característico (Agrios, 2005). *D. dadantii* tiene la particularidad de provocar infecciones de carácter latente, que se activan en postcosecha, donde el tejido vegetal está en fase de senescencia. La susceptibilidad de la planta al ataque de patógenos en esta fase es mayor, lo que hace de *D. dadantii* una bacteria especialmente perniciosa. Su adaptación al interior de la planta depende de mecanismos como la captación de hierro, para lo que sintetiza los sideróforos crisobactina y acromobactina (Münzinger *et al.*, 2000). Asimismo, la defensa contra sustancias tóxicas vegetales juega un papel importante, y *D. dadantii* cuenta con una batería de genes que codifican bombas de eflujo, encargadas de expulsar compuestos tóxicos extraños del espacio intracelular bacteriano (Maggiorani Valecillos *et al.*, 2006). Pero el factor de más peso en la virulencia de este patógeno es la secreción de enzimas hidrolíticas, como pectato liasas, poligalacturonasas y pectin metil esterasas, que degradan la pared celular vegetal y que contribuyen notablemente a la maceración de los tejidos (Barras *et al.*, 1994).

Cancro bacteriano

Aunque los cancros descritos en plantas por causa de patógenos bacterianos son infrecuentes, algunos de ellos pueden tener efectos devastadores en algunos cultivos. Las bacterias implicadas suelen ser patovares de *X. axonopodis* y *P. syringae*, que infectan cítricos y árboles frutales de pepita o hueso, respectivamente (Agrios, 2005). *X. axonopodis* pv. citri es el agente causal de cancro cítrico, una de las enfermedades más perniciosas que afectan al cultivo de agrios, siendo susceptibles a ella la mayoría de las variedades comerciales de estas plantas (Graham *et al.*, 2004). Su distribución geográfica es mundial, aunque se cree que tiene su origen en el sudeste asiático. La sintomatología se caracteriza por defoliación y manchas en los frutos, que caen prematuramente. Las ramas mueren de forma regresiva y se produce un debilitamiento general del árbol. De entre los factores que explican la virulencia de *X. axonopodis* pv. citri destacan la secreción de efectores vía T3SS, así como adhesinas implicadas en la formación de biofilms (Zimaro *et al.*, 2013, 2014). Un estudio reciente demostró que *X. axonopodis* pv. citri sintetiza un péptido natriurético vegetal, llamado XacPNP, que altera las condiciones homeostáticas de la planta, favoreciendo la supervivencia de la bacteria en el interior de la misma (Gottig *et al.*, 2008).

Necrosis apical

La necrosis apical es una patología vegetal distribuida en zonas templadas a nivel mundial, afectando a plantas leñosas. Aunque está asociada principalmente a la bacteria *P. syringae* pv. syringae, como en los casos del peral o el mango (Montesinos, Vilardell,

1991; Cazorla *et al.*, 1998), también se han descrito necrosis apicales producidas por cepas de *Pantoea agglomerans* y *Xanthomonas arboricola* (Yang *et al.*, 2011; Moragrega *et al.*, 2011).

El ejemplo más representativo de esta enfermedad es la necrosis apical del mango, presente en cultivos de esta planta en zonas subtropicales del área mediterránea, Australia y Estados Unidos. El mango (*Mangifera indica* L.) es una especie arbórea de la familia de las anacardiáceas, normalmente de gran altura (25-30 metros) y con una copa densa y redondeada (Bally, 2006). Es una planta leñosa, con un tronco robusto de corteza áspera y gruesa y de hojas verde oscuro. Aunque tiene su origen en India y Birmania, en la actualidad su cultivo se extiende por zonas tropicales y subtropicales de todo el mundo. En España, el número de plantaciones de mango ha aumentado considerablemente en las últimas décadas, concentrándose gran parte de las mismas en la costa subtropical andaluza y Canarias. La necrosis apical del mango es a día de hoy una de las patologías más perniciosas que afectan a su cultivo en esta zona, y fue descrita por Cazorla y colaboradores en 1998. El patógeno causante de esta enfermedad es *P. syringae* pv. syringae, que aunque está presente en los cultivos durante todo el año, en estaciones lluviosas y períodos fríos su capacidad infecciosa aumenta, provocando en ocasiones la pérdida de plantaciones enteras. La sintomatología afecta a brotes, hojas, tallos y panículas, aunque no a frutos. Las yemas terminales se necrosan en los brotes, mientras que las hojas adquieren manchas sin halos cloróticos de distintos tamaños. Los tallos presentan también manchas necróticas, pudiendo producir exudaciones blanco-lechosas. En las panículas, la necrosis se distribuye por los ejes principal y secundario, que pueden llegar a secarse por completo (Cazorla *et al.*, 1998).

Pseudomonas syringae pv. syringae UMAF0158, modelo de estudio de la necrosis apical del mango

P. syringae pv. syringae UMAF0158 es una cepa bacteriana, aislada en plantas de mango, que en las últimas dos décadas ha servido como modelo de estudio de la necrosis apical de dicha planta en la costa subtropical andaluza (Cazorla *et al.*, 1998; Arrebola *et al.*, 2012; Gutiérrez-Barranquero *et al.*, 2013a). Parte esencial de la virulencia de UMAF0158 en plantas de mango reside en la producción de mangotoxina, una toxina antimetabolito que inhibe la enzima ornitina N-acetiltransferasa, interfiriendo en la ruta de biosíntesis de ornitina y arginina (Arrebola *et al.*, 2003). La síntesis de mangotoxina se ha observado en el 87% de las cepas de *P. syringae* pv. syringae aisladas de mango, y en ella intervienen los operones *mgo* y *mbo*, de 4 y 6 genes, respectivamente (Arrebola *et al.*, 2012; Carrión *et al.*, 2013). Tanto la caracterización de esta toxina como el estudio de su contribución a la virulencia y el estilo de vida epífito de *P. syringae* pv. syringae en plantas de mango se ha llevado a cabo usando UMAF0158 como cepa modelo.

Recientemente, se ha obtenido la secuencia completa del genoma de UMAF0158, cuyo análisis bioinformático se ha abordado en el presente trabajo. Dicho análisis ha permitido identificar una batería de factores genéticos potencialmente implicados en la capacidad de esta cepa para infectar mango. Posteriores estudios funcionales ayudarán a diseñar estrategias de control para paliar los efectos de este patógeno en la costa subtropical andaluza.

2.2. Bacterias beneficiosas y control biológico de enfermedades vegetales

Como hemos visto, los agentes fitopatógenos son el origen de multitud de enfermedades infecciosas que afectan a cultivos de frutas y verduras, viéndose reducida significativamente su producción. No sólo bacterias, sino también insectos, hongos, virus y viroides, entre otros, son responsables directos de una amplia suerte de patologías vegetales, en ocasiones con consecuencias agroeconómicas devastadoras (Agrios, 2005). La solución convencional para controlar las poblaciones de estos patógenos es la aplicación de plaguicidas químicos sintetizados artificialmente, que previenen o destruyen las plagas para las que están diseñados, pero pueden tener efectos negativos sobre el medio ambiente, la sanidad y la propia producción agrícola (Santos *et al.*, 2010). Al albur de los intereses comerciales de determinadas empresas, durante años se ha promovido la venta de plaguicidas sin informar convenientemente de los riesgos que generan, lo que ha derivado en un sobreuso cuanto menos irracional. Estas malas prácticas han supuesto un grave perjuicio para la salud pública y un costo social del todo inadmisible. Según la OMS, las intoxicaciones con plaguicidas químicos oscilan entre 500.000 y 1 millón de casos anuales, estimándose el número de fallecidos por intoxicación entre 5.000 y 20.000 (Wesseling *et al.*, 1997; Eddleston *et al.*, 2002). Además, la aplicación de plaguicidas puede ser perniciosa para el agroecosistema en que se emplean, siendo especialmente grave su efecto contra los enemigos naturales de los agentes fitopatógenos. La toxicidad de los plaguicidas puede desproteger a los cultivos de organismos de control biológico, que se dan de forma natural y que proporcionan una ayuda esencial a la planta contra algunos agentes infecciosos. Esto puede derivar en un aumento de las poblaciones de patógenos que lleguen de nuevo al cultivo o que no hayan sido eliminados completamente por los plaguicidas, resultando en el efecto contrario al deseado (Santos *et al.*, 2010).

Dadas las consecuencias que resultan del uso de plaguicidas, emergen con fuerza soluciones alternativas que tratan de equilibrar la eficacia del tratamiento con la sostenibilidad del ecosistema en que se va a aplicar. Tal es el caso de las estrategias de control integrado,

que combinan formas diferentes de control, como control biológico, físico, químico y cultural, para paliar enfermedades vegetales de forma viable económicamente y minimizando el impacto ecológico (Lewis, Papavizas, 1991). En lo que respecta al control biológico, éste introduce un elemento adicional en el triángulo de interacción planta-patógeno-ambiente, los organismos antagonistas. De entre estos, cabe destacar determinados microorganismos, ya sean naturales o modificados, que permiten reducir los efectos indeseables de los patógenos vegetales, al tiempo que favorecen el desarrollo de otros microorganismos beneficiosos para las plantas, como los enemigos naturales. Dado su peso específico en este trabajo, en lo que sigue hacemos énfasis en el uso de bacterias como agentes de biocontrol y promotores del crecimiento vegetal.

Rizobacterias, biocontrol y promoción del crecimiento vegetal

La rizosfera comprende una zona del suelo en que las interacciones entre las raíces de las plantas y los microorganismos existentes son de un dinamismo y especificidad excepcionales. La acumulación de exudados vegetales genera un entorno rico en nutrientes, que sirven de reclamo para los microorganismos que habitan el suelo, resultando en un aumento en la concentración de biomasa y actividad microbiana (Philippot *et al.*, 2013; McNear, 2012). Se cree que entre el 1 y el 2% de las bacterias que habitan la rizosfera promueven directa o indirectamente el crecimiento vegetal (Bonfante, Anca, 2009), destacando los géneros *Bacillus* y *Pseudomonas* como los más prominentes (Kloepper, 1981). Una forma de contribución directa a este crecimiento viene dada por la capacidad de algunas bacterias de incrementar la disponibilidad de nutrientes que la planta necesita. Un ejemplo de ello es la fijación biológica de nitrógeno que, como hemos visto, llevan a cabo determinadas bacterias, que incorporan este elemento a la rizosfera y proveen a las plantas del sustento necesario para crecer (Rodrigues *et al.*, 2008; Hatayama *et al.*, 2005). En ocasiones la contribución bacteriana viene dada por la síntesis de biomoléculas, típicamente hormonas, implicadas directamente en el crecimiento de la planta. Así como la producción de hormonas por parte de patógenos vegetales puede alterar perniciosamente el crecimiento de la planta, en bacterias promotoras del crecimiento vegetal (PGPR, del inglés, *plant growth promoting rhizobacteria*) la síntesis de estas moléculas ayuda al correcto desarrollo de la misma. Algunas rizobacterias sintetizan citoquininas y giberelinas que intervienen en la madurez de los brotes (Van Loon, 2007), o auxinas, que promueven la formación de raíces laterales (Yang *et al.*, 2009).

Por otro lado, los efectos indirectos de las PGPR se centran en su capacidad de paliar la severidad de determinadas infecciones, así como de presentar actividad antagonista frente a determinados patógenos. Para ello, ciertas bacterias producen compuestos como

antibióticos, toxinas o enzimas líticas que generan un efecto deletéreo en microorganismos fitopatógenos. Un estudio reciente mostró que la síntesis de los antibióticos bacilomicina y macrolactina por parte de la cepa *Bacillus amyloliquefaciens* NJN-6 presenta efectos antagonistas contra *Fusarium oxysporum* y *R. solanacearum*, respectivamente (Yuan *et al.*, 2012). Por otro lado, en el proceso de competencia por los nutrientes en la rizosfera, la síntesis de sideróforos puede suponer una ventaja respecto a potenciales competidores. Una capacidad superior de captar hierro, cuya disponibilidad es limitada en ambientes aeróbicos, puede tener un efecto de antibiosis frente a otros microorganismos, cuyo acceso a este nutriente se verá reducido significativamente, así como su capacidad de proliferar. Se ha demostrado que la producción de pioverdinas, un sideróforo fluorescente sintetizado por algunas *Pseudomonas* (Schalk, Guillon, 2013), ayuda a controlar los efectos de los patógenos *F. oxysporum* en patata y *Gaeumannomyces graminis* en tabaco (Voisard *et al.*, 1989; Schippers *et al.*, 1990). Asimismo, algunas rizobacterias, a medida que colonizan la raíz, inducen en las plantas un mecanismo de resistencia llamado ISR (*Induced Systemic Resistance*), mediante el que la parte aérea de las mismas se protege contra un amplio rango de patógenos. Esta propiedad se describió inicialmente al observarse que la cepa *Pseudomonas fluorescens* WCS417r protegía sistemáticamente plantas de clavel contra *F. oxysporum* f. sp. dianthi (Peer van *et al.*, 1991), así como una selección de rizobacterias inducían resistencia en pepino contra *Colletotrichum orbiculare* (Wei, 1991). Se han descrito diversos determinantes bacterianos capaces de desencadenar ISR, como la flagelina (Ausubel, 2005), el antígeno O de los lipopolisacáridos (Leeman et al., 1995), algunos sideróforos (Ran *et al.*, 2005; Loon van *et al.*, 2008) y ciertos lipopéptidos. Tal es el caso de los lipopéptidos surfactina y fengicina, sintetizados por *Bacillus subtilis*, que activan este mecanismo en plantas de judía contra el patógeno *Botrytis cinerea* (Ongena *et al.*, 2007), agente causal de la podredumbre gris.

Dada su relevancia en esta Tesis Doctoral, el siguiente apartado se centra en una cepa bacteriana probada como agente de control biológico en plantas de olivo, *P. fluorescens* PICF7.

Pseudomonas fluorescens PICF7, agente de biocontrol de la verticilosis del olivo

P. fluorescens PICF7 es una cepa endofita natural de raíces de olivo y eficaz agente de biocontrol de la verticilosis (*Verticillium dahliae* Kleb.), una de las enfermedades más devastadoras que afecta al cultivo de dicha leñosa en la cuenca mediterránea. El hongo penetra en el olivo por heridas o a través de las raíces y llega al xilema, desde el que se expande al resto de la planta. Cuando la enfermedad está muy avanzada, el hongo crece

fuera de los tejidos vasculares, dando lugar a una flacidez diurna y una defoliación intensa de la planta, que acaba por marchitarse de forma permanente hasta llegar en muchos casos a la muerte (López-Escudero, Mercado-Blanco, 2011).

P. fluorescens PICF7 se aisló del suelo de olivares junto con otras siete *Pseudomonas* en un estudio llevado a cabo por el Dr. Jesús Mercado-Blanco y sus colaboradores (Mercado-Blanco *et al.*, 2004). En dicho trabajo, todos los aislados presentaron actividad supresora contra *V. dahliae*, siendo la cepa PICF7 la que exhibió mayor capacidad de biocontrol. Estudios posteriores han confirmado que PICF7 es capaz de controlar eficientemente infecciones provocadas por el patotipo defoliante (D) de *V. dahliae* en plantones de la variedad Picual, que presenta una alta susceptibilidad a la enfermedad. Asimismo, mediante microscopía confocal, disección tridimensional de tejidos y marcaje autofluorescente, se demostró la capacidad de esta cepa de colonizar endofíticamente raíces de olivos de la variedad Arbequina (Prieto, Mercado-Blanco, 2008). Sin embargo, los determinantes bacterianos de PICF7 relacionados tanto con su estilo de vida endofítico como su actividad de biocontrol están aún por dilucidarse. Una primera aproximación se ha centrado en estudiar, mediante mutagénesis dirigida, la implicación de fenotipos como la motilidad tipo "*swimming*" o la producción del sideróforo pioverdina en las propiedades naturales de PICF7. Mutantes en ambos fenotipos no mostraron diferencias significativas en la colonización de las raíces ni el antagonismo de PICF7 contra *V. dahliae* (Maldonado-González *et al.*, 2013). Por otro lado, la planta parece responder a la colonización radicular de PICF7, lo que podría explicar su capacidad de biocontrol. Mediante la metodología *Suppression Subtractive Hybridization* (SSH), se pudo confirmar que PICF7 induce ISR en los tejidos aéreos (Gómez-Lama *et al.*, 2014), así como un amplio espectro de respuestas defensivas en las propias raíces (Schilirò *et al.*, 2012).

Con objeto de comprobar el potencial de PICF7 como agente de biocontrol contra otros patógenos de olivo, un estudio reciente examinó la interacción entre esta cepa y *P. savastanoi* NCPPB 3335, cepa modelo de estudio de infección bacteriana en plantas leñosas (Ramos *et al.*, 2012). El estudio mostró que, aunque PICF7 no es capaz de suprimir el desarrollo de tumores, su co-inoculación con el patógeno produce una disminución de la población de este último *in planta*, dando lugar a síntomas necróticos reducidos y a una alteración de la colonización del tejido hiperplásico (Maldonado-González *et al.*, 2013).

La secuencia del genoma de PICF7 se ha caracterizado recientemente, y el estudio de la misma ha sido abordado en esta Tesis Doctoral. Como consecuencia, se han identificado una serie de componentes genéticos potencialmente implicados en las propiedades fenotípicas de este agente de biocontrol.

3. Bioinformática para el análisis genómico de bacterias asociadas a plantas

Como se ha señalado, la evolución de las metodologías de secuenciación genómica ha sido notable en los últimos años (Zhao, Grant, 2011). Los avances en estas técnicas han supuesto tal reducción en tiempo y costes que han cambiado el paradigma de la investigación biológica (Delanty, Goldstein, 2013; Salto-Tellez, Gonzalez De Castro, 2014). La facilidad con que se puede acceder al genoma de prácticamente cualquier organismo vivo ha hecho que especialistas de todas las áreas secuencien, casi rutinariamente, los genomas de aquellos organismos que les son de interés (Rothberg, Leamon, 2008). La fitobacteriología no es una excepción, y en la actualidad existe un repertorio extenso de genomas de bacterias asociadas a plantas accesibles en bases de datos públicas (Hamilton *et al.*, 2011; Winsor *et al.*, 2011). Los mecanismos que dictan el estilo de vida de estas bacterias responden en gran medida a su información genética, por lo que profundizar en el conocimiento de los mismos pasa necesariamente por el estudio de sus genomas.

Aunque el desarrollo de la secuenciación masiva es relativamente reciente (Margulies *et al.*, 2005), los bacteriólogos llevan décadas caracterizando genes involucrados en las interacciones que las bacterias establecen con sus plantas huésped. Los procedimientos de secuenciación usados, rudimentarios en comparación con los actuales, han permitido caracterizar un amplio repertorio de componentes genéticos bacterianos, que han ayudado en gran medida a describir mecanismos específicos de patogénesis, de control biológico, de promoción del crecimiento vegetal, etc. En estos procesos de asociación intervienen multitud de factores, muchos de ellos esenciales para que la interacción surta efecto. Factores como adhesinas o exopolisacáridos son fundamentales en la fase de adherencia a los tejidos de la planta (Pizarro-Cerdá, Cossart, 2006; Mhedbi-Hajri *et al.*, 2011). Asimismo, para subsistir en dichos tejidos muchas bacterias disponen de sideróforos, antibióticos o bombas de extrusión multidroga, necesarios para competir por los nutrientes y protegerse de compuestos tóxicos antimicrobianos (Ahmed, Holmström, 2014; Wang, Raaijmakers, 2004; Martínez *et al.*, 2009). Por otro lado, factores como fitotoxinas, fitohormonas o enzimas degradadoras de la pared celular vegetal juegan un papel esencial en la colonización del huésped (Dudler, 2014; Costacurta, Vanderleyden, 1995; Barras *et al.*, 1994). La determinación de las secuencias genéticas encargadas de la síntesis de estos factores ha contribuido de forma significativa al estudio de las interacciones planta-bacteria, y con los avances en las técnicas de secuenciación, la identificación de nuevos factores se ha facilitado notablemente.

Algunos de los factores anteriormente mencionados son transportados al interior del huésped, para lo que la bacteria dispone de herramientas específicas encargadas de la secreción de moléculas al exterior. Tal es el caso de los sistemas de secreción, que permiten a las bacterias transportar proteínas al medio extracelular, así como al interior de otras células eucariotas o procariotas (Tseng *et al.*, 2009). En el contexto de las interacciones planta-bacteria, son de especial relevancia los sistemas de secreción de tipos III, IV y VI (T3SS, T4SS y T6SS, respectivamente). Algunos ejemplos del papel que juegan tanto el T3SS como el T4SS en las asociaciones que establecen ciertas bacterias con sus plantas huésped se mencionaron en el apartado 2.1. Por otro lado, el T6SS es un sistema de reciente descripción, cuyos componentes se asemejan estructuralmente al bacteriófago T4 (Silverman *et al.*, 2012). Como el T3SS y el T4SS, el T6SS transloca proteínas al exterior, y aunque está presente en una proporción considerable de bacterias Gram-negativas (Bingle *et al.*, 2008; Boyer *et al.*, 2009), existe todavía poca evidencia experimental que describa el modo de ensamblaje de este sistema, así como las funciones específicas de las proteínas secretadas. Estudios funcionales de sus efectores han mostrado que el T6SS no sólo juega un papel importante en la virulencia de ciertas bacterias, sino también en las relaciones mutualistas y comensalistas que éstas establecen con determinados organismos eucariotas (Jani, Cotter, 2010). Asimismo, trabajos recientes han indicado que efectores del T6SS tienen función antibiótica frente a otros microorganismos, ayudando en el proceso de competencia de algunas bacterias (Dong *et al.*, 2013).

Dada la relevancia que tienen los sistemas de secreción en el estilo de vida de las bacterias, se han caracterizado también muchos de los genes implicados en la síntesis tanto de sus componentes estructurales como de sus efectores asociados. En suma, y aunque algunos de los mecanismos involucrados en las interacciones planta-bacteria están aún por dilucisarse, el número de genes bacterianos conocidos que intervienen de una forma u otra no es nada desdeñable. Si consideamos además el conjunto creciente de genomas bacterianos disponibles, la cantidad de información resultante es inmensa. Desde un punto de vista computacional, dicha información resulta muy útil, en tanto que cubre un amplio espectro del conocimiento existente de las interacciones planta-bacteria, y puede usarse para generar nuevas hipótesis en este campo. Una aplicación directa de estos datos es el desarrollo de herramientas específicas de anotación, capaces de escanear secuencias de entrada en busca de genes homólogos a los ya conocidos y descritos en bacterias asociadas a plantas. La anotación generada podría usarse para direccionar posteriores análisis experimentales en base a los genes identificados. El desarrollo de este tipo de herramientas ha sido abordado en esta Tesis Doctoral.

Otro aspecto interesante a la hora de tratar la ingente cantidad de información disponible es que ésta ofrece la oportunidad de abordar la genómica de las interacciones planta-bacteria desde un punto de vista global. Si el modo de vida de una bacteria está de alguna forma codificado en su material genético, sería interesante comprobar hasta qué punto la información genómica disponible es suficiente para identificar las distintas asociaciones que establecen las bacterias con las plantas. En este contexto, los métodos de aprendizaje automático son particularmente útiles, ya que, como se mencionó en el apartado 1.4, permiten usar la información existente para inducir modelos de representación. De este modo, podríamos entrenar modelos en base a los datos genómicos conocidos para generar clasificadores que nos permitan agrupar genomas de bacterias con estilos de vida similares, como patógenas o comensalistas. Este tipo de aproximación ya ha sido aplicada a genomas de bacterias asociadas a humanos con resultados prometedores, aunque el límite parece estar en discernir bacterias patógenas y no patógenas (Iraola *et al.*, 2012; Andreatta *et al.*, 2010; Barbosa *et al.*, 2014). En lo que respecta a las interacciones planta-bacteria, una herramienta de predicción de estilos de vida basada en información genómica sería de gran utilidad, no sólo en el ámbito de la fitopatología, sino también en el de la seguridad alimentaria. En la última década, se han dado diversas epidemias relacionadas con el consumo de productos vegetales, afectando a miles de personas en todo el mundo. Sorprendentemente, los patógenos responsables han sido cepas de bacterias tradicionalmente asociadas al consumo de alimentos de origen animal, como *Escherichia coli* o *Salmonella* (Deering *et al.*, 2012; Lim *et al.*, 2014). Teniendo en cuenta este escenario, en el presente trabajo se ha abordado el diseño de un método que, en base a la información genómica disponible, identifica interacciones potenciales entre bacterias y plantas.

OBJETIVOS

Esta Tesis Doctoral tiene un doble cometido: por un lado, analizar las secuencias genómicas de dos cepas bacterianas asociadas a cultivos de relevancia en la agricultura andaluza, haciendo hincapié en los factores genéticos implicados en la asociación de las mismas con sus plantas huésped; por otro, desarrollar nuevas herramientas computacionales basadas en la información existente de las interacciones planta-bacteria que permitan extraer nuevo conocimiento en este campo. Para ello se establecieron los siguientes objetivos específicos:

1. **Analizar el genoma secuenciado de la cepa *Pseudomonas syringae* pv. syringae UMAF0158, agente causal de la necrosis apical del mango**, poniendo especial énfasis en genes potencialmente implicados en la síntesis de factores de virulencia.

2. **Analizar el genoma secuenciado de la cepa *Pseudomonas fluorescens* PICF7, endofita de olivo y agente de biocontrol contra la verticilosis**, destacando genes potencialmente implicados en biocontrol y endofitismo.

3. **Desarrollar herramientas bioinformáticas de anotación de genes bacterianos implicados en las interacciones planta-bacteria**.

4. Combinar dichas herremientas con técnicas de aprendizaje automático para **implementar un sistema automatizado de clasificación de genomas bacterianos** que permita identificar potenciales asociaciones planta-bacteria.

CHAPTER I

Bioinformatics analysis of the complete genome sequence of the mango tree pathogen *Pseudomonas syringae* pv. syringae UMAF0158 reveals traits relevant to virulence and epiphytic lifestyle

Pedro Manuel Martínez-García, Pablo Rodríguez-Palenzuela, Eva Arrebola, Víctor J. Carrión, José Antonio Gutiérrez-Barranquero, Alejandro Pérez-García, Cayo Ramos, Francisco M. Cazorla and Antonio de Vicente. Bioinformatics analysis of the complete genome sequence of the mango tree pathogen *Pseudomonas syringae* pv. syringae UMAF0158 reveals traits relevant to virulence and epiphytic lifestyle. Under revision with *PLoS ONE*.

Summary

Pseudomonas syringae pv. syringae (Pss) is the causal agent of the apical necrosis of mango trees, a disease causing severe economic losses worldwide. The complete sequence of the Pss UMAF0158 chromosome was determined, annotated and analyzed. The 5.8-Mb chromosome has an average 59.3% GC content and comprises 5,017 predicted protein-coding genes. Bioinformatics analysis revealed the presence of putative genes potentially implicated in the virulence and epiphytic fitness of this strain, including siderophores, adhesins, exopolysaccharides, phytohormones, phytotoxins, secretion systems, plant cell wall degrading enzymes and detoxifying compounds. Comparative genomics with related *P. syringae* strains were also performed. Among the other 3 completely sequenced *P. syringae* isolates, namely, DC3000, 1448A and B728a, the latter showed the highest sequence similarity to UMAF0158. Genetic features of UMAF0158 that are absent in B728a that could explain UMAF0158 ability to infect mango trees include the mangotoxin biosynthetic operon *mbo*, a cellulose production operon and two additional secretion systems of types III and VI, respectively.

Introduction

Pseudomonas syringae is a species of Gram-negative plant pathogenic bacteria that cause disease in many agriculturally important crops (O'Brien *et al.*, 2011). *P. syringae* infection provokes a variety of symptoms, such as leaf spots and blights, soft rots, stem knots, scabs or cankers, and leads to severe economic losses worldwide (Agrios, 2005). Based on plant pathogenicity test and host specificity, strains belonging to the *P. syringae* complex are subdivided into 57 pathovars (Bull *et al.*, 2010). In contrast, DNA hybridization segregated the *P. syringae* complex into at least nine different genomospecies (Gardan *et al.*, 1999; Scortichini, Marcelletti, 2014) and multilocus sequence typing (MLST) permit the classification into 13 phylogroups, including 23 differentiated clades (Berge *et al.*, 2014). Notably, the broad host range of the species as a whole contrasts with that of individual isolates, which can exhibit virulence potential in a diverse range or in a limited set of plant hosts (Morris *et al.*, 2000; Sarkar *et al.*, 2006), but this question frequently is not clear because the experimental data about the host range are not available (Lamichhane *et al.*, 2014). Moreover, many pathovars associated with unrelated plants are grouped together, sometimes even within the same clade (Berge *et al.*, 2014). Among plant pathogenic bacteria, *P. syringae* is an ideal system to study how evolutionary forces shape adaptation to different hosts, which makes it an archetype of plant-pathogen interactions (Sarkar *et al.*, 2006; Baltrus *et al.*, 2011).

To date, the complete genomes of three *P. syringae* isolates have been sequenced, including *P. syringae* pv. tomato DC3000 (DC3000), *P. syringae* pv. syringae B728a (B728a) and *P. syringae* pv. phaseolicola 1448A (1448A) (Buell *et al.*, 2003; Feil *et al.*, 2005; Joardar *et al.*, 2005). DC3000, a pathogen that infects both tomato and *Arabidopsis*, is the causal agent of bacterial speck disease. In addition, both B728a and 1448A infect bean, but show significantly different phenotypes. B728a produces brown spot disease and exhibits an extensive epiphytic phase. 1448A is the seed-borne causal agent of halo blight for bean, which is a calamitous disease in a number of first-world countries. Many other *P. syringae* draft genome sequences exist for isolates from a great diversity of host plants. These genomes exhibit drastic differences from each other, particularly in the presence/absence of virulence factors-associated genes (Baltrus *et al.*, 2011). Of these, the major determinants of pathogenesis include the effector proteins secreted through the type III secretion systems (T3SS) (Lindeberg *et al.*, 2012). Once inside plants, effectors have the ability to promote virulence by disrupting and suppressing host immune signals. The super repertoire of effectors in the pangenome of the *P. syringae* species complex comprises approximately 60 T3SS effector (T3E) protein families with DC3000 having the largest number of validated

T3Es (Baltrus *et al.*, 2011). Furthermore, the host range of a given strain is thought to be mainly structured by the repertoire of the T3Es it encodes (Lindeberg *et al.*, 2012). Which combination of effectors and other virulence-associated genes define the overall plasticity of host ranges is a question that remains unanswered.

P. syringae pv. syringae includes a diverse collection of strains isolated from different environments, many of them pathogenic on a variety of plant hosts, but showing different host range (Morris *et al.*, 2000; Berge *et al.*, 2014). Some strains of pv. syringae have a broad host range, and it includes strains that cause mango tree apical necrosis, a disease affecting buds, leaves, and stems that has important economic impact worldwide (Cazorla *et al.*, 1998). Its pathogenic arsenal includes several T3Es and virulence factors, but the production of phytotoxins plays a crucial role during symptom development. Phytotoxins promote virulence by disrupting host metabolism and mimicking plant hormones (Bender *et al.*, 1999). Such is the case for mangotoxin, an antimetabolite toxin encoded by several pathovars of *P. syringae* genomospecies 1, which is produced in the early exponential growth phase and inhibits the enzyme ornithine N-acetyl transferase (Arrebola *et al.*, 2003). Production of mangotoxin, which has been observed in more than 87% of the *P. syringae* pv. syringae strains isolated from mango tissues, requires involvement of the *mgo* and *mbo* operons, whose contribution to virulence has been mostly analyzed in the model strain UMAF0158 (Arrebola *et al.*, 2012; Carrión *et al.*, 2013). Interestingly, phylogenetic analyses grouped *P. syringae* pv. syringae strains isolated from mango with mangotoxin-producing strains isolated from other plants (Gutiérrez-Barranquero *et al.*, 2013a). Unlike herbaceous hosts, mango trees provide infection and overwintering sites, which are unique to woody perennials. This characteristic is reflected in the lifestyle of mango-associated *P. syringae* strains and the way they induce disease symptoms in plants (Kennelly *et al.*, 2007). In this regard, genome sequencing and comparative genomics are useful tools for the identification of genetic elements of *P. syringae* complex strains that enable colonization of woody plants as olive, horse chestnut, plum, maple tree or kiwi (Rodríguez-Palenzuela *et al.*, 2010; Green *et al.*, 2010; Baltrus *et al.*, 2011; Bartoli *et al.*, 2014), which will provide insight into the interactions of *P. syringae* and woody plants.

Here, we report the complete sequence, annotation and bioinformatics analysis of the *P. syringae* pv. syringae UMAF0158 genome (chromosome and pPSS158 plasmid), highlighting the virulence and plant interaction genetic background of this pathogenic bacteria on mango trees. We also compared its genome with those of other 25 sequenced *P. syringae* strains from the *P. syringae* complex with a special focus on *P. syringae* pv. syringae B728a, which shares approximately 91% of the UMAF1058 protein coding genes,

but they grouped into different pv. syringae phylotypes (Gutiérrez-Barranquero *et al.*, 2013a) and clades of phylogroup 2 (Berge *et al.*, 2014). Our analysis reveals the existence of genetic differences between these two strains, which may confer UMAF0158 its ability to infect mango trees. In addition to the presence of the mangotoxin biosynthetic operon *mbo*, the UMAF0158 genome differs from that of B728a in the codification of a cellulose synthase operon and in harboring two additional secretion systems, i.e., a T3SS and a T6SS. Moreover, UMAF0158 displays a different repertoire of T3Es, which may be a determinant of its association with mango trees. Our data provide the basis for further functional studies of the virulence mechanisms and host specificity determinants of the *P. syringae* pv. syringae UMAF1058, a representative strain of phylotype 1 of this pathovar, which is the fourth *P. syringae* strain whose complete genome sequence has been made available.

Results and Discussion

General features

The *P. syringae* pv. syringae UMAF0158 genome is composed of one circular chromosome of 5,787,986 bp (Table 1; Figure 1) and one plasmid. The pPSS158 plasmid of 63,004 bp has an average GC content of 54.6% and 71 predicted coding sequences (CDSs) (Table 1). Among the latter, *repA*, a T4SS conjugative system and the *rulAB* genes appear as the most relevant features (Table S1). Conjugative plasmids harboring *rulAB* genes have been shown to contribute to UV and solar radiations tolerance, as well as to epiphytic fitness (Cazorla *et al.*, 2008). In total, 5017 CDSs were identified within the UMAF0158 chromosome, which has an average GC content of 59.3% (Table 1). Among the predicted chromosomal CDSs, a putative function was assigned to 4030 (80%), while the remaining 987 CDSs were designated as hypothetical proteins. A total of 18 genes were predicted to be pseudogenes. The classification of the UMAF0158 CDSs into functional categories according to the COG (Clusters of Orthologous Groups) database is summarized in Table 2 in comparison with *P. syringae* pv. syringae B728a, *P. syringae* pv. tomato DC3000 and *P. syringae* pv. phaseolicola 1448A. With the exception of category L (replication, recombination and repair), which includes a reduced number of UMAF0158 CDSs (137) in comparison with those of the other three genomes (187, 272 and 370 CDSs for B728a, 1448A and DC3000, respectively), no significant differences were found regarding the remaining functional categories, further supporting the relatedness of these four strains. In addition to CDSs, a total of 63 tRNAs and 5 rRNA operons were found on the UMAF0158 chromosome.

Phylogeny

In order to establish the phylogenetic relationship between UMAF0158 and other related *P. syringae* strains, we selected 25 genome sequenced strains (Table S2) and compared a set of five protein-coding house-keeping genes, namely *gapA*, *gltA*, *recA*, *rpoA* and *rpoB*. We created an alignment of the proteins and reconstructed the phylogenetic tree shown in Figure 2, using neighbor-joining methods. The strain *Pseudomonas fluorescens* Pf-5 was used as an out-group. The resulting phylogeny clustered UMAF0158 with *P. syringae* Cit 7, a strain originally isolated from a healthy orange tree (Lindow, 1985; Baltrus *et al.*, 2011), and more separate from B728a, the model strain for pv. syringae. This result is in agreement with previous phylogenetic analyses, which clustered both UMAF0158 and Cit 7 in phylotype 1 (Carrión *et al.*, 2013; Gutiérrez-Barranquero *et al.*, 2013a).

Such a phylotype of the pathovar syringae is mainly associated with the mango host and characterized by mangotoxin production. Additionally, UMAF0158 and other strains of phylotype 1 are pathogenic on mango, lilac, tomato or pear (Cazorla *et al.*, 1998), but not on bean; in contrast, B728a is pathogenic on bean but not on mango (Feil *et al.*, 2005; Gutiérrez-Barranquero *et al.*, 2013a). Given that UMAF0158 is the only strain belonging to this group whose complete genome is available, it could be taken as a representative of phylotype 1 for pv. syringae.

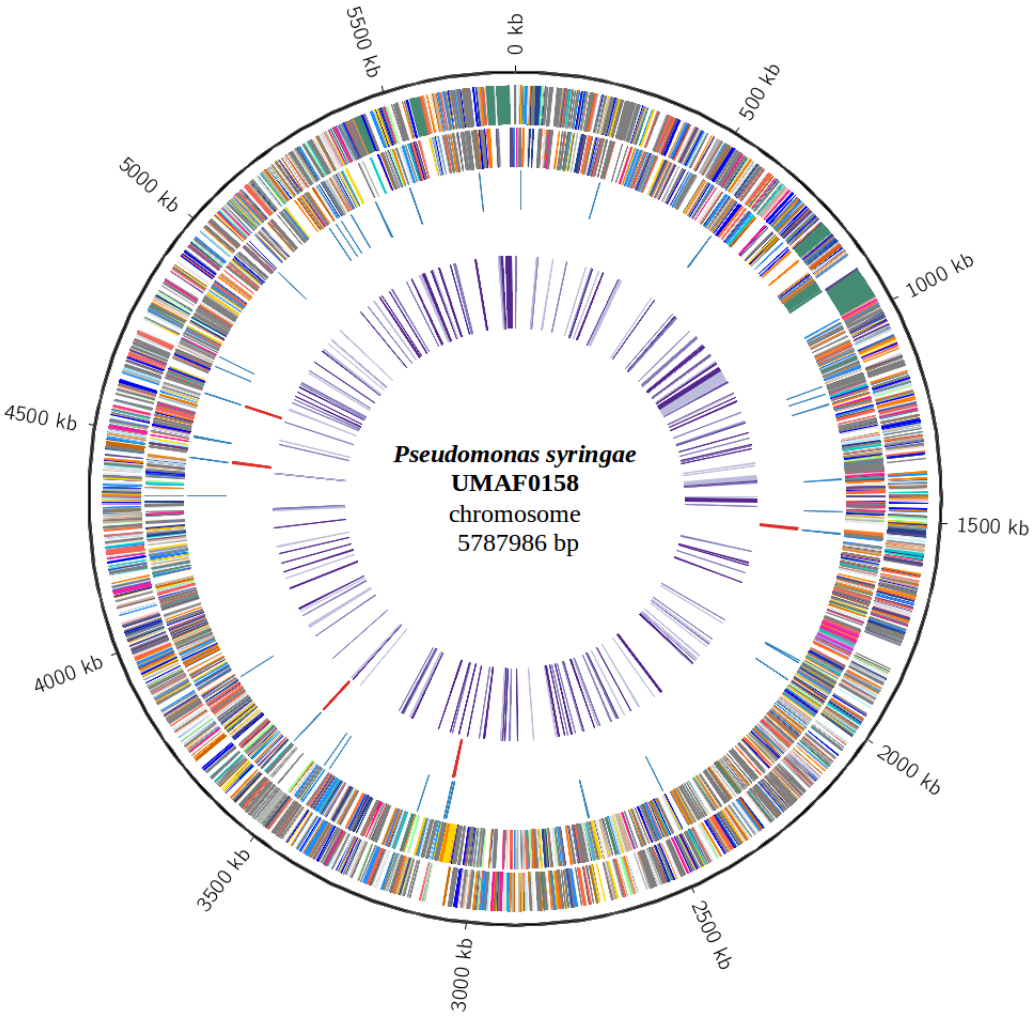


Figure 1. Features of the *P. syringae* pv. syringae UMAF0158 chromosome. From the outside in, the outermost circle (black) shows the scale line; circles 2 and 3 represent predicted coding regions on the plus and minus strand, respectively, which are color coded based on COG categories; circles 4 and 5 show tRNA (blue) and rRNA (red), respectively; circle 6 depicts ORFs associated with virulence (see Virulence factors section and Table S6).

Table 1. General features of the *P. syringae* pv. syringae UMAF0158 genome and comparison with *P. syringae* pv. syringae B728a, *P. syringae* pv. phaseolicola 1448A and *P. syringae* pv. tomato DC3000.

	UMAF0158		**B728a**		**1448A**			**DC3000**	
Molecule	Chromosome	pPSS158	Chromosome	Chromosome	p1448A-A	p1448A-B	Chromosome	pDC3000A	pDC3000B
Size (bp)	5,787,986	63,004	6,093,698	5,928,787	131,950	51,711	6,397,126	73,661	67,473
G+C content (%)	59.3	54.6	59.2	58	54.1	56	58.4	55.1	56.1
No. of predicted CDSs	5,017	71	5,089*	4,985*	127*	60*	5,481*	68*	70*
No. of rRNAs	16	-	16	16	-	-	15	-	-
No. of tRNAs	63	-	64	64	-	-	63	-	-
Reference	This study		Feil *et al.* (2005)	Joardar *et al.* (2005)			Buell *et al.* (2003)		

* The number of predicted CDSs corresponds to those indicated at NCBI for the corresponding genome sequences (March 1st, 2015).

Table 2. Number of CDSs associated with COG functional categories in the *P. syringae* pv. syringae UMAF0158 genome and comparison with *P. syringae* pv. syringae B728a, *P. syringae* pv. phaseolicola 1448A and *P. syringae* pv. tomato DC3000.

	Functional category	UMAF0158	B728a	1448A	DC3000
A	RNA processing and modification	1	1	1	1
B	Chromatin structure	1	1	1	1
C	Energy production and conversion	223	231	210	234
D	Cell cycle control, cell division, chromosome partitioning	38	46	41	44
E	Amino acid transport and metabolism	459	467	458	461
F	Nucleotide transport and metabolism	89	87	88	81
G	Carbohydrate transport and metabolism	276	268	263	264
H	Coenzyme transport and metabolism	177	178	179	173
I	Lipid transport and metabolism	163	162	166	176
J	Translation, ribosomal structure and biogenesis	198	205	204	201
K	Transcription	360	360	343	367
L	Replication, recombination and repair	137	187	272	370
M	Cell wall/membrane/envelope biogenesis	270	288	270	263
N	Cell motility	163	166	168	160
O	Posttranslational modification, protein turnover, chaperones	156	157	152	157
P	Inorganic ion transport and metabolism	272	277	287	278
Q	Secondary metabolites biosynthesis, transport, and catabolism	119	128	112	118
R	General function prediction only	521	522	520	544
S	Function unknown	385	393	356	406
T	Signal transduction mechanisms	343	349	339	358
U	Intracellular trafficking, secretion, and vesicular transport	150	148	156	136
V	Defense mechanisms	48	53	48	58
Z	Cytoskeleton	1	1	1	0
	Total	4,550	4,675	4,635	4,851

Regarding the other *P. syringae* strains in the phylogeny with complete genome sequences, B728a was the closest to UMAF0158. Accordingly, this strain shares the highest number of CDSs predicted in UMAF0158 (see next section). These data, together with the fact that both B728a and UMAF0158 belong to *P. syringae* pv. syringae, prompted us to pay special attention to the genomic differences between these two strains.

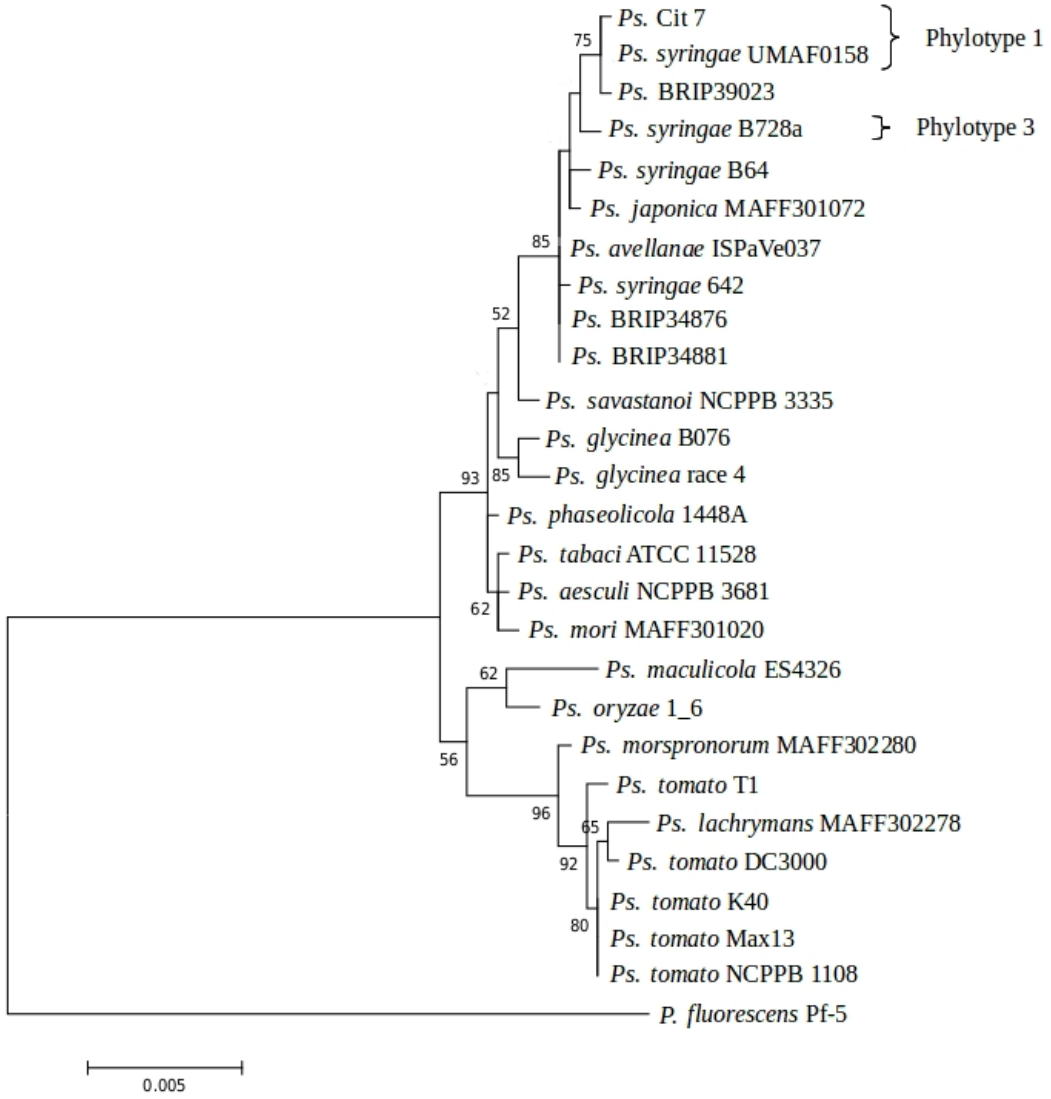


Figure 2. Phylogenetic analysis of *P. syringae* pv. syringae UMAF0158 and 25 selected strains of the *P. syringae* complex. Multilocus sequence analysis were performed using a concatenated dataset for *gapA*, *gltA*, *recA*, *rpoA* and *rpoB*. The evolutionary history was inferred using the Maximum Likelihood method based on the JTT matrix-based model. The percentage of trees in which the associated taxa clustered in the bootstrap test (1000 replicates) is shown next to the branches. Some strains are included in phylotypes 1 and 3 of pv. syringae according to Carrión *et al.* (2013) and Gutiérrez-Barranquero *et al.* (2013a).

Comparative genomics

The sequence of the UMAF0158 chromosome was compared to that of selected *P. syringae* strains (Figure 3). Of the 5017 CDSs predicted in UMAF0158, 4912 (98%) have orthologs (BLASTP E-value $\leq 1e^{-10}$) in other *P. syringae* and 3570 (71%) are present in all strains. All of the 105 genes found to be unique to UMAF0158 are heavily enriched in hypothetical proteins (103). The other two genes include a membrane protein (PSYRMG_17725) and a flavodoxin (PSYRMG_09680).

Among the selected strains, Cit 7, BRIP39023 and 642 have the closest number of CDSs compared to UMAF0158. These three strains, which belong to pv. syringae or are closed to it and whose complete genome sequences are not yet available, share 93.6, 93.4 and 91.2% of the CDSs predicted in UMAF0158, respectively (Figure 3). Regarding *P. syringae* with complete genome sequences, B728a shares 90.7% of the CDSs followed by 1448A and DC3000, which share 88.4 and 88.1%, respectively (Figure 3).

Figure 3 shows some sequence features associated with mechanisms of horizontal transfer, including regions with differential distributions of trinucleotides and GC-content, predicted prophages and putative horizontally transferred genes. In most cases, these features match with non-conserved regions of the UMAF0158 chromosome (white-colored in the six most outer rings).

Comparison of CDSs between UMAF0158 and B728a

In order to compare the genomes of UMAF0158 and B728a, we proceeded to identify regions enriched in coding genes in either strain that are not present in the other. The search was performed so that only regions spanning at least 4000 kb were retained (the whole set of differential protein coding genes are listed in Tables S3 and S4). These regions are summarized in Tables 3 and 4. Thirteen regions were identified in UMAF0158 with sizes ranging from 6051 to 20822 kb. Eight of such regions are highly enriched in hypothetical proteins (at least 80% of their CDSs). Two of the remaining five regions contain a combination of mobile genetic elements and hypothetical proteins. The remaining regions correspond to three potential operons: an additional T3SS, a cellulose production operon (PSYRMG_20805-20845), and the well-described mangotoxin biosynthetic operon *mbo* (PSYRMG_10110-10135) (Carrión *et al.*, 2012). This operon is present in only a limited number of strains belonging to genomospecies 1, and it has been acquired once during evolution by horizontal transfer (Carrión *et al.*, 2013). In addition, ten regions were identified in B728a with sizes ranging from 4968 to 43402 kb. Most of these regions contain mobile genetic elements. It is worth noting a region containing the streptomycin

resistance transposon Tn*5393* (Feil *et al.*, 2005). Two other regions are enriched in secretion components with one of them corresponding to a T4SS, which is addressed in the next section.

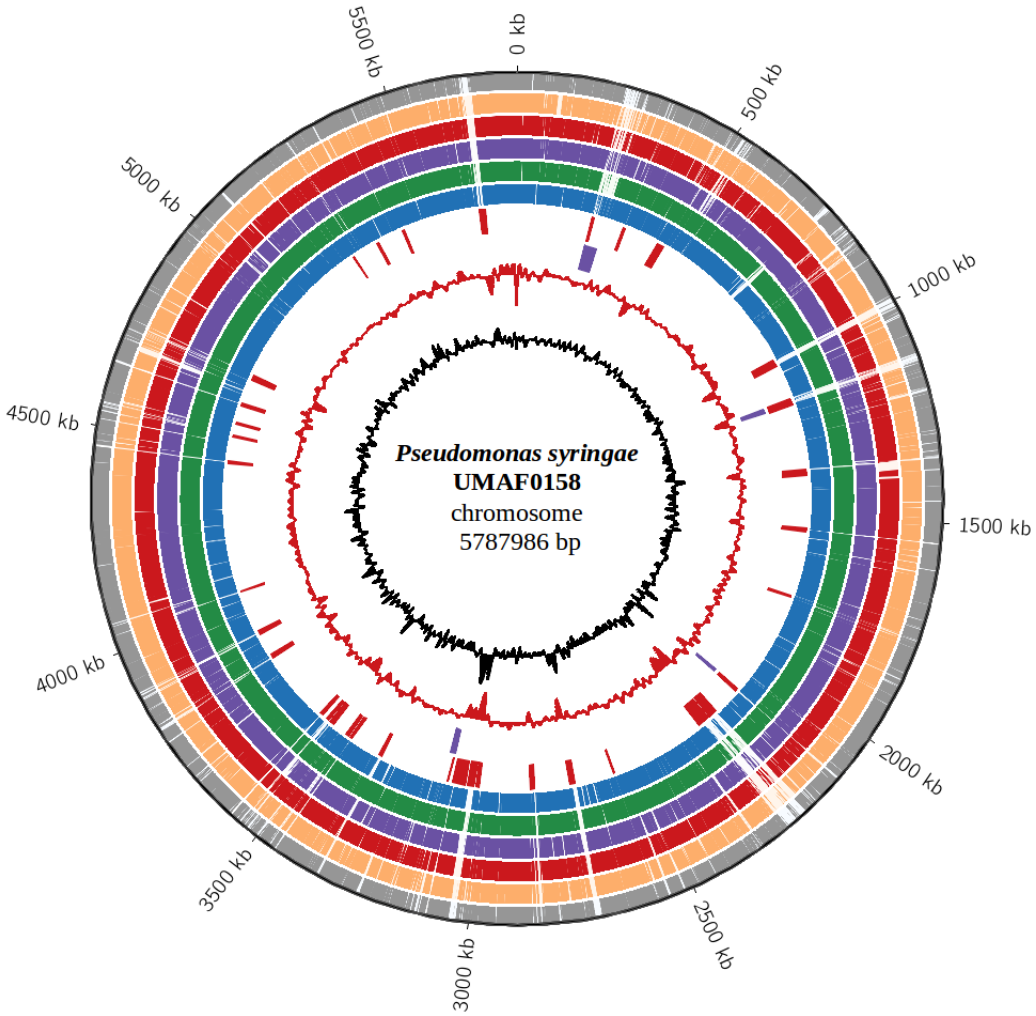


Figure 3. Conservation analysis of the *P. syringae* pv. syringae UMAF0158 chromosome. From the outside in, the outermost circle (black) shows the scale line. Circles 2 to 4 display CDSs homology (E-value $\leq 1e^{-10}$) among UMAF0158 and the three *P. syringae* with complete genome sequences: DC3000 (grey), 1448A (orange) and B728a (red). Circles 5 to 7 display CDSs homology (E-value $\leq 1e^{-10}$) among UMAF0158 and the draft genomes of the three phylogenetically closest *P. syringae* strains among the 25 selected in this study: 642 (purple), BRIP39023 (green) and Cit 7 (blue). Circles 8 and 9 display putative horizontally transferred regions (red) and prophages (purple), respectively; circle 10 shows G+C in relation to the mean G+C in 2 kb windows (red); circle 11 shows trinucleotide composition (black).

Table 3. Regions of *P. syringae* syringae pv. syringae UMAF0158 chromosome with low homology to *P. syringae* pv. syringae B728a.

Location (bp)	Length	No. of CDSs	No. of hypothetical	No. of CDSs not present in B728a	Relevant features
248304-257525	9221 kb	16	9 (56%)	15 (94%)	mobile genetic elements
258593-272300	13707 kb	21	11 (52%)	19 (90%)	mobile genetic elements
519318-537403	18085 kb	20	8 (40%)	15 (75%)	T3SS components
996632-1011913	15281 kb	17	17 (100%)	17 (100%)	hypothetical proteins
1133844-1143487	9643 kb	20	19 (95%)	18 (90%)	mobile genetic elements
1357822-1378644	20822 kb	5	4 (80%)	5 (100%)	hemolysin secretion/activation
2201422-2221045	19623 kb	22	18 (82%)	19 (86%)	mobile genetic elements
2233554-2244999	11445 kb	33	28 (85%)	31 (94%)	thiamin biosynthesis, peptidase
2324538-2330589	6051 kb	6	2 (33%)	6 (100%)	mangotoxin biosynthetic operon
2710838-2721328	10490 kb	10	10 (100%)	10 (100%)	hypothetical proteins
3032080-3050678	18598 kb	14	13 (93%)	14 (100%)	hypothetical proteins
4684145-4696189	12044 kb	7	0	6 (86%)	cellulose synthase
5668482-5685647	17165 kb	21	21 (100%)	21 (100%)	hypothetical proteins

Table 4. Regions of *P. syringae* syringae pv. syringae B728a chromosome with low homology to *P. syringae* pv. syringae UMAF0158.

Location (bp)	Length	No. of CDSs	No. of hypothetical	No. of CDSs not present in UMAF0158	Relevant features*
248304-257525	9221 kb	16	9 (56%)	15 (94%)	HP/MGE
258593-272300	13707 kb	21	11 (52%)	19 (90%)	HP/MGE T3 effector
519318-537403	18085 kb	20	8 (40%)	15 (75%)	T4SS components
996632-1011913	15281 kb	17	17 (100%)	17 (100%)	HP membrane transport
1133844-1143487	9643 kb	20	19 (95%)	18 (90%)	MGE secretion pilus proteins
1357822-1378644	20822 kb	5	4 (80%)	5 (100%)	HP/Virulence protein
2201422-2221045	19623 kb	22	18 (82%)	19 (86%)	streptomycin resistance
2233554-2244999	11445 kb	33	28 (85%)	31 (94%)	HP/phage-related proteins
2324538-2330589	6051 kb	6	2 (33%)	6 (100%)	HP/MGE/T3 effector
2710838-2721328	10490 kb	10	10 (100%)	10 (100%)	HP membrane transport plasmid-related proteins phage-related proteins T3 effector

* HP and MGE refer to hypotetical proteins and mobile genetic elements, respectively.

Secretion systems

The T6SS, which was first described in pathogenic bacteria such as *Vibrio cholerae*, *Pseudomonas aeruginosa* and *Burkholderia mallei* (Mougous *et al.*, 2006; Pukatzki *et al.*, 2006; Schell *et al.*, 2007), has been widely identified in Gram-negative bacteria, including *P. syringae* (Sarris *et al.*, 2010). This versatile secretion system has been proposed to promote symbiotic, commensal or mutualistic relationships between bacteria and eukaryotes and to intervene in cooperative or competitive interactions between different bacteria (Jani, Cotter, 2010). UMAF0158 contains two putative gene clusters that are associated with the T6SS, similarly to what it has been previously described for other strains of pvs. tomato, tabaci and oryzae, and in contrast with the genomes of B728a and 1448A, which carry only one T6SS (Sarris *et al.*, 2010). Both clusters in UMAF0158 are composed of 14 genes (PSYRMG_02245 to PSYRMG_02310 and PSYRMG_15560 to PSYRMG_15625) and range from coordinates 470416 to 487611 (Figure 4A) and 3502757 to 3522874 (Figure 4B), respectively. The first cluster is highly conserved in DC3000, while the second shows large similarity to that found in B728a (Barret *et al.*, 2011).

The T4SS can be involved in the translocation of proteins and genetic material, thus contributing to genome plasticity and virulence of bacteria harboring them (Voth *et al.*, 2012). B728a contains a conjugative G-type T4SS (Juhas *et al.*, 2008) that was not found in UMAF0158. However, UMAF0158 encodes a putative P-type T4SS in its plasmid, which is presumably involved in DNA conjugation (Cazorla *et al.*, 2008) (see Table S1).

The role of T3SSs in the virulence of pathogenic bacteria has been widely investigated Coburn *et al.*, 2007; Buttner, 2012. Two T3SS clusters were observed in the UMAF0158 chromosome. A complete T3SS (here named T3SS-1) similar to the *Hrp-1* family of T3SS (Egan *et al.*, 2014) found in pathogenic *P. syringae* strains (Mur *et al.*, 2008; Gazi *et al.*, 2012) was identified, ranging from coordinates 5000786 to 5047952 and consisting of 42 genes (PSYRMG_22290 to PSYRMG_22510) (Figure 4C). This was not unexpected as this strain has been shown to induce the hypersensitive response (HR) in tobacco plants (Cazorla *et al.*, 1998), a process dependent on a functional T3SS in *P. syringae* (Mur *et al.*, 2008). This cluster showed high similarity to that found in B728a, whose role in virulence has been widely reported (Vinatzer *et al.*, 2006; Lee *et al.*, 2012). The second putative T3SS (here called T3SS-2), which ranges from coordinates 523773 to 547384 in the UMAF0158 chromosome and consists of 24 genes (PSYRMG_02470 to PSYRMG_02585) (Figure 4D), shows high similarity to the rhizobial-like T3SS *Rhc* of Rhizobiales family of T3SS (Gazi *et al.*, 2012; Egan *et al.*, 2014). Among these 24 genes are included the complete set of core components to form the minimal apparatus (Abby, Rocha, 2012). This cluster is also

encoded in *P. syringae* pv. phaseolicola 1448A (Joardar *et al.*, 2005), *P. syringae* pv. tabaci 11528 (Studholme *et al.*, 2009), *P. syringae* pv. oryzae 1–6 (Reinhardt *et al.*, 2009), *P. syringae* pv. syringae 642 (Clarke *et al.*, 2010) and *P. savastanoi* NCPPB 3335 (Matas *et al.*, 2014) among other strains of the *P. syringae* complex but not in *P. syringae* pv. tomato DC3000 and *P. syringae* pv. syringae B728a. In agreement with data reported for other *P. syringae* genomes, regulatory sequences typical of the HrpL regulon (Fouts *et al.*, 2002) were not found preceding the genes in this second T3SS cluster. As deduced from analysis of specific mutants in the canonical T3SS in other *P. syringae* strains also encoding this second T3SS (Lindgren *et al.*, 1986; Pérez-Martínez *et al.*, 2010), it has been suggested that this atypical T3SS is not essential for pathogenicity, but a possible role in plant surface colonization or interaction with insects has been postulated (Clarke *et al.*, 2010; Silby *et al.*, 2011).

T3SS effectors

The predicted proteomes of UMAF0158 and the 25 selected *P. syringae* strains were screened for known *P. syringae* T3Es. Based on the presence or absence (E-value $\leq 1e^{-10}$) of the effector proteins currently described in http://Pseudomonas-syringae.org/, a matrix was constructed by evaluating whether each T3E was found as a complete ORF, an incomplete ORF, or not present (Figure 5; Table S5). File S1 contains detailed information on these analysis. By selecting T3Es with complete ORFs, a binary matrix was built based on Figure 5 data and used to generate a dendrogram shown in Figure 6. The results showed that UMAF0158 encodes 10 putative T3Es, including 6 which are also encoded in the B728a genome. Notably, and in agreement with the phylogeny showed in Figure 2, Cit 7 was the closest strain to UMAF0158 in terms of T3Es repertoire, both containing the same set with the exception of *hopI1*, only present in UMAF0158, and *hopAH1*, only present in Cit 7.

Virulence factors

The UMAF0158 genome was analyzed to detect known genes potentially implicated in virulence. We interpreted virulence to include factors such as siderophores, adhesins, phytotoxins, phytohormones, detoxifying compounds, plant cell wall degrading enzymes (PCWDEs) and exopolysaccharides (EPSs). In total, we identified 107 putative orthologs involved in the production of any of the above virulence factors (Table S6), which are addressed below.

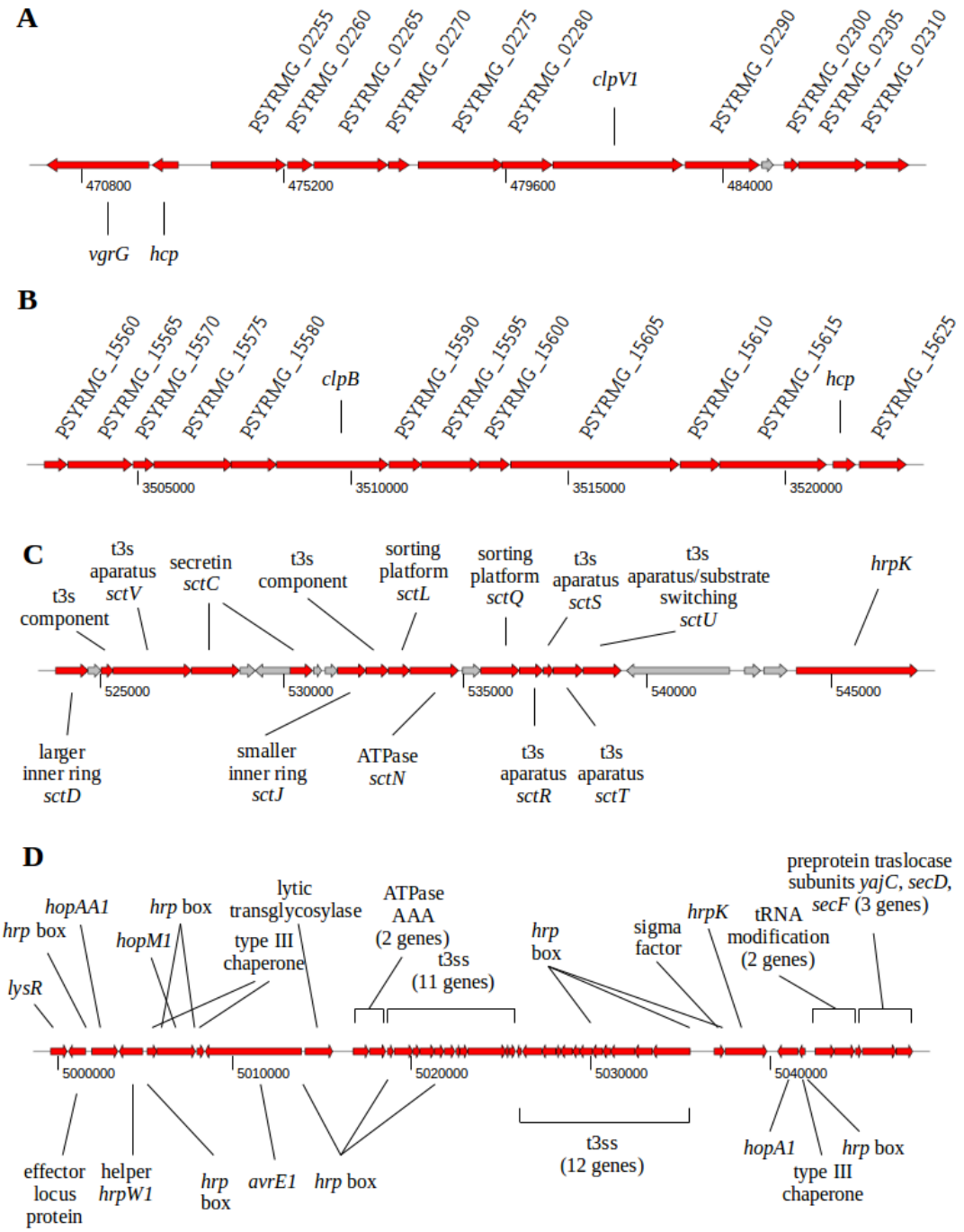


Figure 4. Genomic organization of putative *P. syringae* pv. syringae UMAF0158 secretion systems clusters potentially involved in effector translocation. **A**, T6SS-1. **B**, T6SS-2. **C**, T3SS-1 (*Hrc-1*). **D**, T3SS-2 (*Rhc*). Genes presumably involved in secretion are shown in red. Components of the T6SS with no consensual name are labelled with their corresponding NCBI-annotated locus tags. Numbers below the arrows refer to bp positions in the chromosome.

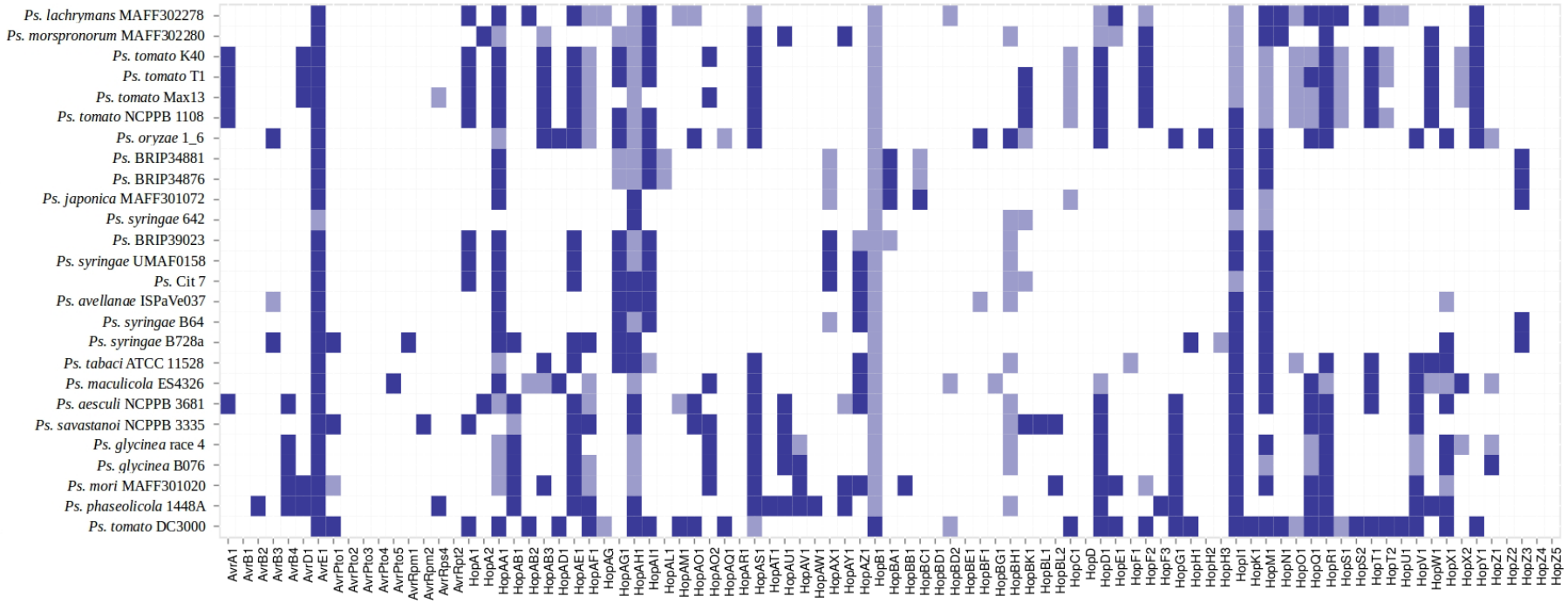


Figure 5. Presence of T3Es across 26 *P. syringae* strains. Known T3Es are listed across the bottom. Blue boxes indicate presence of complete ORFs within each genome; light blue boxes indicate that genes were found by similarity searches but they seem to be incomplete (see Material and Methods); white boxes indicate that no significant matches (E-value $\leq 1e^{-10}$) were found (Table S5; File S1). Searches were performed by BLASTp of the T3E proteins from the *P. syringae* Genome Resource (*http://pseudomonas-syringae.org/*) against the predicted CDSs of the selection of *P. syringae* strains.

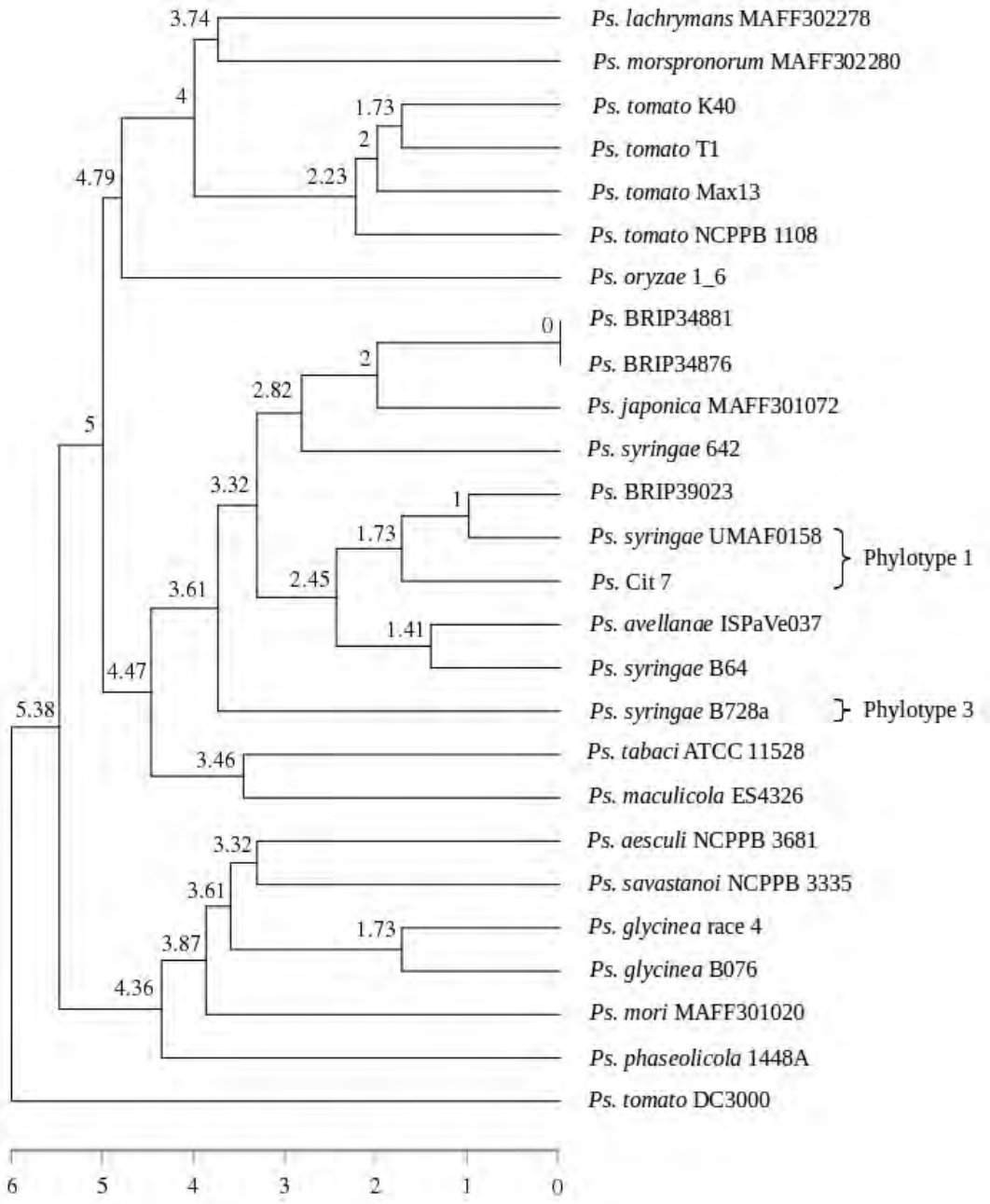


Figure 6. Dendrogram analysis of P. syringae pv. syringae UMAF0158 and 25 selected strains of the *P. syringae* complex based on the presence of T3Es. A binary matrix was created based on the presence/absence of T3E proteins. Then, the corresponding distance matrix was inferred, and the R package APE was used to generate the tree (see Material and Methods). The scale shows the joining distance between each point (strains) in the T3E matrix (Table S5; File S1).

Siderophores

Phytopathogenic bacteria synthesize and secrete a number of low molecular weight iron-chelating compounds called siderophores, which allow them to grow in iron-limited host environments (Guerinot, 1994; Krewulak, Vogel, 2008). UMAF0158 encodes genes orthologous to those required for the synthesis of pyoverdin and achromobactin, which have been widely described in other *Pseudomonas* (Feil *et al.*, 2005; Berti, Thomas, 2009; Zhang, Rainey, 2013). These two siderophores have also been proposed to increase the epiphytic fitness of *P. syringae* (Wensing *et al.*, 2010).

Adhesins and EPSs

Once a target host is reached, bacteria activate machinery for adhesion to plant tissues, which is thought to be necessary for the pathogenesis of many strains (Pizarro-Cerdá, Cossart, 2006; Mhedbi-Hajri *et al.*, 2011). Attachment factors previously identified in the genome of *P. syringae* strains include type IV pili, alginate, non-alginate capsular polysaccharide, EPSs, and filamentous hemagglutinin (Joardar *et al.*, 2005). Several genes predicting adhesins have been found in the UMAF0158 genome, including a filamentous hemagglutinin (PSYRMG_05745) and several fimbrial proteins, with the latter being clustered together with a number of pilus assembly proteins. Another gene predicting an adhesin is PSYRMG_14835, which resembles the gene encoding XadM, whose role in attachment and the formation of biofilms has been previously reported in *Xanthomonas oryzae* (Pradhan *et al.*, 2012). Interestingly, UMAF0158 contains two genes showing high similarity with the *attC* and *attG* genes in *Agrobacterium* (PSYRMG_02675 and PSYRMG_09205, respectively), whose mutation leads to virulence and lack of attachment on tomato, carrot, and *Bryophyllum daigremontiana* (Matthysse *et al.*, 2008).

Synthesis and secretion of EPSs is a common mechanism used by phytopathogenic bacteria, particularly in *P. syringae* strains (Sarkar *et al.*, 2006). They contribute to virulence by helping in attachment to host tissues and protecting bacterial cells from external stress (Denny, 1995). All of the genes required for alginate biosynthesis (Vázquez Peñaloza *et al.*, 1997) are present in the UMAF0158 gene cluster (PSYRMG_21640-21695). Notably, an ortholog of *gpsX*, a gene that encodes a glycosyltransferase that is involved in EPS production and is essential for the full virulence of *Xanthomonas citri* (Li, Wang, 2012), was also identified (PSYRMG_21000).

Phytotoxins

The production of small phytotoxic compounds by *P. syringae* pv. syringae is a well-known virulence mechanism that contributes to plant disease (Völksch, Weingart, 1998;

Arrebola *et al.*, 2003). UMAF0158 contains orthologs of genes participating in the synthesis of syringopeptin and syringomycin. These two toxins induce necrosis in plant tissues and have been shown to be the major virulence determinants of *P. syringae* pv. syringae (Bender *et al.*, 1999; Scholz-Schroeder *et al.*, 2001). The two clusters encoding these toxins form a larger cluster (PSYRMG_03860 – 03910), which is consistent with previously reported data (Feil *et al.*, 2005; Scholz-Schroeder *et al.*, 2001). The production of syringomycin by UMAF0158 has been experimentally validated in previous studies by our group (Arrebola *et al.*, 2003).

Syringolins are another family of phytotoxins synthesized by a number of *P. syringae* pv. syringae (Krahn *et al.*, 2011). UMAF0158 has a gene cluster resembling that of the production of syringolin A (PSYRMG_24250-24275), a toxin that has been shown to counteract stomatal innate immunity in beans and *Arabidopsis* (Schellenberg *et al.*, 2010).

Phaseolotoxin and coronatine are two chlorosis-inducing toxins that also represent major virulence factors for some *P. syringae* isolates (Aguilera *et al.*, 2007; Zheng *et al.*, 2012). UMAF0158 lacks orthologs for most of the genes involved in the production of coronatine; however, analysis of the 23 genes required for the synthesis of phaseolotoxin (Aguilera *et al.*, 2007) showed that orthologs of 17 of these genes are included in its genome. Given that no inhibition halos were observed in the bioassay for toxin detection when ornithine was added (Arrebola *et al.*, 2003) and no chlorosis was detected among the symptoms of UMAF0158 infection, it is likely that the lack of the other six genes prevent the synthesis of phaseolotoxin by this strain.

The production of mangotoxin by UMAF0158 and its contribution to the virulence of this strain has been widely described (Arrebola *et al.*, 2007; Carrión *et al.*, 2012). Two operons are involved in the synthesis of this toxin, namely, *mgo* (PSYRMG_15820-15835) and *mbo* (PSYRMG_10110-10135) (Carrión *et al.*, 2012; Arrebola *et al.*, 2012). The latter has been shown to be absent in B728a (Carrión *et al.*, 2013). Unlike other toxins, mangotoxin is thought to be associated with host specificity, as it has been found to be synthesized by strains of phylotypes 1 and 2 of pv. syringae, which were mainly isolated from mango trees and other woody crops (Gutiérrez-Barranquero *et al.*, 2013a).

It is worth noting that, even though the metabolic cost of toxin production generally prohibits bacteria from producing more than one, the synthesis of at least two of them has been experimentally validated in UMAF0158 (i.e., syringomicin and mangotoxin; Arrebola *et al.*, 2003). Whether this strain is capable of producing the rest of the phytotoxins mentioned above is a question that requires further investigation.

Phytohormones

Bacterial-produced phytohormones are typically transported to the plant cell to regulate plant biological processes, providing a beneficial context for the pathogen (Costacurta, Vanderleyden, 1995). Such is the case of auxin, which is predominantly represented by indole-3-acetic acid (IAA), a key plant growth regulator that is also involved in plant-bacteria interactions. The downregulation of this hormone in plants has been shown to restrict *P. syringae* growth in *Arabidopsis* (Navarro *et al.*, 2006), suggesting that bacteria may have evolved the production of auxin to overcome this plant response. Accordingly, auxin production has been demonstrated to promote susceptibility to *P. syringae* (Mutka *et al.*, 2013). The genome of UMAF0158 contains orthologs of two of the genes involved in the biosynthesis of IAA, namely, *iaaH* and *iaaM*. Further analyses will elucidate whether this strain is able to produce IAA.

Detoxifying compounds

As a response to bacterial infections, plants synthesize reactive oxygen species (ROS), such as hydrogen peroxide (H2O2), superoxide (O2-) and hydroxyl radical (OH). These molecules have a toxic effect on invading bacteria (Cabiscol *et al.*, 2000), which have evolved mechanisms to counterattack by detoxification. DC3000 typically makes use of the catalases KatB and KatE together with the catalase-peroxidase KatG to detoxify plant-produced H2O2. Interestingly, orthologs of these three gene products are present in the UMAF1058 genome. Moreover, an ortholog of the gene that encodes Dps, a ferritin-like protein that has been shown to protect plant-associated bacteria against oxidative stress (Colburn-Clifford *et al.*, 2010), has been also identified (PSYRMG_23380). Other genes possibly implicated in detoxification found in UMAF0158 correspond to a cluster presumably encoding a cbb(3)-type cytochrome C oxidase (PSYRMG_07490-07505) and a proline iminopeptidase (PSYRMG_18265). The latter has been reported to be required for pathogenicity of *Xanthomonas campestris* (Zhang *et al.*, 2007) and to have dealanylating activity toward ascomycin, an antibiotic produced by *Streptomyces* that inhibits protein synthesis (Sudo *et al.*, 1996). Genes involved in copper resistance have also been identified, including a cluster containing *copA* and *copB* (PSYRMG_23630 and PSYRMG_23625, respectively), and a locus with high similarity to the *cueAR* system in *Pseudomonas putida* (Adaikkalam, Swarup, 2005). This system consists of a copper-transporting ATPase transmembrane protein and its transcriptional regulator (PSYRMG_19635 and PSYRMG_19630, respectively). Additionally, the *copABCD* operon described in other *P. syringae* (Cazorla *et al.*, 2002) is absent in the UMAF0158 chromosome, and also any

copper resistant genes are present in the UMAF0158 plasmid, in agreement with the copper sensitivity of this strain (Gutiérrez-Barranquero *et al.*, 2013b).

PCWDEs

Some phytopathogenic bacteria need to overcome the plant cell wall in the process of accessing the host cytoplasm. Therefore, many plant pathogens harbor a collection of genes encoding PCWDEs, which are considered important virulence determinants (Barras *et al.*, 1994). The UMAF1058 genome contains genes predicting several PCWDEs, such as a cellulase (PSYRMG_06950), a lipoyl synthase (PSYRMG_12655), a xylanase (PSYRMG_13355) and a pectin lyase (PSYRMG_10750).

Conclusions

Summarizing, bioinformatics analysis of the complete genome of *P. syringae* pv. syringae UMAF0158, a pathogen of mango trees, revealed a high degree of conservation with other pseudomonads belonging to the *P. syringae* complex, including the model strain *P. syringae* pv. syringae B728a. However, the resulted phylogeny clustered UMAF0158 with *P. syringae* Cit 7 and more separately from B728a. Indeed, our data revealed a number of genetic factors that could be involved in the differential pathogenic and epiphytic lifestyle of UMAF0158, in comparison with the model strain B728a. The mangotoxin biosynthetic operon *mbo* is included among these factors, whose role in the pathogenicity of UMAF0158 has been previously reported (Carrión *et al.*, 2012, 2013). Moreover, UMAF0158 harbors an operon presumably involved in cellulose production and two clusters predicting additional secretion systems of types III and IV, as well as displays a particular T3Es repertoire. Additionally, the conjugative plasmid pPSS158 contains *rulAB* genes, involved in UV resistance and epiphytic fitness (Cazorla *et al.*, 2008). This work provides the basis for further analyses on the specific mechanisms that enable this strain to infect mango trees as well as for functional analyses of the factors governing host specificity in pv. syringae strains from different phylotypes.

Material and Methods

Bacterial growth and DNA methods

The bacterial strain UMAF0158 (CECT 7752) belonging to *Pseudomonas syringae* pv. syringae was routinely grown in KB medium with 48 h of incubation at 28°C. UMAF0158 was inoculated on 100 ml of LB medium and grown for 15 h at 28°C with shaking (150 r.p.m.). After this period, the OD600nm of the culture was 1.8. Serial dilutions of this culture and plating on LB plates yielded 1.3×109 CFU/ml of pure bacterial culture. The rest of the culture was divided into 54 aliquots of 1.5 ml, and DNA was extracted from all of the cultures using the Jet-Flex genomic DNA purification kit (Genomed GmbH, Germany). DNA samples were collected together and further purified by extraction with 1:1 phenol:chloroform and precipitated with 4 M NaCl and 13% PEG. DNA was suspended in 500 μl MilliQ H2O, and NanoDrop measurements indicated 3.6 μg/μl (in total 1800 μg of DNA) with an A260/A280 of 1.85. The extracted DNA was visualized in 1% agarose after digestion with the restriction enzymes EcoRI and PstI.

Whole Genome Sequencing

The finished UMAF1058 genome was generated at the Beijing Genomics Institute (BGI-HK) using an Illumina HiSeq 2500 system. Briefly, the isolated DNA was used to generate three libraries of 500 bp, 2000 bp and 6000 bp, producing 1336, 1312 and 1352 Mb of raw data, respectively. These were passed through a filtering pipeline that removed known sequencing and library preparation artifacts. After data treatment, the SOAPdenovo 1.05 software package (Li *et al.*, 2008, 2010) was used for sequence assembly and quality assessment. Assembly results were then combined and mapped to the genome of *P. syringae* B728a, yielding to two scaffolds corresponding to one chromosome and one plasmid. Finally, a PCR gap closure and three circle PCR verification were performed to obtain the final complete sequences of the chromosome and the pPSS158 plasmid.

Genomic data and annotation

The assembled genome of UMAF0158 was submitted to the NCBI Prokaryotic Genome Annotation Pipeline for automatic annotation and manually reviewed. Gene locations and protein products were generated from above annotation (ASN.1 file) using the script "asn2all" belonging to the NCBI ToolKit (*http://www.ncbi.nlm.nih.gov/toolkit*). Genome sequences (DNA, proteins and predicted genes locations) of DC3000, B728a and 1488A were downloaded from the NCBI respository of completely sequenced bacterial

genomes (*ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria*), while the corresponding sequences of *P. savastanoi* pv. savastanoi strain NCPPB3335 were obtained from our own sequencing project. The rest of the *P. syringae* genomes were downloaded from the NCBI draft bacterial genome repository (*ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria_DRAFT/*). In this case, proteins and gene locations were generated from DNA sequences (fna files) using Glimmer v3.02 (Delcher *et al.*, 2007). All genomes were downloaded on July 15, 2014. Accession numbers and references for all genome sequences used in this work are summarized in Table S2.

The UMAF0158 annotation of COGs was performed by aligning the set of predicted protein sequences against the COG PSSM of the CDD (Marchler-Bauer *et al.*, 2014) using rps-BLAST. Hits with an E-value ≤ 0.001 were first retained. Then, only the best hit was selected for each protein. The same procedure was used to assign COG categories to the repertoire of predicted proteins of the B728a, DC3000 and 1448A strains. Predictions of horizontally transferred regions and prophages were computed using Alien Hunter 1.7 (Vernikos, Parkhill, 2006) and Prophage Finder (Bose, Barber, 2006), respectively. T346Hunter (Martínez-García *et al.*, 2015b) was used to identify secretion systems clusters. Virulence factors were predicted using a customized annotation tool developed by our group (not published). Further details on such a tool are provided in Chapter IV.

The finished genomic sequences of UMAF0158 were deposited in GenBank under accession numbers CP005970 (chromosome) and CP005971 (plasmid pPSS158)

Trinucleotide composition

The distribution of all 64 trinucleotides was determined for the whole chromosome and 2 kb sub-windows. Then, the χ^2 statistic of the difference between the trinucleotide composition of each window and that of the whole chromosome was computed. Large values for χ^2 in a given window denote different trinucleotide compositions from the rest of the chromosome. Probability values were computed assuming uniform distribution of the DNA composition along the genome. Because this assumption may have been incorrect, large χ^2 values were interpreted as indicators of unusual regions on the chromosome that require further investigation.

Phylogenetic tree

Phylogeny was determined by multilocus sequence analysis using a concatenated dataset for the housekeeping genes *gapA*, *gltA*, *recA*, *rpoA* and *rpoB*. Then, the Maximum Likelihood method based on the JTT (Jones-Taylor-Thornton) matrix-based model (Jones *et al.*, 1992) was applied. The percentage of trees in which associated taxa clustered in the

bootstrap test (1000 replicates) is shown next to the branches in Figure 2 (Felsenstein, 1985). Multiple alignments and evolutionary analyses were conducted using MEGA5 software (Tamura *et al.*, 2011).

Comparative genomics

Each of the predicted proteins in UMAF0158 was compared to those of the other *P. syringae* strains using BLASTP (E-value $\leq 1e^{-10}$). The same procedure was used to compare predicted proteins in B728a with those in UMAF0158.

Distribution of T3Es

We performed BLASTP searches of the T3Es in *http://Pseudomonas-syringae.org/* against the 26 *P. syringae* predicted proteomes. First, only hits with an E-value $\leq 1e^{-10}$ were retained. If no hits were found for a given T3E, it was considered absent. For a given strain, when a gene product was found to match with several T3Es, the one with the best E-value was selected. If there were more than one T3E with the best E-value, the alignment with the greatest number of identities was retained. Then, lengths of the query T3Es were compared to those of the alignments. We labelled a potential T3E as incomplete when the alignment was $\geq$ 25% smaller than the length of the original T3E. Otherwise, the T3E was labelled as complete. Based on the presence of complete T3Es, a binary matrix was created and used to generate a dendrogram by means of the R package APE (Paradis *et al.*, 2004). Further information on this analysis can be found in Table S5 and File S1.

Circular genome visualization

Circular layouts were generated using Circos (Krzywinski *et al.*, 2009).

Supplementary Material

Table S1. Predicted ORF in PssUMAF0158 plasmid (pPSS158). ORFs were annotated by the NCBI Prokaryotic Genome Annotation Pipeline and manually curated. Available: http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter1/Table_S1.docx

Table S2. Accession numbers and references for the genome sequences of 26 *Pseudomonas syringae* strains used in this work. Available: http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter1/Table_S2.docx

Table S3. Regions of UMAF0158 chromosome with low homology to B728a. Available: http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter1/Table_S3.docx

Table S4. Regions of B728a chromosome with low homology to UMAF0158. Available: http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter1/Table_S4.docx

Table S5. T3Es repertoires found in 26 *Pseudomonas syringae* strains. Columns provide information on BLASTp alignments between T3Es from *http://pseudomonas-syringae.org/* and strains gene products, such as E-value, fraction and number of identical positions, alignment length, coordinates (start-end) for the query effector and subject gene product in the alignment and number of gaps. Other relevant features are provided, such as lengths of both the effector and the gene product, the rate between effector length and alignment length and whether the searched effectors are considered complete. Available: http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter1/Table_S5.xlsx

Table S6. Relevant virulence factors found in UMAF0158 chromosome. Available: http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter1/Table_S6.xlsx

File S1. BLASTp alignments (stockholm format) generated by the T3Es searches. Available: http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter1/File_S1.zip

CHAPTER II

Complete genome sequence of *Pseudomonas fluorescens* strain PICF7, an indigenous root endophyte from olive (*Olea europaea* L.) and effective biocontrol agent against *Verticillium dahliae*

Pedro Manuel Martínez-García, David Ruano-Rosa, Elisabetta Schilirò, Pilar Prieto, Cayo Ramos, Pablo Rodríguez-Palenzuela and Jesús Mercado-Blanco. Complete genome sequence of *Pseudomonas fluorescens* strain PICF7, an indigenous root endophyte from olive (*Olea europaea* L.) and effective biocontrol agent against *Verticillium dahliae*. *Standards in Genomic Sciences*. 2015. doi: 10.1186/1944-3277-10-10.

Summary

Pseudomonas fluorescens strain PICF7 is a native endophyte of olive roots. Previous studies have shown this motile, Gram-negative, non-sporulating bacterium is an effective biocontrol agent against the soil-borne fungus *Verticillium dahliae*, the causal agent of one of the most devastating diseases for olive (*Olea europaea* L.) cultivation. Here, we announce and describe the complete genome sequence of *Pseudomonas fluorescens* strain PICF7 consisting of a circular chromosome of 6,136,735 bp that encodes 5,567 protein-coding genes and 88 RNA-only encoding genes. Genome analysis revealed genes predicting factors such as secretion systems, siderophores, detoxifying compounds or volatile components. Further analysis of the genome sequence of PICF7 will help in gaining insights into biocontrol and endophytism.

Introduction

Pseudomonas fluorescens PICF7 is a native colonizer of olive (*Olea europaea* L.) roots and an *in vitro* antagonist of the soil-borne fungal phytopathogen *Verticillium dahliae* Kleb. (Mercado-Blanco *et al.*, 2004), the causal agent of Verticillium wilts in a large number of plant species (Pegg, Brady, 2003). This strain has been demonstrated to be an effective BCA against Verticillium wilt of olive (Mercado-Blanco *et al.*, 2004; Prieto *et al.*, 2009), one of the most important biotic constraints for olive cultivation (López-Escudero, Mercado-Blanco, 2011). Moreover, strain PICF7 is able to display an endophytic lifestyle within olive root tissues under different experimental conditions (Prieto, Mercado-Blanco, 2008; Prieto *et al.*, 2009, 2011) and induces a broad range of defence responses at both local (roots) and systemic (above-ground organs) level, as well as to activate diverse transcription factors known to be involved in systemic defence responses (Schilirò *et al.*, 2012; Gómez-Lama *et al.*, 2014). Accordingly, a recent study has shown the ability of PICF7 to influence the establishment of the pathogen *Pseudomonas savastanoi* pv. savastanoi in olive stems and to affect the normal development of olive knots (Maldonado-González *et al.*, 2013), its associated disease (Ramos *et al.*, 2012). In this report, we summarize the complete genome sequence and annotation of PICF7. We also describe its genomic properties, highlighting genes encoding plant-associated factors, colonization abilities and well-known bacterial biocontrol traits. The genome sequencing of PICF7 and its comparison with related published genomes will provide a framework for further functional studies of its rhizosphere competence, biocontrol effectiveness and endophytic lifestyle.

Results and Discussion

Classification and features

P. fluorescens PICF7 is a motile, Gram-negative, non-sporulating rod in the order *Pseudomonadales* of the class *Gammaproteobacteria*. Rod-shaped cells are approximately 0.5 μm in width and 2.0-2.5 μm in length (Figure 1 left and Centre). The strain is moderately fast-growing, forming 2 mm colonies within 2-3 days at 28°C. Colonies formed on King's B (KB) (King *et al.*, 1954) agar plates are yellow-green opaque, domed and moderately mucoid with smooth margins (Figure 1 Right).

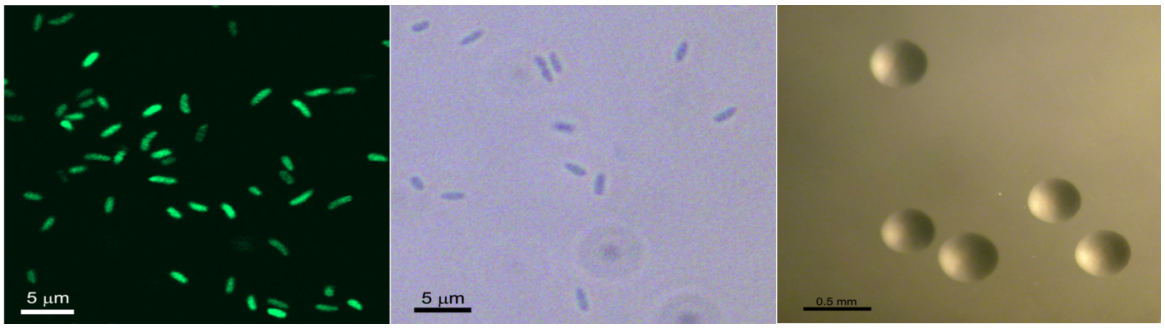


Figure 1. Image of *P. fluorescens* PICF7 cells using confocal laser scanning (Left) and phase-contrast (Centre) microscopy (CLSM and PCM, respectively) and the appearance of colony morphology after 24 h growing on KB agar medium at 28°C (Right). CLSM image was obtained using a PICF7 derivative carrying a plasmid with an enhanced green fluorescent protein (EGFP) (Prieto, Mercado-Blanco, 2008).

PICF7 was isolated from the roots of healthy nursery-produced olive plants cv. Picual in Córdoba province (Southern Spain) (Mercado-Blanco *et al.*, 2004). It grows in complex media such as LB (Bertani, 1951) or KB, as well as in minimal media such as Standard Succinate Medium (SSM; pH 7.0) (Meyer, Abdallah, 1978). Even though the optimal growth temperature is 28°C, PICF7 can also slightly replicate at 5°C in liquid LB and KB. However, growth at 37°C was not observed in these culturing media after 24 h. The bacterium is an efficient colonizer of the olive rhizosphere (Mercado-Blanco *et al.*, 2004) and displays an endophytic lifestyle (Prieto, Mercado-Blanco, 2008; Prieto *et al.*, 2009, 2011). It does not cause any deleterious effect on its original host (olive) (Mercado-Blanco *et al.*, 2004; Prieto, Mercado-Blanco, 2008; Maldonado-González *et al.*, 2013). Strain PICF7 has natural resistance to kanamycin (50 mg/L) and nalidixic acid (25 mg/L), and it is possible to develop spontaneous rifampicin-resistant mutants (Mercado-Blanco *et al.*, 2004).

Minimun Information about the Genome Sequence (MIGS) of *P. fluorescens* PICF7 is summarized in Table 1, and its phylogenetic position is shown in Figure 2.

Table 1. Classification and general features of *P. fluorescens* PICF7 according to the MIGS recommendations (Field *et al.*, 2008).

MIGS ID	Property	Term
	Classification	Domain *Bacteria*
		Phylum *Proteobacteria*
		Class *Gammaproteobacteria*
		Order *Pseudomonadales*
		Family *Pseudomonadaceae*
		Genus *Pseudomonas*
		Species *Pseudomonas fluorescens*
		Strain PICF7
	Gram stain	Negative
	Cell shape	Rod-shaped
	Motility	Motile
	Sporulation	None
	Temperature range	Mesophilic
	Optimum temperature	28°C
MIGS-22	Oxygen requirement	Aerobic
	Carbon source	Heterotrophic
	Energy metabolism	Chemoorganotrophic
MIGS-6	Habitat	Soil, olive root-associated
MIGS-6.3	Salinity	NaCl 1-4%
MIGS-10	Extrachromosomal elements	None
MIGS-11	Estimated size	6.14 Mb
MIGS-15	Biotic relationship	Rhizospheric, root endophytic
MIGS-14	Pathogenicity	Non-pathogenic
	Host	*Olea europaea*
	Host taxa ID	4146
	Isolation source	Root
	Biosafety level	1
MIGS-4	Geographic location	Córdoba, Spain
MIGS-5	Sample collection time	1998
MIGS-4.1	Latitude	37°4128N
MIGS-4.2	Longitude	4°2858W
MIGS-4.4	Altitude	230 m.a.s.l

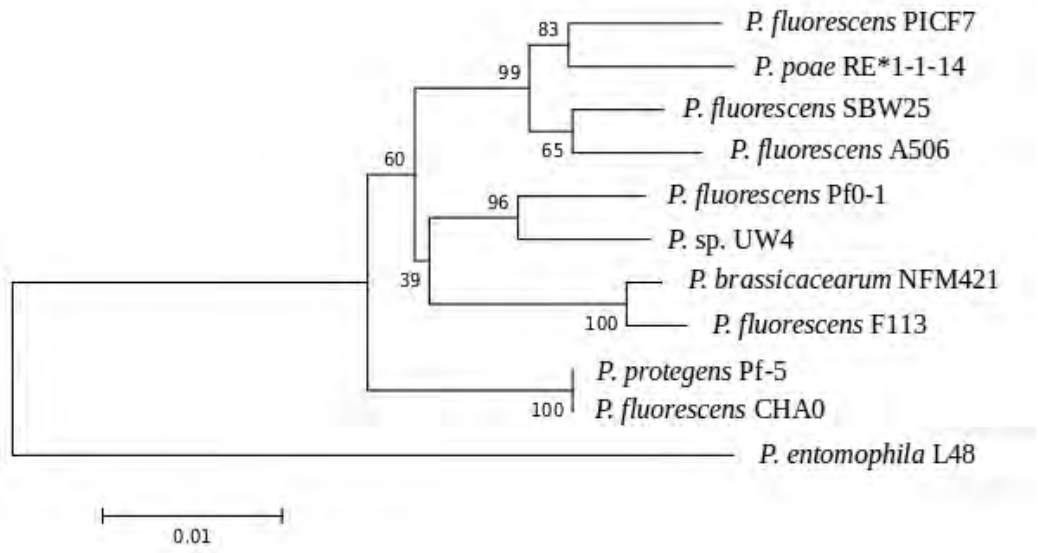


Figure 2. Phylogenetic tree highlighting the position of P. fluorescens strain PICF7 relative to its closest Pseudomonas strains for which complete genomes are available. *P. entomophila* strain L48 was used as an outgroup. For the construction of the tree, five protein-coding house-keeping genes were first aligned, namely: *argF*, *atpA*, *nusA*, *pyrH* and *rpoH*. Then, the Maximum Likelihood method based on the JTT (Jones-Taylor-Thornton) matrix-based model was used (Jones *et al.*, 1992). The percentage of trees in which the associated taxa clustered in the bootstrap test (1000 replicates) is shown next to the branches (Felsenstein, 1985).

Genome project history

P. fluorescens PICF7 was selected for sequencing due to its ability to exert biocontrol against Verticillium wilt of olive (Mercado-Blanco *et al.*, 2004; Prieto *et al.*, 2009) and to develop an endophytic lifestyle within olive root tissues (Prieto and Mercado-Blanco, 2008; Prieto et al., 2011). The genome project is deposited in the Genomes OnLine Database (Liolios *et al.*, 2006) and the NCBI's BioProject database. The finished genome sequence is in GenBank. A summary of the project information is shown in Table 2.

Genome properties

The genome of PICF7 is composed of one circular chromosome of 6,136,735 bp with an average GC content of 60.4% (Table 3, Figure 3), which is similar to that of other *P. fluorescens* strains. Among the 5,655 predicted genes, 5,567 were identified as protein coding genes. Of the last, 4,573 (82.1%) were assigned a putative function, while the other 994 (17.9%) were designated as hypothetical proteins. The classification of CDSs into functional categories according to the COG database (Tatusov *et al.*, 1997) is summarized in Table 4.

Table 2. Project information.

MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished
MIGS-28	Libraries used	Three libraries of 500 bp, 2,000bp and 6,000bp, respectively
MIGS-29	Sequencing platforms	Solexa
MIGS-30	Assemblers	SOAPdenovo 1.05
MIGS-32	Gene calling method	NCBI Prokaryotic Genome Annotation Pipeline
	Locus Tag	PFLUOLIPICF7
	Genbank ID	CP005975
	Date of Release	May 31, 2017
	GOLD ID	Gi0079402
	BIOPROJECT	PRJNA203247
	NCBI taxon ID	1334632
	Project relevance	Plant-bacteria interaction Model for endophytic lifestyle Agricultural, Environmental

Table 3. Genome statistics.

Attribute	Genome (total)	
	Value	% of total
Genome size (bp)	6,136,735	100
DNA coding region (bp)	5,439,499	88.6
DNA G+C content (bp)	3,706,588	60.4
DNA scaffolds	1	-
Total genes	5,655	100
Protein-coding genes	5,567	98.4
RNA genes	68	1.6
Pseudo genes	30	0.8
Protein-coding genes with function prediction	4,573	82.1
Protein-coding genes assigned to COGs	4,581	82.3
Proteins with signal peptides	644	11.6
Proteins with transmembrane helices	1,319	23.7

Table 4. Number of genes associated with general COG functional categories.

	Functional category	Value	% of total
A	RNA processing and modification	1	0.02
B	Chromatin structure	5	0.09
C	Energy production and conversion	280	5.03
D	Cell cycle control, cell division, chromosome partitioning	41	0.74
E	Amino acid transport and metabolism	554	9.95
F	Nucleotide transport and metabolism	96	1.72
G	Carbohydrate transport and metabolism	307	5.51
H	Coenzyme transport and metabolism	196	3.52
I	Lipid transport and metabolism	219	3.93
J	Translation, ribosomal structure and biogenesis	200	3.59
K	Transcription	501	9
L	Replication, recombination and repair	156	2.8
M	Cell wall/membrane/envelope biogenesis	267	4.8
N	Cell motility	162	2.9
O	Posttranslational modification, protein turnover, chaperones	177	3.18
P	Inorganic ion transport and metabolism	301	5.41
Q	Secondary metabolites biosynthesis, transport, and catabolism	151	2.71
R	General function prediction only	592	10.63
S	Function unknown	446	8.01
T	Signal transduction mechanisms	366	6.57
U	Intracellular trafficking, secretion, and vesicular transport	153	2.75
V	Defense mechanisms	67	1.2
Z	Cytoskeleton	-	-
W	Extracellular estructures		
-	Not in COGs	986	17.7

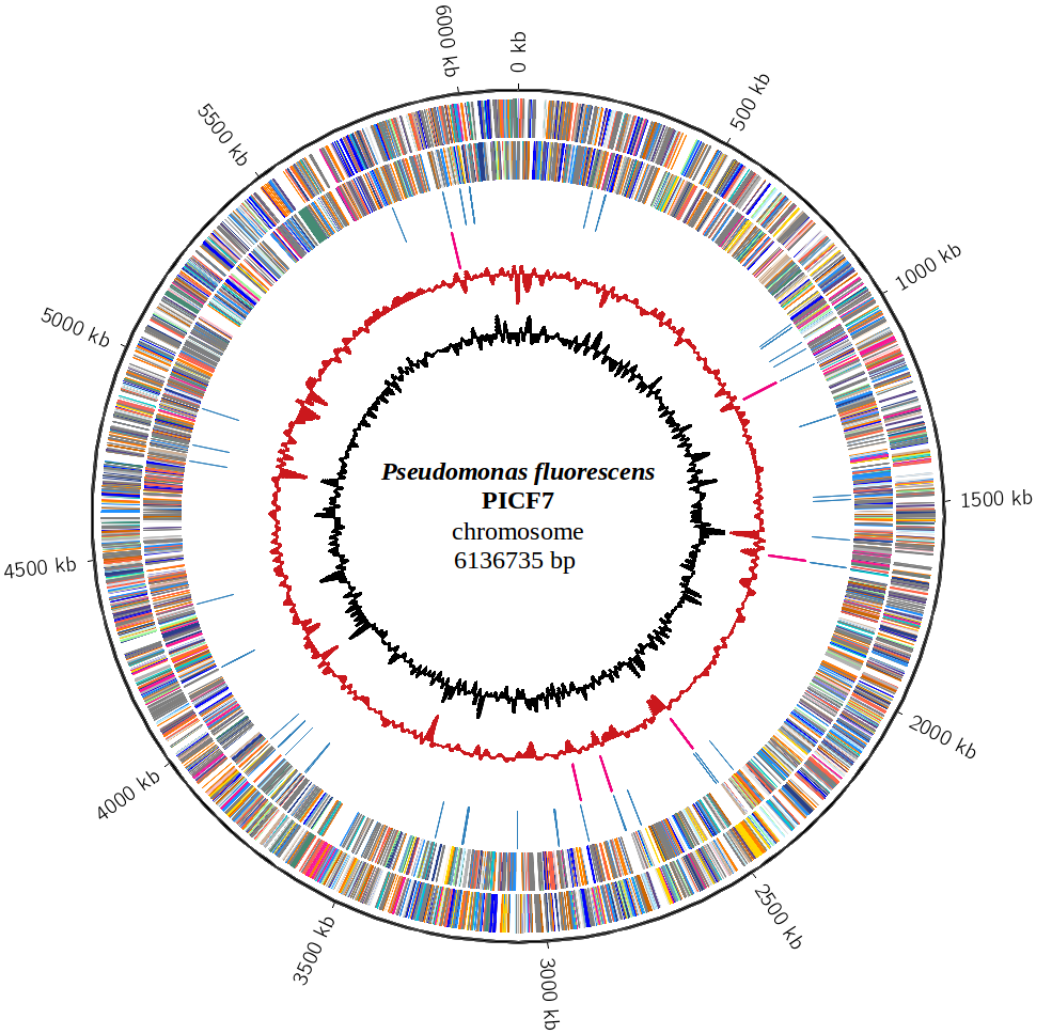


Figure 3. Graphical map of the chromosome. From outside to the centre: genes on forward strand (coloured by COG categories), genes on reverse strand (coloured by COG categories), RNA genes: tRNAs - blue, rRNAs – pink, G+C in relation to the mean G+C in 2kb windows and trinucleotide distribution in 2kb windows. The latter was defined as the χ^2 statistic on the difference between the trinucleotide composition of 2kb windows and that of the whole chromosome.

Insights from the genome sequence

The genome contains a complete canonical type III secretion system and two known effector proteins, namely, AvrE1 and HopB1. In addition, two complete type VI secretion system (T6SS) clusters were identified. T6SS has been described to promote antibacterial activity against a wide range of competitor bacteria (Hood *et al.*, 2010). PICF7 genome also encodes gene clusters for the synthesis of the siderophores pyochelin and pyoverdine and the hemophore HasAp. A repertoire of cell adhesion proteins has been also identified,

including two filamentous hemagglutinin proteins and several fimbrial proteins clustered together with a number of pilus assembly proteins. Notably, two genes have been found to show high similarity with *attC* and *attG* genes from *Agrobacterium*, whose mutation leads to lack of attachment on tomato, carrot, and *Bryophyllum daigremontiana* (Matthysse *et al.*, 2008).

It is worth mentioning the presence of genome components presumably involved in the synthesis of detoxifying compounds. Such is the case of two clusters containing genes for copper resistance and for production of a cbb(3)-type cytochrome C oxidase, respectively. An ortholog of the gene that codes for Dps, a ferritin-like protein reported to protect plant-associated bacteria against oxidative stress (Colburn-Clifford *et al.*, 2010), has also been found. Additional identified traits involved in detoxification are orthologs of catalase KatB and hydroperoxidase KatG, which detoxify plant-produced H_2O_2 (Guo *et al.*, 2012), and a gene coding for a proline iminopeptidase, which has been shown to have dealanylating activity toward the antibiotic ascamycin (Sudo *et al.*, 1996). A gene predicting a salycilic hydroxylase has been also identified in PICF7 genome. This gene could be involved in the degradation of the plant defence hormone salicylic acid, thus disrupting the systemic response against colonizing bacteria. In addition, all genes required for biosynthesis of the exopolysaccharide alginate (Vázquez Peñaloza *et al.*, 1997) are present in a gene cluster.

Genes predicting volatile components are present in PICF7 genome as well. Volatile components have been shown to act as antibiotics and to induce plant growth (Ryu *et al.*, 2003; Ren *et al.*, 2010). An example is hydrogen cyanide (HCN), an inorganic compound with antagonistic effects against soil microbes (Ahmad *et al.*, 2008). Orthologs of genes required for the biosynthesis of other volatile components such as 2,3-butanediol and acetoin were also found. Further genome analysis revealed other factors presumably involved in the endophytic fitness of PICF7. Such is the case of enzymes like a cellulase and a phytase, as well as the gene coding for aminocyclopropane-1-carboxylate deaminase suggested to be key in the modulation of ethylene levels in plants by bacteria (Hardoim *et al.*, 2008).

Conclusions

In this report we describe the complete genome sequence of *Pseudomonas fluorescens* strain PICF7, a "Pseudomonadales" in the order Gammaproteobacteria that was originally isolated from the roots of healthy nursery-produced olive plants cv. Picual in Córdoba province, Spain. This strain was selected for sequencing based on its ability to exert biocontrol against Verticillium wilt of olive and to develop an endophytic lifestyle within olive root tissues. Such properties likely have origins in a repertoire of genes including a putative T3SS, two putative T6SS, and several genes presumably implicated in siderophore production. It also has a collection of genes predicting adhesion proteins, detoxifying compounds, volatile components and enzymes such as a cellulase, aphytase and a deaminase. Further functional studies and comparative genomics with related isolates will provide insights into biocontrol and endophytism.

Material and Methods

Growth conditions and DNA isolation

P. fluorescens strain PICF7 was grown in 50 ml of LB medium and incubated for 16 h at 28°C. After this period of time, the OD600 of the culture was 1.2. Serial dilutions from this culture and plating on LB plates yielded 2.8 x 108 CFU/mL of a pure bacterial culture (colonies showed uniform morphology and kanamycin resistance). The culture was divided into two 25-ml aliquots and total genomic DNA was extracted using the 'Jet-Flex genomic DNA purification' kit (Genomed GmbH, Löhne, Germany), according to the manufacturer's indications. DNA samples were further purified by extraction with phenol:chloroform and precipitation with ethanol. DNA quality and quantity were checked by agarose gel electrophoresis, spectrophotometry using a ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE), and digestion with different restriction enzymes. Two DNA aliquots (0.6 μg/μL, ~20 μg each) were sent in a dry ice container to the sequencing service.

Genome sequencing and assembly

The genome of PICF7 was sequenced at the Beijing Genomics Institute (BGI) using Solexa paired-end sequencing. Draft assemblies were based on 3,482,351 reads with a length of 500 bp resulting in 1,200 Mb, 2,456,221 reads with a length of 2,000 bp resulting in 1,209 Mb and 1,924,515 reads with a length of 6,000 bp resulting in 1,309 Mb. The SOAPdenovo 1.05 software package (Li *et al.*, 2008, 2010) developed by BGI was used for sequence assembly and quality assessment.

Genome annotation

Automatic annotation was performed using the NCBI Prokaryotic Genome Annotation Pipeline (Angiuoli *et al.*, 2008). Identification of known type III effectors effectors was conducted by BLASTP searches of the effectors described in *http://pseudomonas-syringae.org/* against the predicted protein sequences of PICF7. Functional annotation was performed by aligning the latter CDSs against the COG PSSM of the CDD (Marchler-Bauer *et al.*, 2014) using RPS-BLAST. Hits with an E-value <= 0.001 were first retained. Then, only the best hit was selected for each protein. Signal peptides and transmembrane helices were predicted using SignalP (Emanuelsson *et al.*, 2007) and TMHMM (Krogh *et al.*, 2001), respectively. T346Hunter (Martínez-García *et al.*, 2015b) was used to identify secretion systems clusters. Genetic factors potentially implicated in bacteria-plant associations

were predicted using a customized annotation tool developed by our group (not published). See Chapter IV for further details on such a tool.

CHAPTER III

T346Hunter: A novel web-based tool for the prediction of type III, type IV and type VI secretion systems in bacterial genomes

Pedro Manuel Martínez-García, Cayo Ramos and Pablo Rodríguez-Palenzuela. T346Hunter: A novel web-based tool for the prediction of type III, type IV and type VI secretion systems in bacterial genomes. *PLoS ONE*. 2015. doi: 10.1371/journal.pone.0119317.

Summary

T346Hunter (Type Three, Four and Six secretion system Hunter) is a web-based tool for the identification and localisation of type III, type IV and type VI secretion systems (T3SS, T4SS and T6SS, respectively) clusters in bacterial genomes. Non-flagellar T3SS (NF-T3SS) and T6SS are complex molecular machines that deliver effector proteins from bacterial cells into the environment or into other eukaryotic or prokaryotic cells, with significant implications for pathogenesis of the strains encoding them. Meanwhile, T4SS is a more functionally diverse system, which is involved in not only effector translocation but also conjugation and DNA uptake/release. Development of control strategies against bacterial-mediated diseases requires genomic identification of the virulence arsenal of pathogenic bacteria, with T3SS, T4SS and T6SS being major determinants in this regard. Therefore, computational methods for systematic identification of these specialised machines are of particular interest. With the aim of facilitating this task, T346Hunter provides a user-friendly web-based tool for the prediction of T3SS, T4SS and T6SS clusters in newly sequenced bacterial genomes. After inspection of the available scientific literature, we constructed a database of hidden Markov model (HMM) protein profiles and sequences representing the various components of T3SS, T4SS and T6SS. T346Hunter performs searches of such a database against user-supplied bacterial sequences and localises enriched regions in any of these three types of secretion systems. Moreover, through the T346Hunter server, users can visualise the predicted clusters obtained for approximately 1700 bacterial chromosomes and plasmids. T346Hunter offers great help to researchers in advancing their understanding of the biological mechanisms in which these sophisticated molecular machines are involved. T346Hunter is freely available at *http://bacterial-virulence-factors.cbgp.upm.es/T346Hunter*.

Introduction

The secretion of large molecules across the cell envelope is an essential bacterial mechanism involved in their survival and adaptation to diverse environments. Proteins are transported from the bacterial cell to the environment or directly into eukaryotic or prokaryotic cells (Pukatzki *et al.*, 2007; Wandersman, 2013). The general secretory pathway (Sec) and the two-arginine (Tat) translocation pathway, which are universal machineries shared by bacteria, archaea and eukaryotes (Tseng *et al.*, 2009), are sufficient for protein secretion in Gram-positive bacteria. Meanwhile, in Gram-negative didermic bacteria, these two pathways translocate proteins into the periplasm but not across the outer membrane (OM). This second membrane system serves as a protective structure against antibiotics and antimicrobial host compounds and enables the colonisation of host environments. However, it also presents an impediment to protein secretion, and throughout evolution, Gram-negative bacteria have developed sophisticated mechanisms for the translocation of proteins across the cell envelope. Similarly, Gram-positive bacteria with a cell wall heavily modified by lipids, such as mycobacteria, have also evolved refined machineries for protein secretion (Houben *et al.*, 2014). So far, seven general classes of secretion systems have been identified, numbered T1SS to T7SS (Tseng *et al.*, 2009). All these systems play a crucial role in the interaction of bacteria with the environment, particularly during the relationships established with eukaryotic host cells.

In terms of pathogenicity, secretion systems represent major virulence determinants for bacteria harbouring them. A broad range of secretion systems has been described for plant, animal, human and fish pathogens (Gerlach, Hensel, 2007; Tseng *et al.*, 2009). By means of secreted enzymes such as proteases, lipases and pectate lyases, bacteria are able to degrade eukaryotic cell wall components and metabolise host polymers by decomposing them. These enzymes are exported to the environment, and their secretion is executed mainly by T1SS, T2SS and T5SS (Wandersman, 2013). On the other hand, effector proteins are injected into host cells by T3SS, T4SS and T6SS (Lindeberg *et al.*, 2012; Zechner *et al.*, 2012; Russell *et al.*, 2014). Effectors have the ability to produce physiological changes in the host, fulfilling essential functions during the interaction between bacteria and eukaryotes. T3SS effectors secreted by *Salmonella enterica*, a pathogen that causes gastroenteritis and typhoid fever in humans, help bacteria to modulate host immune signals (Figueira, Holden, 2012). VirB effectors delivered by the T4SS VirB system of *Brucella* contribute to the intracellular growth of this pathogen, as well as to its persistence in the livers of mice (Myeni *et al.*, 2013). Meanwhile, effector protein delivery via T6SS has been shown to provoke actin cytoskeleton disruption and apoptosis in HeLa cells (Suarez *et al.*, 2010). Toxins,

which are secreted by all secretion systems, help pathogenic bacteria to promote infection by damaging host tissues and by modulating the host immune response (Henkel *et al.*, 2010). Such is the case for the cholera toxin, secreted by *Vibrio cholerae* via the T2SS (Davis *et al.*, 2000), and AexT, an *Aeromonas hydrophila* T3SS-secreted toxin (Vilches *et al.*, 2008). *Agrobacterium tumefaciens* promotes tumour formation in plants by transferring cytokinin- and auxin-coding genes to the plant from the T-DNA plasmid, which also encodes a T4SS required for its transfer (Vergunst *et al.*, 2000).

Current research on secretion systems of pathogenic bacteria is targeted at identifying such systems in bacterial genomes and at characterising the functions of their secreted effectors, with a special focus on T3SS, T4SS and T6SS. There exists a wide variety of each of these systems and the effectors they deliver. Distribution of effector proteins varies both among different species and among different strains of the same species, and strains isolated from diverse locations may have significantly divergent effector repertoires (Lindeberg *et al.*, 2012). Therefore, the design of new control strategies against bacterial pathogens first requires the identification of T3SS, T4SS and T6SS, and of their secreted effectors (Baron, 2010). Functional studies will then allow a deeper understanding of the molecular mechanisms of action of these secretion systems and their targets in eukaryotic cells.

Due to the rapid progression of advances in genome sequencing and with plunging costs, a large amount of genomic data is being generated every day. Bacteriologists routinely make use of high-throughput sequencing technologies to obtain the whole genome sequences of strains of interest, requiring computational methods for automatic identification of bacterial virulence machineries. Nonetheless, few web-based applications have been developed to predict secretion systems components. SSPred is a web server based on support vector machine (SVM) for the prediction of proteins involved in bacterial secretion (Pundhir, Kumar, 2011). It takes a set of amino acid sequences as input and classifies it into T1SS, T2SS, T3SS, T4SS or Sec secretion system. However, a maximum of four amino acid sequences can be submitted, and genome-wide analyses cannot be performed. T3DB is a T3SS related database that provides a tool for the identification of T3SS genes given user-supplied genomic sequences (Wang *et al.*, 2012). Again, genome-wide searches are not supported, and sequences need to be supplied one by one. Guglielmini *et al.* (2011) built the CONJscan-T4SSscan web server, which uses hidden Markov model (HMM) profiles to scan a set of protein sequences for T4SS components. Along similar lines, Abby and Rocha (2012) implemented a web server, called T3SSscan-FLAGscan, for the identification of T3SS components. The latter makes use of HMM profiles to not

only identify T3SS components but also discriminate between flagellar and non-flagellar ones, given a set of protein sequences. To our knowledge, no specific tool is available to predict T6SS components, and users typically rely on general annotation tools for their identification (Angiuoli *et al.*, 2008; Overbeek *et al.*, 2014).

One limitation of the above servers is that each input sequence is analysed individually, which makes them not suitable for the localisation of genomic clusters. Such a feature is of special interest given that secretion systems components are typically encoded within virulence-associated plasmids or pathogenicity islands. Besides, most of those tools either generate raw outputs from BLAST (Altschul, 2005) (T3DB) and HMMER (Eddy, 2011) (T3SSscan-FLAGscan) or just produce a shallowly informative output of the predictions (SSPred). Only CONJscan-T4SSscan presents the results in a tab-delimited schema, which is certainly a more helpful format, but still requires some user manipulation. None of the before drawbacks is found in the annotation tool provided by SecReT4, an open-access web database of information on the T4SS (Bi *et al.*, 2013). A modest limitation of this tool is that it only accepts one input sequence, being not suitable for the analysis of draft genomes.

In this context, we developed T346Hunter (Type Three, Four and Six secretion system Hunter), a novel web-based tool designed to facilitate the identification of T3SS, T4SS and T6SS encoded in newly sequenced bacterial genomes. T346Hunter makes use of a database of HMM profiles and protein sequences to automatically annotate and localise T3SS, T4SS and T6SS in user-supplied bacterial genomes. By exploring the available scientific literature, we constructed a database of protein components that captures the diversity of these three types of secretion systems. Once the database search is performed and the secretion systems clusters have been localised, the system presents the results in a comprehensive and user-friendly formatted document, which can be accessed online or downloaded. Furthermore, T346Hunter accepts submissions of both complete and unfinished genomes. We think T346Hunter represents a valuable tool for researchers to help further their understanding of secretion systems in the context of pathogenesis.

Results and Discussion

Using T346Hunter

T346Hunter provides a simple and user-friendly web interface (Figure 1A) for the search of T3SS, T4SS and T6SS in user-supplied genomic sequences. A user can upload the sequence of interest at *http://bacterial-virulence-factors.cbgp.upm.es/T346Hunter* in two ways. First, the user can provide raw DNA sequences in FASTA format, in which case the system makes use of Glimmer v3.02 (Delcher *et al.*, 2007) to predict coding regions. When more than one sequence is submitted, T346Hunter interprets the set of sequences as a draft genome (Figure S1). Alternatively, the user can upload an NCBI-formatted sequence. In such a case, three types of data have to be uploaded to the server, including DNA sequence, protein sequences and the location of protein-coding genes. The format of the data should be similar to that used by the NCBI Genome FTP Server (fna, faa and ptt files). Users are encouraged to submit their sequences in this format, since T346Hunter does not take phylogeny into account to predict coding regions, leading to a potential decrease in the quality of gene calls. Besides the input sequence, several parameters are provided for the user to configure the search, including HMMER3 (Eddy, 2011) and BLASTp (Altschul, 2005) E-value thresholds, secretion systems to be predicted and sequence shape (circular or linear). Once the prediction is completed, secretion systems loci are displayed in an intuitive HTML document containing tabulated and graphical information of the regions, along with a whole-genome graphical view. A tab-delimited summary of the gene-by-gene search, which can be easily incorporated into data pipelines, is also provided. These results are stored on the T346Hunter server for a week.

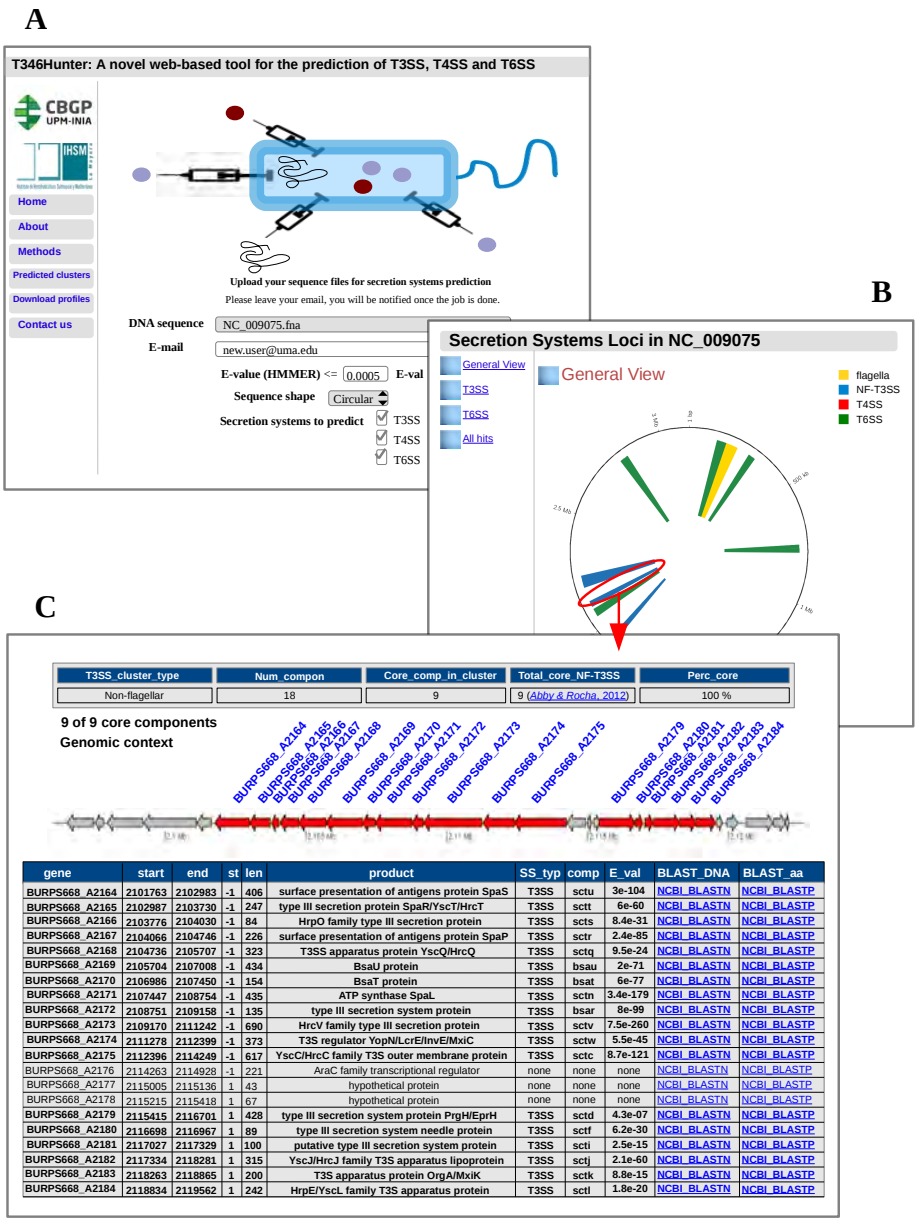
T346Hunter output

Here, the result of the execution of T346Hunter on the sequence of *Burkholderia pseudomallei* strain 668 is shown as an example (Figure 1). The genome of *B. pseudomallei* is typically comprised of two circular chromosomes, which encode several secretion systems (Holden *et al.*, 2004; D'Cruze *et al.*, 2011; Burtnick *et al.*, 2011). *B. pseudomallei* is an aerobic, Gram-negative bacterium that infects humans, animals and even plants. It is the causative agent of melioidosis, an often-fatal disease that is endemic to Southeast Asia and Northern Australia, and whose infection can take place by ingestion, inhalation and skin abrasion. There is no vaccine available to protect against this pathogen, which is also highly resistant to antibiotics. All this makes it a potential organism to be used as a bioterrorism agent (Estes *et al.*, 2011), and a growing interest on its virulence

mechanisms has recently emerged. Consequently, hundreds of genome sequences are nowadays available for *B. pseudomallei* (Nandi *et al.*, 2014), and the role of T3SS and T6SS in the virulence of this species has been previously reported (D'Cruze *et al.*, 2011; Burtnick *et al.*, 2011). Figure 1B shows the whole-genome graphical overview generated by T346Hunter displaying the predicted secretion systems clusters for *B. pseudomallei* 668 chromosome 2 (RefSeq NC_009075). The system localises three clusters of NF-T3SS, one cluster of flagellar T3SS and five clusters of T6SS. These predictions are consistent with previously reported *in silico* analyses (Boyer *et al.*, 2009; Abby, Rocha, 2012). The *bsa* NF-T3SS of *B. pseudomallei* has been shown to be an important part of the virulence armoury of this strain (Stevens *et al.*, 2004). Figure 1C shows the output generated by T346Hunter containing detailed information of such cluster, including a tabulated output and a gene map graphic representing its genomic context.

Core components

The sets of core components used for T3SS, T4SS and T6SS were as described by Abby and Rocha (2012), Bi *et al.* (2013) and Shrivastava and Mande (2008), respectively. However, there is no consensus on the definition of core in terms of secretion systems components. As long as we understand, "core" is the minimum set of components experimentally proven to be necessary for a secretion system to be functional. That appears to be the meaning used by Abby and Rocha (2012) and Shrivastava and Mande (2008) when they suggest a set of T3SS core components and T6SS components of major requirement, respectively. On the other hand, Bi *et al.* (2013) do not explicitly describe minimum required sets of components, but suggest a list of core components for each of the 18 T4SS they collect in their database. It is not clear though whether such proteins are indispensable for these T4SS to be functional. For instance, the trb T4SS encoded by *A. tumefaciens* C58 (Li *et al.*, 1998) lacks TrbN, which belongs to the above core list. Given this controversy, T346Hunter makes no discrimination regarding the different uses of the core set, leaving to the user the role of interpreting the results. We chose 4 as the minimum threshold of core components after trying different values and manually screening the predicted secretion systems, since it offers a trade-off between false positives and false negatives. Again, it is the user who has to sift through the predictions. Further information about core components can be found in Table S1.



gene	start	end	st	len	product	SS_typ	comp	E_val	BLAST_DNA	BLAST_aa
BURPS668_A2164	2101763	2102983	-1	406	surface presentation of antigens protein SpaS	T3SS	sctu	3e-104	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2165	2102987	2103730	-1	247	type III secretion protein SpaR/YscT/HrcT	T3SS	sctt	6e-60	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2166	2103776	2104030	-1	84	HrpO family type III secretion protein	T3SS	scts	8.4e-31	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2167	2104066	2104746	-1	226	surface presentation of antigens protein SpaP	T3SS	sctr	2.4e-85	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2168	2104736	2105707	-1	323	T3SS apparatus protein YscQ/HrcQ	T3SS	sctq	9.5e-24	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2169	2105704	2107008	-1	434	BsaU protein	T3SS	bsau	2e-71	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2170	2106986	2107450	-1	154	BsaT protein	T3SS	bsat	6e-77	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2171	2107447	2108754	-1	435	ATP synthase SpaL	T3SS	sctn	3.4e-179	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2172	2108751	2109158	-1	135	type III secretion system protein	T3SS	bsar	8e-99	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2173	2109170	2111242	-1	690	HrcV family type III secretion protein	T3SS	sctv	7.5e-260	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2174	2111278	2112399	-1	373	T3S regulator YopN/LcrE/InvE/MxiC	T3SS	sctw	5.5e-45	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2175	2112396	2114249	-1	617	YscC/HrcC family T3S outer membrane protein	T3SS	sctc	8.7e-121	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2176	2114263	2114928	-1	221	AraC family transcriptional regulator	none	none	none	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2177	2115005	2115136	1	43	hypothetical protein	none	none	none	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2178	2115215	2115418	1	67	hypothetical protein	none	none	none	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2179	2115415	2116701	1	428	type III secretion system protein PrgH/EprH	T3SS	sctd	4.3e-07	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2180	2116698	2116967	1	89	type III secretion system needle protein	T3SS	sctf	6.2e-30	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2181	2117027	2117329	1	100	putative type III secretion system protein	T3SS	scti	2.5e-15	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2182	2117334	2118281	1	315	YscJ/HrcJ family T3S apparatus lipoprotein	T3SS	sctj	2.1e-60	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2183	2118263	2118865	1	200	T3S apparatus protein OrgA/MxiK	T3SS	sctk	8.8e-15	NCBI_BLASTN	NCBI_BLASTP
BURPS668_A2184	2118834	2119562	1	242	HrpE/YscL family T3S apparatus protein	T3SS	sctl	1.8e-20	NCBI_BLASTN	NCBI_BLASTP

Figure 1. Example of execution of T346Hunter using the sequence of *B. pseudomallei* 668 chromosome 2 as input. **A**. Web interface of T346Hunter. **B**. Genome-wide graphical view showing the predicted secretion systems of *B. pseudomallei* 668 chromosome 2. **C**. Genomic representation of one of the three NF-T3SS clusters identified, including a graphical gene map and a tabulated gene list with detailed information of each component. Hyperlinks to NCBI for direct execution of BLASTn and BLASTp against the non-redundant nucleotide and protein databases are provided for each gene within the loci. Some other relevant information is also included, such as the percentage of core components found in the cluster and PubMed hyperlinks to the studies we have based our methods on to build the component profiles found in such a cluster.

Identification of T3SS, T4SS and T6SS gene clusters in sequenced bacterial genomes

Complete bacterial genomic sequences of 2,997 chromosomes and 2,164 plasmids sequenced available as of 14 February 2014 were downloaded from the NCBI RefSeq project Pruitt *et al.* (2012). T346Hunter was executed on these sequences and localised clusters enriched in either NF-T3SS, T4SS or T6SS components. In total, 2,814 clusters were identified (512 NF-T3SS, 1,466 T4SS and 836 T6SS) across 1,121 organisms. Predicted clusters are summarised in Table S2 and can be queried at the *Predicted Clusters* interface from the T346Hunter website. Sequences with negative predictions are listed in Table S3.

Comparison with currently available tools

In order to validate the performance of T346Hunter, systematic comparisons with other available applications for secretion systems prediction would certainly be the best choice. By crossing our predictions with loci identified by other servers one could have a measure of the relative accuracy of our tool. But, in practise, this is not straightforward to carry out. On the one hand, predicted data are not always available in a format that are ready to be systematically analysed. Servers usually provide their data in a way that either they have to be queried using some kind of criteria (e.g. strain name) or they are just embedded in the webpage. On the other hand, few servers are available to predict secretion systems clusters as such. To our knowledge, only SecReT4 (Bi *et al.*, 2013) provides a specific tool for genomic localisation of T4SS clusters. Despite these difficulties, and given the need for assessing the accuracy of our predictions relative to others' work, we attempted to accomplish comparisons either by systematic processing, when possible, or by manual inspection, when not.

First, we aimed to compare our predictions of T3SS with those of T3SSscan-FLAGscan (Abby, Rocha, 2012). Such server does not localise T3SS clusters in user-submitted sequences, but does keep a repository of predicted NF-T3SS loci that can be accessed using different features. Therefore, we randomly queried clusters for 100 genomic sequences and manually compared them with our predictions. Since negative predictions are not provided in the server, this selection was restricted to positive predictions. We found that T346Hunter predicted any NF-T3SS cluster in the 100 sequences examined. More precisely, T346Hunter and T3SSscan-FLAGscan identified the same number of clusters in 95 sequences (Table S4). For each of the resting five sequences, the number of predicted clusters just differs in one. This difference is probably explained by the different searching criteria used by both methods, particularly in defining contiguous genes within a cluster and setting the minimum required number of core components. In total, 125 clusters were

identified by T3SSscan-FLAGscan and 128 by T346Hunter.

To test how T346Hunter performs in predicting T4SS, we compared it with SecReT4 (Bi *et al.*, 2013). This server provides a summary of T4SS predictions on a number of sequences by means of a table in the webpage. We could, then, proceed to systematically compare the two methods. Again, such a comparison was necessarily restricted to positive predictions. In total, 387 sequences were examined and all of which were found to contain at least one T4SS cluster. Of 387 such replicons, 324 (84%) were predicted by both methods to contain the same number of T4SS (Table S5). In this case, differences in searching criteria may have had a stronger impact in the predictions. SecReT4, for instance, does not restrict T4SS genes to be located in a specific cluster. In contrast, T346Hunter identifies a cluster whenever orthologues of 4 core components are found in a window of up to 70 kb. Despite such divergent parametrisation, 53 out of the 63 sequences with discordant predictions (84%) only differ in one cluster. Accounting for the 387 sequences analysed, SecReT4 and T346Hunter identified 522 and 595 T4SS, respectively.

We went ahead to inspect the accuracy of our T6SS predictions. As a specific server for the prediction of T6SS is not available, tool-by-tool comparison was precluded in this case. However, systematic localisation of T6SS loci has been previously performed (Shrivastava, Mande, 2008; Boyer *et al.*, 2009; Barret *et al.*, 2011), reporting data we could use to compare our predictions with. We focused on Boyer *et al.* (2009), which completed the widest analysis. Since no readily processable summary of the predictions was provided, comparisons had to be manually performed. Among the 100 sequences reported with positive predictions, T346Hunter identified at least one T6SS in 98 of them. Furthermore, both approaches predicted the same number of T6SS clusters in 93 sequences (Table S6). Such a subtle difference is explained by the requirement of 4 core components in a cluster imposed by T346Hunter, which was not applied by Boyer *et al.* (2009). Nonetheless, predictions on the resting 7 sequences of both approaches differed in only one cluster. Summing up all identified clusters in the 100 sequences examined, T346Hunter and Boyer *et al.* (2009) predicted 170 and 175, respectively.

Regarding general annotation engines, some of the most widely used tools are the NCBI Prokaryotic Genome Annotation Pipeline (Angiuoli *et al.*, 2008) and the RAST server system (Overbeek *et al.*, 2014). In the last few years, RAST has become particularly popular and is now frequently used to rapidly annotate bacterial genomes against its comprehensively curated subsystem database. Due to its constant growth, RAST automated annotations are nowadays of a great quality, having reached a high degree of specificity. Indeed, T3SS, T4SS and T6SS are included among the subsystems

collection of RAST, and thorough reports of related genes are provided within general annotations. Such reports include visual and tabular information of the corresponding genomic clusters, thus offering an exhaustive output. However, bacterial strains that has not been incorporated into RAST database are reported with no subsystems, and users need to manually inspect individual features to infer the existence of secretion systems clusters. This makes RAST subsystems search dependent on its database of bacterial isolates, and makes it particularly not suitable for the analysis of newly characterised bacteria. Furthermore, when subsystems are reported, the number of secretion systems clusters are not directly shown in the output and it rather needs to be derived from reported tables. Besides, no information regarding core components is provided, and some conjugal T4SS are not categorised as such. Therefore, even though RAST performs quite well in detecting genes encoding secretion systems when compared to other general annotation tools, its annotations lack some relevant information on T3SS, T4SS and T6SS, and do not directly offer the whole picture of the underlying genomic clusters.

Conclusions

The development of web-based tools for the prediction of virulence factors is crucial for allowing researchers to identify the bacterial pathogenic arsenal. Here, we present T346Hunter, an online tool for annotation and localisation of secretion systems clusters in sequenced bacterial genomes. Because they are distinctive features of pathogenesis, T346Hunter searches for T3SS, T4SS and T6SS, whose identification is of particular interest in the development of strategies against bacterial-mediated diseases. The server will be continuously updated as new experimental and bioinformatics information on secretion systems becomes available. We believe T346Hunter will help researchers uncover the mechanisms of bacterial secretion as a virulence trait.

Material and Methods

Protein sequences and profiles

Sequence profiles of secretion systems components were generated by selecting orthologues of each component in order to capture the diversity of the T3SS, T4SS and T6SS. Following the approach described by Abby and Rocha (2012), we selected protein sequences corresponding to the components of the flagellar and non-flagellar T3SS (NF-T3SS). Meanwhile, to build profiles that represent the variety of T4SS, protein sequences of the components of 18 archetypal T4SS (Bi *et al.*, 2013) were also selected. In both cases, sequences of each component were extracted from a set of model organisms representative of the diversity of all these types of systems. We based the construction of T6SS components profile on a previously reported list of coding sequences belonging to several bacterial genomes; sequences that were found to be orthologues of the components of the first described T6SS, namely, *V. cholerae*, *Pseudomonas aeruginosa* and *Burkholderia mallei* (Shrivastava, Mande, 2008). The sequences of these orthologues were also included in our set of protein families. All these sequences were downloaded from Uniprot (Boutet *et al.*, 2007) and the NCBI website (*http://www.ncbi.nlm.nih.gov/*), and selected based on their corresponding genome annotations. Then, sequences corresponding to each component were aligned with Muscle (Edgar, 2004) and manually adjusted with Seaview (Gouy *et al.*, 2010). Finally, protein profiles were built with HMMER3 (Eddy, 2011). When fewer than five representative model organisms were found to code for a given component, no profile was built; instead, files were generated in multi-FASTA format.

We extended the above set by collecting protein sequences from AtlasT4SS (Souza *et al.*, 2012) and proceeding in the same way as above to generate profiles of orthologue clusters described in that database. Secretion systems loci identified in this study were manually screened, and additional profiles were incorporated based on the RefSeq genome annotation. Consequently, our final dataset comprises sequence information for a total of 364 components of the T3SS, T4SS and T6SS (65, 449 and 20, respectively). Further information about each of the component profiles can be found in Table S1.

Identification of secretion systems clusters

T346Hunter performs BLASTp (Altschul, 2005) and HMMER3 (Eddy, 2011) searches of the protein sequences and profiles described above against user-supplied genomic sequences. Regions containing homologous genes (E-value ≤ 0.0005 by default) of at least 4 different core components of T3SS, T4SS or T6SS and spanning up to 70 kb are retained

and included in the output report. We consider core components of T3SS, T4SS and T6SS as described by Abby and Rocha (2012), Bi *et al.* (2013) and Shrivastava and Mande (2008), respectively (see Table S1 for details).

Implementation

T346Hunter runs on a Linux platform with an Apache web server. The web interface was implemented using HTML and CSS, and data pipelines were developed using PHP, Perl, R and shell scripts. DNA and protein sequences are processed by means of Bioconductor (Gentleman *et al.*, 2004) and BioPerl (Stajich *et al.*, 2002). Circular genome images are generated using circos (Krzywinski *et al.*, 2009), and gene maps are produced using the R package genoPlotR (Guy *et al.*, 2010). Open reading frame predictions are generated with Glimmer v3.02 (Delcher *et al.*, 2007).

Supplementary Material

Figure S1. Overview of T346Hunter prediction workflow.

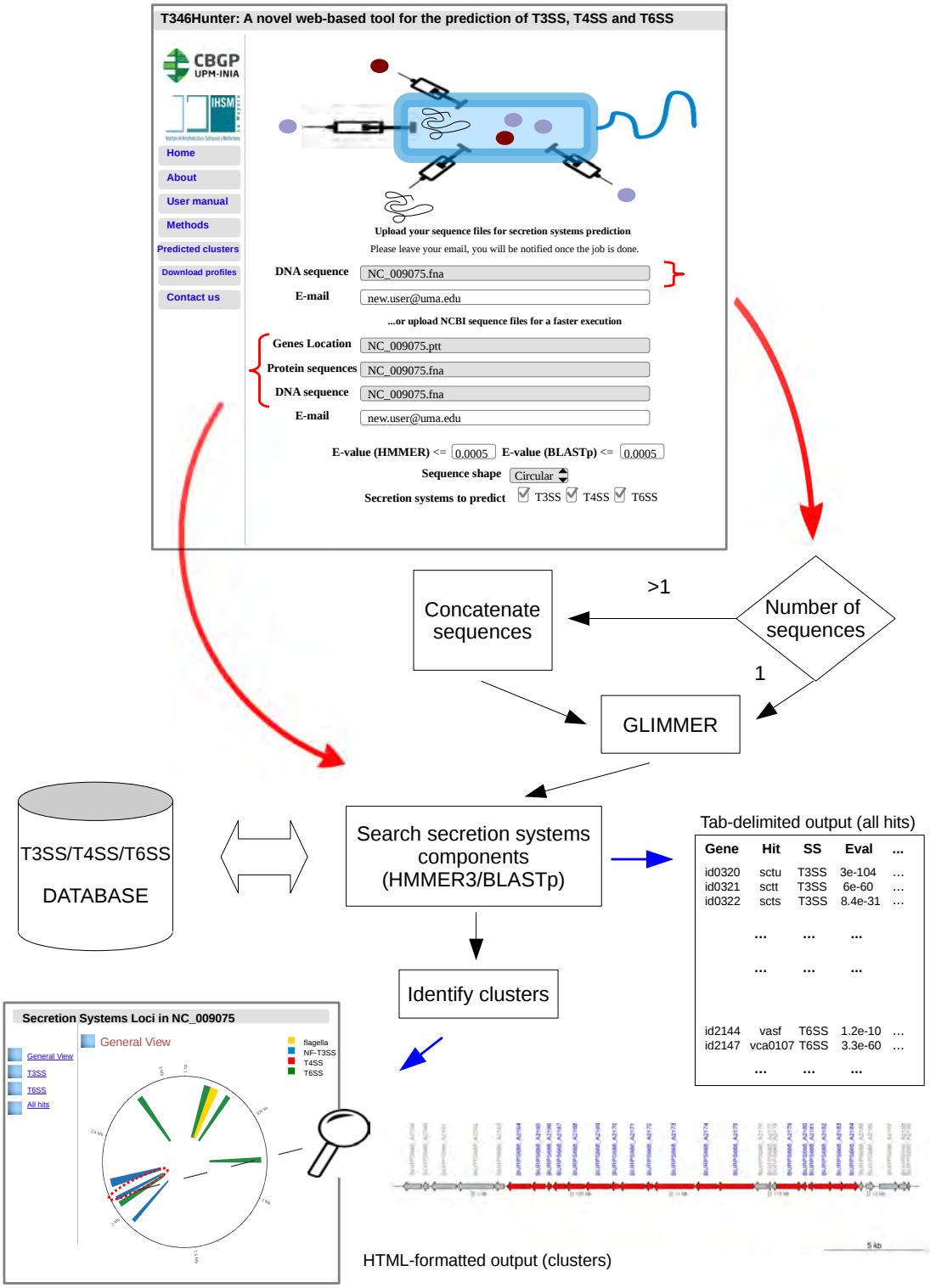


Table S1. Protein profiles and sequences used in this study. Available:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0119317.s002

Table S2. Summary of predicted clusters. Available:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0119317.s003

Table S3. List of complete bacterial genomic sequences and plasmids not found to encode NF-T3SS, T4SS or T6SS. Available:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0119317.s004

Table S4. Summary of NF-T3SS clusters predicted by T3SSscan-FLAGscan and T346-Hunter on 100 sequences. Available:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0119317.s005

Table S5. Summary of T4SS clusters predicted by SecReT4 and T346Hunter on 387 sequences. Available:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0119317.s006

Table S6. Summary of T6SS clusters identified in Boyer *et al.* (2009) and by T346Hunter on 100 sequences. Available:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0119317.s007

Table S7. Predictions of secretion systems clusters not fulfilling the restriction of 4 core components. Available:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0119317.s008

Data S1. Hidden Markov Model profiles and sequences used by T346Hunter. Available:
http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0119317.s009

CHAPTER IV

A supervised machine-learning approach for the prediction of bacterial associations with plants

Pedro Manuel Martínez-García, Emilia López-Solanilla, Cayo Ramos and Pablo Rodríguez-Palenzuela. A supervised machine-learning approach for the prediction of bacterial associations with plants. Under revision with *PLoS Pathogens*.

Summary

In the present study we base on genes reported to affect the outcome of plant-bacteria interactions to develop a classifier of plant-associated bacterial genomes. With this aim, we inspected the existing literature and collected 170 bacterial factors involving 603 gene products (accessible online at *http://bacterial-virulence-factors.cbgp.upm.es/PIFAR*). These factors were used in an *ad hoc* bioinformatics pipeline that works in two stages. First, they were searched against a selection of 420 bacterial genomes known to develop diverse interactions with plants. Then, the results of this search was assembled to train probabilistic classifiers by means of supervised machine-learning methods. After evaluating classification performances, we concluded that our collection of factors is best suited to separate plant-associated from non-plant-associated bacteria, achieving a precision rate over 94%. Adhesion, deconstruction of the plant cell wall and detoxifying activities were highlighted as the most predictive features when classifying plant-associated bacteria. Finally, we applied our classifier to the whole set of sequenced bacterial genomes available at NCBI and focused on a selection of bacterial genera. A number of well-known human pathogens were assigned significant plant-associated probabilities, enteric bacteria being the most prominent. Our data agrees with recent scenarios of microbial contamination of food plants that have resulted in important foodborne illnesses. We expect the application of our approach in epidemiological studies can help researchers in gaining insights into plant pathology and food safety.

Introduction

Bacteria that live in association with plants are diverse in several aspects, such as the habitats they colonize, their phylogeny and their influence on plant growth and yield (Gnanamanickam, 2006). Some of them establish in the rhizosphere or within plant root tissues and may be beneficial to their hosts. These bacteria are usually referred to as plant growth-promoting rhizobacteria (PGPR), and they can contribute to plant fitness by helping to prevent the deleterious effects of phytopathogenic organisms (Van Loon, 2007; Lugtenberg, Kamilova, 2009). On the contrary, plant pathogenic bacteria typically enter the plant through natural openings and wounds, colonize host tissues and cause diseases in crops, leading to important economic losses (Agrios, 2005; Mansfield *et al.*, 2012). Other phytobacteria are ubiquitous, and possess a virulence potential that allows them to infect a broad range of organisms (Rahme *et al.*, 1995). Such is the case of *Pseudomonas aeruginosa*, a well-known opportunistic pathogen that has been described to cause significant infections in plants, insects, nematodes and mammals (Mahajan-Miklos *et al.*, 1999; Jander *et al.*, 2000; He *et al.*, 2004). In the last few years, a number of cases involving microbial contamination of food plants have resulted in important outbreaks, and a growing interest in the implicated factors has recently emerged (Holden *et al.*, 2009; Lim *et al.*, 2014). Traditional human pathogens such as *Salmonella* and *Escherichia coli* have been identified as sources of such foodborne illnesses, which opens a door to speculation about whether other human-associated bacteria have also developed strategies to colonize plant hosts (Deering *et al.*, 2012; Schikora *et al.*, 2012; Fletcher *et al.*, 2013). Bacterial associations with plants requires complex interactions that involve multiple molecular processes in which a wide range of factors are implicated. Our motivation in this study has been to address these associations basing on the current genomic knowledge in the field.

Whether a given bacterium is able to develop some kind of interaction with a host is influenced by a number of elements that intervene to varying degrees, including environment, bacterial gene components and regulatory networks, or the host itself. The vast amount of data that is nowadays available as a result of countless sequencing projects has motivated researchers to explore whether bacteria-host associations can be identified by means of genomic information. Previous works have based on sequence similarity to classify bacterial pathogenicity, typically by drawing upon pre-established virulence-related genes (Iraola *et al.*, 2012) or genes that are distinctive of a collection of genomes (Andreatta *et al.*, 2010; Barbosa *et al.*, 2014). These approaches perform quite well in distinguish pathogenic from non-pathogenic human-associated bacteria. However, when predicting different pathogenicity classes (i.e. opportunistic), sequence information seems

to be limited (Barbosa *et al.*, 2014). Although a wide set of gene components has been reported to be implicated in bacterial associations with plants, their predictive potential to classify those associations has not been tested. Current knowledge on these genes are stored in several resources that hold comprehensive information of different aspects of plant-associated bacteria. These resources include the Comprehensive Phytopathogen Genomics Resource (CPGR) (Hamilton *et al.*, 2011), the Plant Associated and Environmental Microbes Database (PAMDB) (Almeida *et al.*, 2010), the *Pseudomonas* Genome Database (Winsor *et al.*, 2011), the *Pseudomonas syringae* Genome Resources (Lindeberg, 2011), the PhytoPath project (*http://www.phytopathdb.org*), the Virulence Factor Database (VFDB) (Chen *et al.*, 2005) and The Pathogen Host Interactions Database (PHI-base) (Urban *et al.*, 2014). The above initiatives are the result of impressive efforts of data curation and have an unquestionable value for researchers in plant bacteriology. To address the problem of classifying bacterial associations with plants, though, it would be convenient to base on gene products that allow to capture the whole picture of such associations. This implies considering the underlying processes in a generic mode, and none of the above resources is structured in such a way. Bacterial association to plants requires genetic factors that could be framed in three possible steps: adhesion to plant surfaces, survival in the environment inside or in the vicinity of plants and active processes of colonization (Agrios, 2005; Gnanamanickam, 2006). Therefore, a well suited classifier may account for processes involved in these scenarios.

In this study, we have inspected the available literature to collect a set of genetic products found to be implicated in any of the above steps. Then, we used a bioinformatics strategy based on supervised machine learning to test the predictive potential of such repertoire in grouping bacteria with similar lifestyles. To the best of our knowledge, this is the first approach for a systematic identification of plant-associated bacteria based on genomic content, and we envisage it can have innovative applications in plant pathology and food safety.

Results and Discussion

PIFAR, an open-access web-based resource for plant-bacteria interaction factors

Although our aim for collecting bacterial factors was to test their predictive potential for the classification of plant-associated bacteria, we have also deposited them in a web server, called PIFAR (Plant-bacteria Interaction FActors Resource), so that they can be accessed by external users (Figure 1). PIFAR is freely available at *http://bacterial-virulence-factors.cbgp.upm.es/PIFAR*. The website has two main purposes: to hold comprehensive information on gene products described as implicated in bacterial interactions with plants and to help researchers to identify such products in input genome sequences.

PIFAR holds a curated set of products that cover a broad spectrum of the current genomic information on plant-associated bacteria, including pathogenic (i.e. *Pseudomonas syringae*) and non pathogenic species (i.e. *Rhizobium leguminosarum*). Entries in the database are structured in terms of the molecular processes they are involved. To account for adhesion and attachment to plant tissues, genes implicated in the synthesis of factors such as adhesins or exopolysaccharides (EPSs) were included (Pizarro-Cerdá, Cossart, 2006; Mhedbi-Hajri *et al.*, 2011). Regarding bacterial survival, genes involved in the production of factors like siderophores, antibiotics, lipopolysaccharides (LPSs) or multidrug efflux pumps (MDRs) were considered (Erbs, Newman, 2003; Wang, Raaijmakers, 2004; Martínez *et al.*, 2009; Ahmed, Holmström, 2014). Meanwhile, to account for plant-host colonization, genes implicated in the synthesis of factors such as phytotoxins, phytohormones or plant cell wall degrading enzymes (PCWDEs) were also incorporated (Dudler, 2014; Costacurta, Vanderleyden, 1995; Barras *et al.*, 1994). The final collection comprised 170 factors involving 603 gene products and 16 PFAM domains (Finn *et al.*, 2014).

Factors were selected by manual inspection of the available scientific literature. The curation based on selecting relevant genes reported to affect the outcome of bacteria-plant interactions. We are aware that the literature is somehow biased towards virulence factors, since pathogenesis has been more extensively studied than any other bacterial association to plants. Similarly, information on some bacterial species and/or genera is also over-represented. However, we do not conceive our repertoire of factors as finished. On the one hand, it will be updated as new experimental data becomes available. On the other hand, we expect it to grow by means of the contribution of other experts in plant bacteriology. To this end, PIFAR allows data submissions of missing factors or newly charaterized ones. Further information on how to use the web server can be found in File S1.

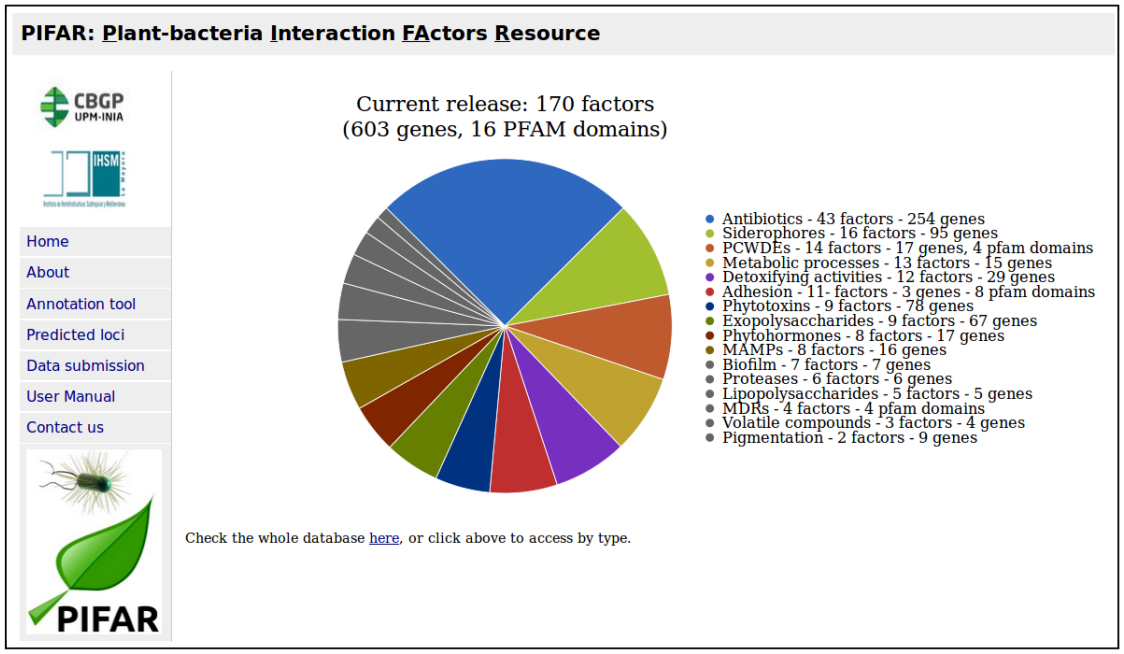


Figure 1. Front page of PIFAR. Several interfaces are available for the user to navigate throught the website (left; see File S1). The pie chart represents the distribution of bacterial factors included in the database. By clicking on the different segments, users can access the database by factor type.

A supervised machine learning strategy to classify plant-associated bacterial lifestyles

Three classifiers were trained to discriminate between different bacterial lifestyles (Figure 2). To that aim, we selected 420 sequenced bacterial strains (Figure 2A; Table S1) classified into three classes: plant-pathogenic (PP, 109 strains), plant-associated-non-pathogenic (PANP, 93 strains) and non-plant-associated (NPA, 218 strains). This classification was based on literature searches and NCBI annotations (see Material and Methods). Then, we performed a systematic screening of our database against such strains (Figure 2B). For each genome analyzed, we obtained its repertoire of plant-associated bacterial factors. Based on presence/absence of such factors, a vector was generated for each genome so that each position contained the count of the identified factors by type (Figure 2C). Therefore, a given feature in a vector represents the total number of factors of a certain type (i.e. phytotoxins) found in a given bacterial strain. These features were extended using T346Hunter (Martínez-García *et al.*, 2015b). T346Hunter is an annotation tool, developed by our group, which identifies genomic clusters potentially involved in the synthesis of secretion systems of types III, IV and VI (T3SS, T4SS and T6SS, respectively). Thus, three more features were incorporated to the vectors consisting of the numbers of T3SS, T4SS and T6SS genomic clusters that a given genome harbors (T4SS involved in conjugation and

DNA uptake/release were excluded). The result was a 420x19 matrix (Figure 2D) with no repeated rows (a unique vector resulted from each strain) that can be checked in Table S2. Finally, we used random forests (Breiman, 2001) to train three supervised machine-learning classifiers (Figure 2E and Figure 3).

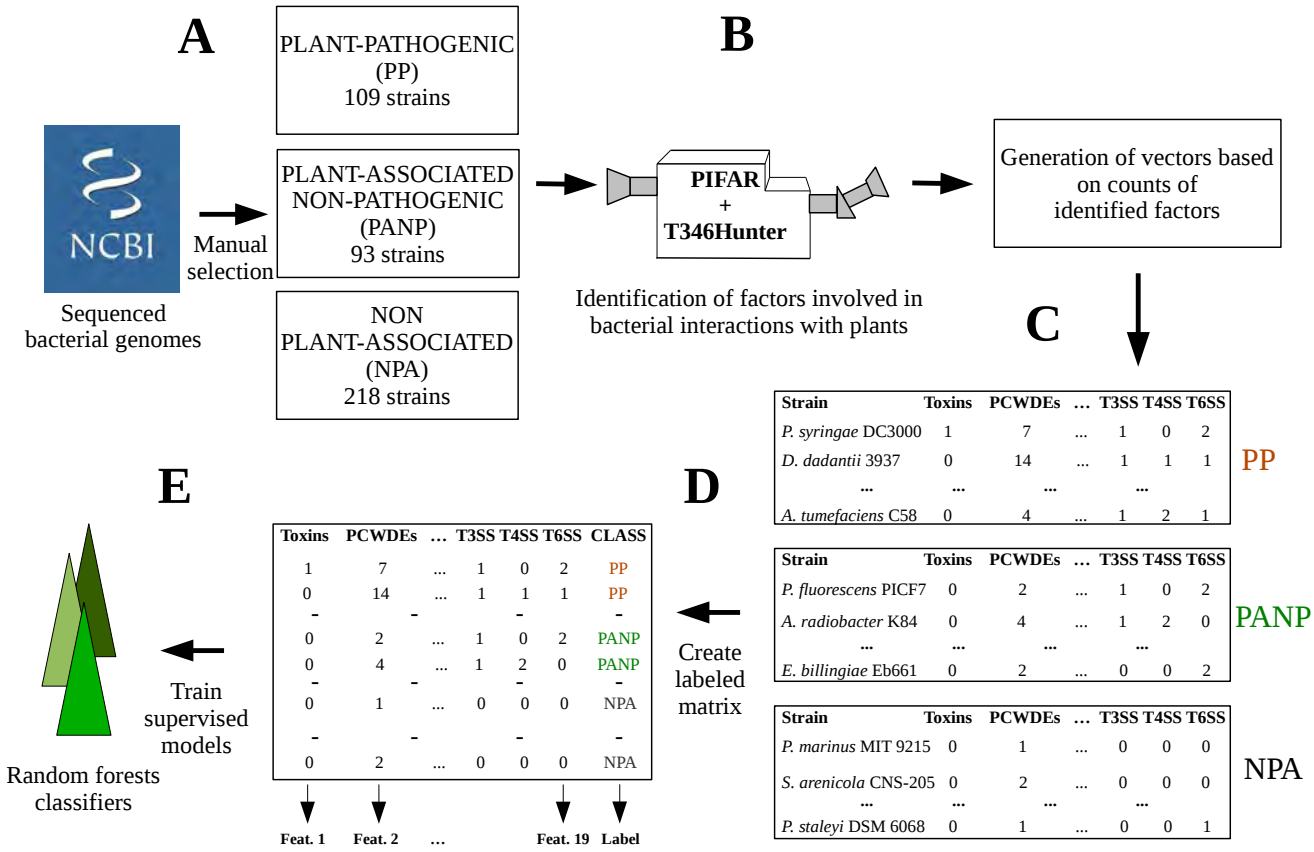


Figure 2. Schematic representation of the bioinformatics pipeline used for the generation of the training set. A. The genomes of a selection of 420 bacterial strains were retrieved from NCBI. Strains were selected according to three categories: plant-pathogenic (PP), plant-associated non-pathogenic (PANP) and non-plant-associated (NPA). **B**. PIFAR annotation tool was used to screen these strains for factors in the database. T346Hunter (Martínez-García *et al.*, 2015b) was used to identify genomic clusters encoding T3SS, T4SS and T6SS. **C**. The results of above searches were used to create vectors containing the repertoire of factors of each genome. **D**. Each vector was labeled as PP, PANP or NPA and a matrix was created consisting of all 420 vectors. **E**. The resulting matrix was used as the training set of three random forests classifiers.

First, we generated a 3-class model, which classified bacterial genomes into PP, PANP and NPA with a precision rate of 90% (Figure 3A; Table 1). The classifier seems to perform quite better in predicting NPA strains (212 out of 218; 97.3%) than PP (94 out of 109; 86.2%) or PANP (72 out of 93; 77.4%). The latter presents the highest error rate in this study (22.6%), which may be explained by the bias towards virulence factors existing in our database. This feature could be showing that, to date, more data on pathogenicity than on other kind of plant-bacteria interactions is available in the literature. Anyway, given its error rate, we consider this PP-PANP-NPA classifier as limited.

By combining PANP and NPA classes into a new class consisting of non-plant-pathogenic (NPP) strains, we generated a second model that classified bacterial genomes into PP and NPP with a precision rate of 94% (Figure 3B, Table 2). PP classes were correctly classified in 85.3% of the cases (93 out of 109), while NPP classification reached an accuracy of 97.7% (304 out of 311). Such a significant difference in error rates could barely be explained by the lightly unbalanced dataset (109 PP versus 311 NPP). More likely, it may be regarded as a consequence of the database content itself. It is possible that we are missing unknown factors that could help to better capture the pathogenic lifestyle of several phytobacteria, particularly those misclassified by our models. Such factors, yet to be described, would complement the repertoire of well-known virulence factors in these organisms, and their characterization would help improving the performance of this classifier. In any case, the available information on bacterial genetic factors seems to be limited to accomplish PP-NPP classification.

Finally, a third model was built by merging PP and PANP classes into a new class that harbors all plant-associated (PA) bacterial strains included in the dataset. Such model classified bacterial genomes into PA and NPA with a precision rate of 94.3% (Figure 3C, Table 3). In this case, the general error rate is the most similar to that of false positives and false negatives of the three classifiers. PA classes were correctly assigned in 188 out of 202 strains (93.1%), while NPA classes were correctly classified in 208 out of 218 strains (95.4%). This slight difference may be due to the fact that factors included in our database are by definition involved in associations to plants, and thus capture plant-associated bacterial traits. Since this is the positive class to be predicted, the unavoidable lack of information may be prejudicing PA classification. We expect this tendency will decrease as new experimental data is incorporated into the database. Nevertheless, the classifier shows a good performace, with a consistent number of correctly/incorrectly classified genomes (396/24) and an acceptable balance between false positives and false negatives (6.9%-4.6%). In this case, our repertoire of genetic components has enough predictive strength to separate

PA and NPA bacteria with high accuracy (94%). Therefore, we focused on this classifier for further analyses.

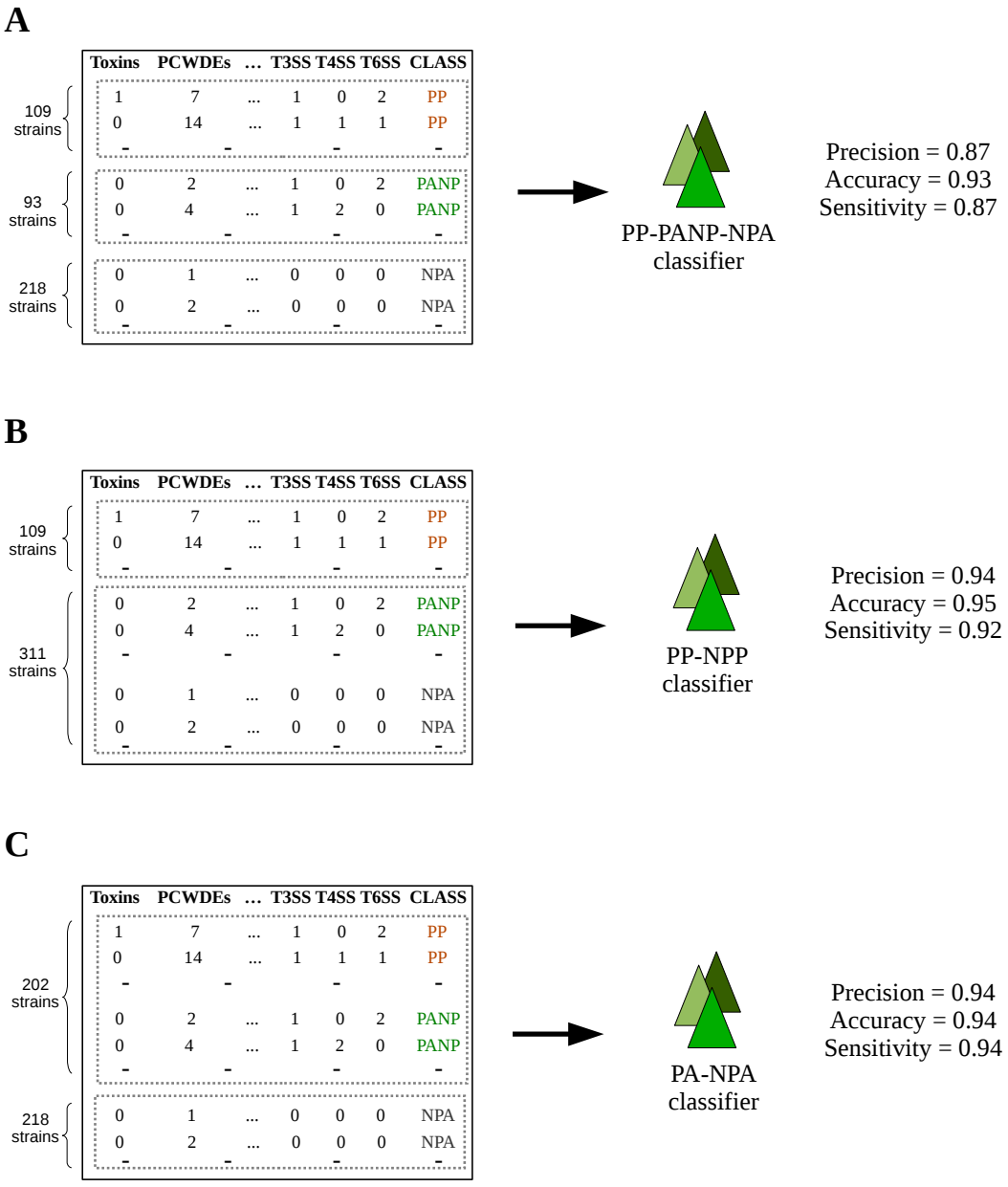


Figure 3. Training data assembly to train the three classifiers. A. PP VS PANP VS NPA classifier. Assembly of training data as generated in Figure 2. **B**. PP VS NPP classifier. PANP and NPA are combined to form the NPP class. **C**. PA VS NPA classifier. PP and PANP are combined to form the PA class.

Table 1. Confusion matrix showing classification performance of the three-classes classifier. PP, PANP and NPA refer to plant-pathogenic, plant-associated non-pathogenic and non-plant-associated, respectively.

	Total	Predicted as PP	Predicted as PANP	Predicted as NPA
Plant-pathogenic	109	94 (86.2%)	8 (7.3%)	7 (6.4%)
Plant-associated (NP)	93	5 (5.4%)	72 (77.4%)	16 (17.2%)
Non-plant-associated	218	2 (0.9%)	4 (1.8%)	212 (97.3%)

Table 2. Confusion matrix showing classification performance of the PP-NPP classifier. PP and NPP refer to plant-pathogenic and non-plant-pathogenic, respectively.

	Total	Predicted as PP	Predicted as NPP
Plant-pathogenic	109	93 (85.3%)	16 (14.7%)
Non-plant-pathogenic	311	7 (2.3%)	304 (97.7%)

Table 3. Confusion matrix showing classification performance of the PA-NPA classifier. PA and NPA refer to plant-associated and non-plant-associated, respectively.

	Total	Predicted as PA	Predicted as NPA
Plant-associated	202	188 (93.1%)	14 (6.9%)
Non-plant-associated	218	10 (4.6%)	208 (95.4%)

Adhesion, plant cell wall degradation and detoxification as the most predictive features for identifying plant-associated bacteria

An attractive aspect of random forests (Breiman, 2001) is that it assesses an importance measure to each input variable in the final predictions. This importance is computed by measuring the decrease of classification performance when the values of a given variable in a given tree of the forest are randomly permuted (Chen, Ishwaran, 2012). The use of random forests importances to rank variables in our prediction model can give us hints on which repertoire of bacterial factors are the most relevant in the classification, and therefore, are of biological interest. Figure 4 shows a ranking with the importance measure assigned to each of the features in the PA-NPA classifier. Factor types having the largest importance are those related to bacterial adhesion. Adhesion mechanisms are widely conserved in both pathogenic and non-pathogenic plant associated bacteria, and are used by multiple

species as a step previous to colonization or biofilm aggregation (Pizarro-Cerdá, Cossart, 2006; Rigano *et al.*, 2007). The second feature in the ranking of importances correspond to PCWDEs. The secretion of PCWDEs is a widely distributed machanism of plant-associated bacteria, and are used to penetrate the cell wall prior to plant infection. PCWDEs have been extensively studied in soft-rot enterobacteria, such as *Pectobacterium* and *Dickeya* (Barras *et al.*, 1994; Toth *et al.*, 2003). The third type of factors in the ranking correspond to those involving bacterial detoxification. In the last two decades, thorough research has been dedicated to the mechanisms that enable plant-associated bacteria to resist the action of plant products such as antimicrobial agents (López-Solanilla *et al.*, 1998; García-Olmedo *et al.*, 2001) or active oxygen species (Hassouni *et al.*, 1999; Miguel *et al.*, 2000; Colburn-Clifford *et al.*, 2010). These findings show that defense against plant-derived toxic compounds plays a key role for bacterial survival in host cells.

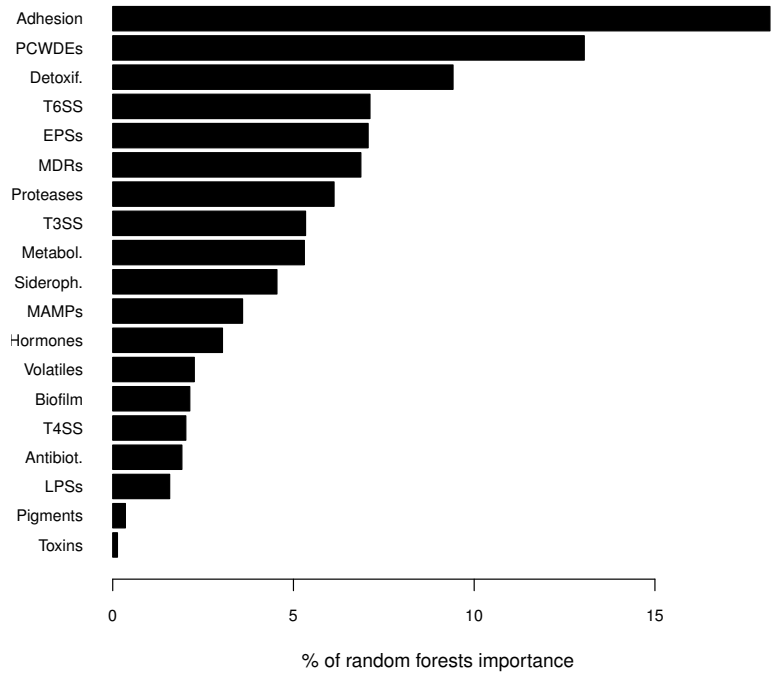


Figure 4. Ranking of predictive features based on random forests importances. Random forests importances (Gini index) were normalized dividing individual values by the sum of all importances. Horizontal axis correspond to the percentage of such normalized values.

Regarding features with low importance measures, it is worth mentioning that phytotoxins are assigned the smallest value. Phytotoxins are products of the host-pathogen interaction that directly injure plant cells, but they are not required for pathogenicity .

Moreover, synthesis of phytotoxins is not a common trait, and only some bacteria such as *P. syringae* are well-known to produce them. Another significant factor that is assigned a low importance is the T4SS. Note that we have only considered in the analysis T4SS involved in effector translocation. Although the T4SS is an important machinery in plant-microbe interactions (e.g. *Agrobacterium*), it is not a common feature of plant-associated bacteria. Nevertheless, we must be cautious when interpreting importance measures, since it is likely that multiple sets of lowly predictive factors are in the end jointly predictive (Chen, Ishwaran, 2012).

The application of the PA-NPA classifier to sequenced bacterial genomes reveals potential associations between human-pathogenic bacteria and plants

The application of our PA-NPA classifier offers the possibility of systematically determining whether a given sequenced bacterial strain could potentially associate with plants. With the increasing number of foodborne poisoning outbreaks that have been linked to fresh produce (Holden *et al.*, 2009; Lim *et al.*, 2014), prediction of potential interactions between human pathogens and vegetables can be very useful for clinical and industrial purposes. Therefore, we went ahead by screening with our model the whole set of complete and draft bacterial genomes available as of 22 December 2014 in the NCBI Genome database (*http://www.ncbi.nlm.nih.gov/genome*). A total of 9,446 strains were analyzed (see Material and Methods), obtaining their corresponding outcomes (plant-associated/non-plant-associated) together with a probability (0-1; Table S3). All assigned probabilities separated by genus were deposited in File S2. Figure 5 shows the distribution of probabilities assigned to all strains. We observe that most of them are given quite low plant-associated probabilities, 50% of them being lower than 0.15 (median) and 75% lower than 0.47 (3rd quartile). Random forests' default threshold is 0.5 when labeling a strain as PA; however, for purposes of discussion, and aiming to maintain a conservative approach, we will not discriminate basing on random forests' labels and will just focus on assigned probabilities.

All in all, paradigmatic pathogenic and non-pathogenic plant-associated bateria are classified quite accurately, with associated probabilities ranging from 0.9 to 1. That is the case of species such as *Agrobacterium fabrum*, *Agrobacterium radiobacter*, *P. syringae*, *Pseudomonas fluorescens*, *Burkholderia cepacia*, *Xanthomonas campestris*, *R. leguminosarum*, *Dickeya dadantii*, *Xylella fastidiosa*, *Bacillus amyloliquefaciens* and *Bacillus subtilis*, among others. Since strains belonging to these species have been used to train the classifier, it seems reasonable that plant-associated bacterial variations of them are identified as such. It

is worth mentioning the resulted outcomes relative to the genus *Pseudomonas*. *Pseudomonas* are well-known for their ability to occupy a wide spectrum of environments as free-living soil organisms or as prominent opportunistic pathogen (Auling, 2001). Among the 221 analyzed *Pseudomonas* strains (Figure 6), 200 (91%) are assigned plant-associated probabilities above 0.85, and only 10 (5%) below 0.7. As a case example, the 48 *P. aeruginosa* isolates are given values ranging from 0.93 to 1. This bacterium is a well documented opportunistic human pathogen, but is also capable of associate and even cause infections to plants (He *et al.*, 2004; Walker *et al.*, 2004).The lowest probability (0.06) among the *Pseudomonas* genus is assigned to the strain *Pseudomonas thermotolerans* DSM 14292, which was isolated from a industrial cooking water and is able to growth at 55º C (Manaia, Moore, 2002).

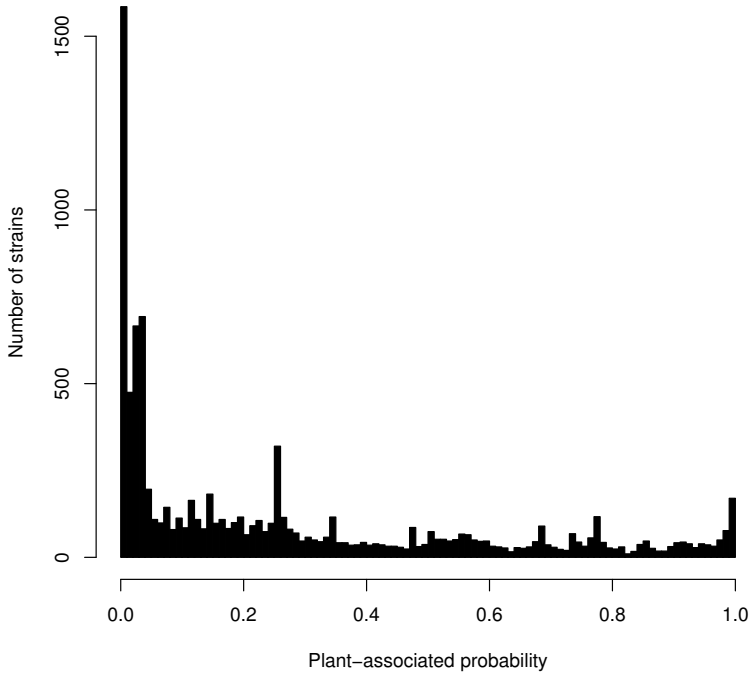


Figure 5. Distribution of probabilities assigned by the PA-NPA classifier to all sequenced bacterial strains in NCBI.

It is also worth to note the outcomes corresponding to the genus *Bacillus*. *Bacillus* comprises a number of species and exhibits a wide range of physiologic abilities, which make them being present in any natural environment (Todar, 2005). This feature is somehow observed when analyzing *Bacillus*' plant-associated probabilities (Figure 6). The genus is divided into two: one cluster containing strains being assigned considerable high PA-probabilities (ranging from 0.78 to 1), such as *B. subtilis*, *B. amyloliquefaciens* and *Bacillus*

licheniformis, and other cluster harbouring strains with relatively low plant-associated probabilities (0.2-0.5), such as *Bacillus thuringiensis*, *Bacillus cereus* and *Bacillus anthracis*.

Human pathogens deserve special attention. We observe that several well-known bacteria such as *Staphylococcus*, *Streptococcus*, *Helicobacter* or *Legionella* are sharply assigned low PA-probabilities, which generally range from 0 to 0.25 (Figure 6; Table S3). In contrast, other human pathogens present significantly higher values, specially strains belonging to Enterobacteriaceae. Such is the case of *E. coli* strain O157:H7, a fatal enterohemorrhagic bacterium whose infection has been traditionally associated to bovine food (Ayscue *et al.*, 2009). Back in 1996, this pathogen caused an important outbreak of foodborne disease, with thousands of cases reported, apparently because of the ingestion of contaminated radish sprout salad (Watanabe *et al.*, 1999). Interestingly, all the 32 *E. coli* O157:H7 strains analyzed were assigned significant plant-associated probabilities, ranging from 0.75 to 0.82. Similarly, another serotype of *E. coli* (strain O104:H4) that recently provoked a large epidemic of the hemolytic-uremic syndrome in Germany (Stockman, 2013) seems to present relatively large values. The total of 49 sequenced isolates of *E. coli* O104:H4 show an average probability of 0.69. In general, the 756 bacteria of the Escherichia genus show a spread distribution of their assigned probabilities (Figure 6), with two prominent clusters with values ranging from 0.5 to 0.6 and from 0.7 to 0.9, respectively. This is in agreement with an increasing opinion that plants are secondary reservoirs for commensal and pathogenic *E. coli* strains (Méric *et al.*, 2013).

Other well-known pathogen that has been suggested not to be exclusively devoted to human hosts is *Salmonella*. *Salmonella* is the causal agent of several worldwide diseases such as gastroenteritis or typhoid fever, provoking around 1.4 million human illnesses annually only in the United States (Mead *et al.*, 1999). A number of studies have shown that this bacterium is able to associate with plants, colonize the phyllosphere and even infect and cause death to plant organs (Kirzinger *et al.*, 2011; Schikora *et al.*, 2012). Like *E. coli*, many cases have been recently reported linking food poisoning with the ingestion of *Salmonella*-contaminated raw fruits and vegetables. It is not surprising, then, that around half of the 325 analyzed *Salmonella* strains show plant-associated probabilities above 0.7 (Figure 6).

Although strains of the *Vibrio* genus are mostly assigned low probabilities (around 0.3), members of the species *Vibrio vulnificus*, *Vibrio harveyi*, *Vibrio parahaemolyticus* or *Vibrio nigripulchritudo* generally show quite high values. Interestingly, *V. parahaemolyticus* has been described as associated with plant roots as well as to fix nitrogen, suggesting this organism could use the rhizosphere as a refugium until an eventual transfer to the overlying waters, potentially contributing to infectious outbreaks (Criminger *et al.*, 2007).

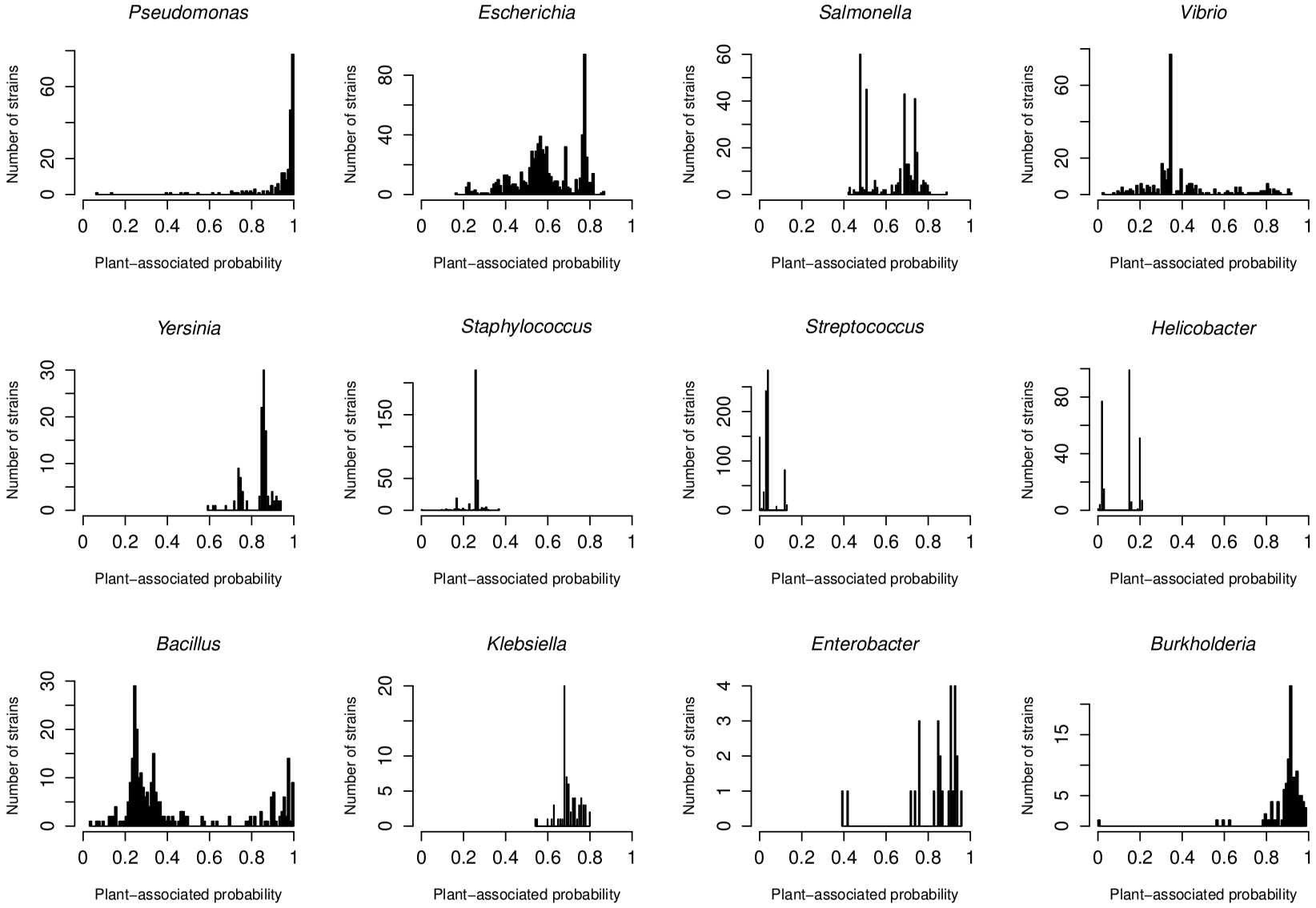


Figure 6. Distribution of probabilities assigned by the PA-NPA classifier to a selection of bacterial genera.

All 117 *Yersinia* strains analyzed exhibited remarkably high plant-associated probabilities (averaging 0.84; Figure 6). Some members of *Yersinia* are well-known human pathogens, such as *Yersinia pestis*, *Yersinia pseudotuberculosis* and *Yersinia enterocolitica*. Notably, both *Y. enterocolitica* and *Y. pseudotuberculosis* have been isolated from soil, water and plants (Massa *et al.*, 1988; Merhej *et al.*, 2008). A recent study has hypothesized that *Y. enterocolitica* could use PCWDEs to guarantee a source of energy from plant cells, allowing the survival of this organism as a soft-rotting bacteria (Tils von *et al.*, 2012).

Other bacteria that have been mostly assigned high PA-probabilities are strains of the genus *Klebsiella*. This is consistent with several studies showing that *Klebsiella* isolates are able to fix nitrogen and live in association with plants (Iniguez *et al.*, 2004; Rosenblueth *et al.*, 2004). Similarly, strains of the genus *Serratia*, a human pathogen typically found in the respiratory and urinary tracts, have been also found to be plant-pathogenic (Kirzinger *et al.*, 2011). Among the 20 analyzed *Serratia* isolates, 18 are assigned probability values above 0.78 (Table S3), with the only exceptions of the two endosymbionts *S. symbiotica* SCc and *S. symbiotica* Tucson (Lamelas *et al.*, 2011). *Enterobacter cloacae*, another bacterium showing cross-kingdom pathogenesis (Kirzinger *et al.*, 2011), is also mostly assigned quite high PA-probabilities by our classifier (8 out of 9 strains).

We also draw attention to *Cronobacter*, which comprises a diverse group of opportunistic food-borne pathogens that are responsible for newborn and infant cases of meningitis, necrotizing enterocolitis, bacteraemia or sepsis (Jaradat *et al.*, 2014). In the last few years, evidence has emerged that plants can be used as natural habitats by members of this genus, which are able to colonize tomato and maize roots both epiphytic and endophytically (Schmid *et al.*, 2009). In line with these findings, all of the six *Cronobacter* strains screened with our classifier are assigned relatively high PA-probabilities (∼0.6; Table S3). Interestingly, the strain that is given the highest value (0.89) corresponds to *Cronobacter turicensis* LMG 23827, a particularly pernicious pathogen that caused the deaths of two neonates in a children's hospital in Zürich in 2005 (Stephan *et al.*, 2011).

With some exceptions, we observe that a wide range of enterobacteria show high plant-associated probability values. Most of these bacteria are frequent human pathogens, although, as pointed out here, many of them have been found to develop some kind of association with plants. A growing interest is arising around the mechanisms for the establishment and interactions of enteric pathogens with vegetable and fruit plants, as countless foodborne illnesses have been caused by these bacteria in the last decades (Holden *et al.*, 2009; Lim *et al.*, 2014; Martínez-Vaz *et al.*, 2014). It is likely, then, that the proposed classifier is identifying associations between some of these strains and plant hosts, which

would be used as alternative reservoirs until new routes back to humans are provided. Therefore, the application of our model to sequenced bacterial isolates presumably involved in food poisoning may serve as a first screening to decipher the cause of the contamination.

Conclusions

Advances in genome sequencing technologies has led to an overwhelming amount of genomic information that is nowadays available to the scientific community. In this work, we took advantage of such a treasure of data to address the problem of bacterial genome classification. We exploit a bioinformatics strategy based on supervised machine-learning that allows us to predict bacterial associations with plants with a precision rate over 94%. The application of the classsifier to the whole set of sequenced bacterial strains in NCBI reveals potential interactions between well-known human pathogenic bacteria and plants. This is the first method for a systematic identification of plant-associated bacteria using genomic information alone, and we believe it can be of great help to researchers in help further their understanding of plant pathology and food safety.

Material and Methods

PIFAR implementation

PIFAR runs on a Linux platform with an Apache web server. The web interface was implemented using HTML and CSS, and data pipelines were developed using PHP, Perl, R and shell scripts. DNA and protein sequences are processed by means of Bioconductor (Gentleman *et al.*, 2004) and BioPerl (Stajich *et al.*, 2002). Dynamic pie charts are generated with the JavaScript library Raphaël (*http://raphaeljs.com/*), and gene maps are produced using the R package genoPlotR (Guy *et al.*, 2010). Open reading frame predictions are generated with Glimmer v3.02 (Delcher *et al.*, 2007).

PIFAR annotation tool

PIFAR performs BLASTp (Altschul, 2005) and HMMER3 (Eddy, 2011) searches of the protein sequences and PFAM profiles (Finn *et al.*, 2014) belonging to the database. To determine whether matches are significant, PIFAR uses threshold E-values for both BLASTp and HMMER3 searches. By default, they are set to $1e^{-15}$ and $1e^{-10}$, respectively. We chose these values after manually screening predicted genes on a number of well-known plant-associated strains, since they offer a trade-off between false positives and false negatives. Anyhow, users are allowed to modify these parameters. Note that the maths behind these E-values depends on the program. To know more about how E-values are defined, we refer the reader to Altschul (2005) and Eddy (2011).

Besides bacterial factors belonging to our database, PIFAR annotation tool searches for type III effectors basing on three other platforms, namely, The *Xanthomonas* Resource (*http://www.xanthomonas.org/*), Ralsto T3E Peeters *et al.* (2013) and the *Pseudomonas syringae* Genome Resources (Lindeberg, 2011).

Genomes and plant-associated classes

Complete and draft bacterial genome sequences available as of 22 December 2014 in the NCBI Genome database (*http://www.ncbi.nlm.nih.gov/genome*) were considered. From such a set, we selected as plant pathogenic (PP) those bacteria that have been reported to cause disease in any plant host. As plant-associated-non-pathogenic (PANP), we selected plant-associated strains reported to be non virulent, including those involved in plant growth-promotion, biological control or simply not found to provoke plant disease. Of these two plant-associated classes, we manually chose as many strains as were available in the initial set. For the selection of non-plant-associated classes (NPA), we based on

the NCBI genome summary tables (lproks). Such tables hold specific information of all sequenced bacterial strains available in NCBI until March 15, 2012. Among other features, information on whether a strain is pathogenic to a specific host is included. In a first step, we excluded strains annotated as able to infect plants and strains lacking annotation on pathogenicity. Then, we manually selected a set of strains by trying to incorporate a diverse and representative ensemble of species. This set was then examined and curated in order to avoid plant-associated strains. The final training set comprised 420 organisms, among which 109 were annotated as PP, 93 as PANP and 218 as NPA. Genome sequences of the biocontrol agents *Pseudomonas putida* PICP2 (Prieto *et al.*, 2011) and *P. fluorescens* PICF7 (Martínez-García *et al.*, 2015a) and the plant pathogenic strains *Pseudomonas savastanoi* NCPPB 3335 (Rodríguez-Palenzuela *et al.*, 2010) and *P. syringae* UMAF0158 (Arrebola *et al.*, 2012) were obtained from our own sequencing projects. Table S1 contains the complete list of strains and their corresponding plant-associated classes.

Our PA-NPA classifier was applied to the whole set of bacterial genome sequences available as of 22 December 2014 in the NCBI Genome database. Entries corresponding to 9,755 strains were first considered (2,785 complete and 6,970 draft). A filtering was performed on this dataset that removed empty entries (folders harboring no sequences) and entries composed by DNA sequences with sizes below 1Mb. The latter aimed to avoid genomes that are only partially sequenced. The final dataset comprised 9,446 sequenced strains (Table S3).

Training set assembly

Full amino acid sequences of the 420 strains were searched using BLASTp (Altschul, 2005) and HMMER3 (Eddy, 2011) against PIFAR database. Based on presence/absence of factors in the database, a vector was generated for each genome so that each position contained the count of identified factors by type. The last three positions, corresponding to secretion systems clusters of type III, IV and VI, were obtained using T346Hunter (Martínez-García *et al.*, 2015b). T4SS involved in conjugation and DNA uptake/release were excluded. The resulting 420x19 matrix was stored in Table S2.

Measures of Performance

The following performance statistics were considered when evaluating supervised machine learning classifiers: accuracy, error rate, precision, sensitivity and F-score. For binary classifiers, we also used ROC (Receiver Operating Characteristic) curves. A systematic analysis of these and other performance measures commonly used in machine

learning classification can be found in (Sokolova, Lapalme, 2009).

Feature selection

We tested whether there are features in the vectors that are uninformative and can be removed. We obtained random forests (Breiman, 2001) importance measures and ranked all features in the vectors. We kept the whole set of features and run 100 random forest classifiers. The classifier with the best precision rate was recorded. This process was repeated by decreasing the number of features based on their rank. In the first step, the feature with lowest importance measure was removed. In the second step, the two features with lowest importance measures were removed, and so on. After the whole process, all classifiers were compared and the one with the best precision rate was selected. This was repeated for the three types of classifiers. In the three cases, the best classifiers were those based on the whole set of features, and consequently, no feature was removed.

Machine Learning: algorithms, configuration and assessment

Cross-validation analyses were used to assess the performance of the trained classifiers. We carried out several methods in order to cover different supervised machine learning algorithms, such as decision trees, bayesian or lazy learners. Six classification algorithms were evaluated, namely, bayesian networks, simple logistic regression, J48, naive bayes, IBk and random forests. Bayesian network and simple logistic regression were used as base classifiers, while J48, naive bayes and IBk were performed as ensemble classifiers. Random forests is a combination of bagging and a random selection of features (Breiman, 2001). These algorithms are reviewed in detail in Larrañaga *et al.* (2006). Each supervised classifier was run 100 times using 10-fold cross-validation, each iteration with a randomly generated seed. In the specific case of random forests, default parameters were used: the number of variables randomly sampled as candidates at each split was set as the square root of the number of features ($\sqrt{19} \sim 4$), the number of total trees in each iteration as 500 and the percentage of OOB (out-of-bag) data used in each tree as 37% of the training data. Performance statistics for the best classifiers are listed in Table 4. ROC curves of binary classifiers are shown in Figure 7. Basing on the precision rate, random forests was chosen as the best strategy.

Training and evaluation of the classifiers were performed using the R packages RWeka (Hall *et al.*, 2009; Hornik *et al.*, 2009) and randomForest (Liaw, Wiener, 2002). ROC curves were produced using the R package ROCR (Sing *et al.*, 2005). For the ranking of random forests' importances, the Gini index was used (Chen, Ishwaran, 2012).

Table 4. Classification performance of each learned classifier.

Classifiers	Precision	Accuracy	Error rate	Sensitivity	F-score
PP-PANP-NPA					
Bayesian Network	0.78	0.86	0.14	0.77	0.77
Simple logistic regression	0.82	0.89	0.11	0.81	0.81
Bagging (J48)	0.81	0.88	0.12	0.78	0.80
Bagging (Naive bayes)	0.78	0.87	0.13	0.78	0.78
Bagging (IBk)	0.81	0.88	0.12	0.77	0.79
Random forests	0.90	0.93	0.07	0.87	0.88
PP-NPP					
Bayesian Network	0.79	0.84	0.16	0.84	0.82
Simple logistic regression	0.87	0.90	0.10	0.86	0.87
Bagging (J48)	0.88	0.91	0.09	0.87	0.88
Bagging (Naive bayes)	0.83	0.88	0.12	0.86	0.85
Bagging (IBk)	0.88	0.89	0.11	0.82	0.85
Random forests	0.94	0.95	0.05	0.92	0.93
PP-PANP-NPA					
Bayesian Network	0.89	0.89	0.11	0.89	0.89
Simple logistic regression	0.89	0.89	0.11	0.89	0.89
Bagging (J48)	0.91	0.91	0.10	0.91	0.91
Bagging (Naive bayes)	0.90	0.90	0.10	0.90	0.90
Bagging (IBk)	0.90	0.89	0.11	0.89	0.89
Random forests	0.94	0.94	0.06	0.94	0.94

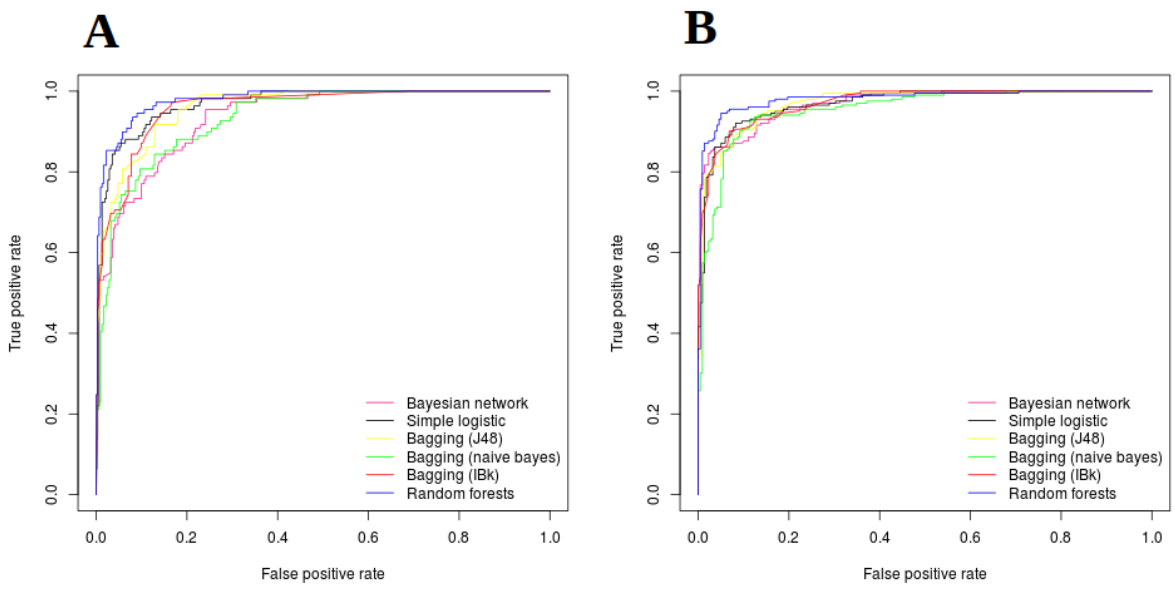


Figure 7. Comparison of ROC curves for the machine-learning schemes used in this study.

Supplementary Material

Table S1. Bacterial strains used to train the classifiers. Available:
http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter4/Table_S1.xlsx

Table S2. Training data (420x19 matrix). Available:
http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter4/Table_S2.xlsx

Table S3. Plant-associated probabilities assigned by the PA-NPA classifier to the whole set of sequenced bacterial genomes from NCBI. Available:
http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter4/Table_S3.xlsx

File S1. PIFAR User Manual. Available:
http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter4/userManual.pdf

File S2. Tables containing plant-associated probabilities assigned by the PA-NPA classifier separated by genus (compressed folder). Available:
http://bacterial-virulence-factors.cbgp.upm.es/supplementary/chapter4/File S2.zip

CONCLUSIONES

1. El conjunto de secuencias codificantes predichas en el genoma de *Pseudomonas syringae* pv. syringae (Pss) UMAF0158 presenta una homología del 91% respecto al del genoma de la cepa filogenétcamente más cercana de la que se conoce su secuencia completa, Pss B728a, patógena de judía.

2. Pss UMAF0158 alberga en su genoma determinantes bacterianos diferenciales respecto a Pss B728a que podrían explicar su capacidad para infectar mango, como el operón *mbo* de síntesis de mangotoxina, un *cluster* potencialmente implicado en la producción de celulosa, dos sistemas de secreción diferentes de tipos III y VI, así como un repertorio particular de efectores del sistema de secreción de tipo III.

3. El genoma de Pss B728a contiene regiones no presentes en el genoma de Pss UMAF0158, la mayoría de ellas correspondientes a elementos genéticos móviles.

4. La anotación del genoma de *Pseudomonas fluorescens* PICF7 reveló genes potenciales que podrían explicar su asociación con olivo, como los codificantes de sideróforos, enzimas detoxificadoras, compuestos volátiles, así como de los sistemas de secreción de tipos III y VI.

5. T346Hunter es una herramienta web, pública y de fácil manejo, que permite identificar regiones genómicas relacionadas con la síntesis de los sistemas de secreción de tipos III, IV y VI.

6. PIFAR (*Plant-bacteria Interaction FActors Resource*) es una aplicación web de acceso público que permite consultar y descargar una base de datos curada de factores de interacción planta-bacteria, así como la anotación de los mismos en genomas bacterianos.

7. La combinación de T346Hunter, PIFAR y el método de aprendizaje automático *random forests* ha permitido generar un clasificador de genomas bacterianos asociados a plantas.

8. La aplicación del clasificador descrito en la conclusión 7 sobre aproximadamente 9.500 genomas bacterianos ha revelado posibles asociaciones con plantas de patógenos humanos, particularmente enterobacterias.

BIBLIOGRAFÍA

Bibliografía

Abby, S. S. and Rocha, E. P. C. The Non-Flagellar Type III Secretion System Evolved from the Bacterial Flagellum and Diversified into Host-Cell Adapted Systems // PLoS Genetics. 2012. 8, 9.

Adaikkalam, V. and Swarup, S. Characterization of *copABCD* operon from a copper-sensitive *Pseudomonas putida* strain. // Canadian journal of microbiology. 2005. 51, 3. 209–216.

Agrios, G. Plant Pathology 5th Edition // San Diego: Academic Press. 2005. p. 922.

Aguilera, S., López-López, K., Nieto, Y., Garcidueñas Piña, R., Hernández-Guzmán, G., *et al.* Functional characterization of the gene cluster from *Pseudomonas syringae* pv. phaseolicola NPS3121 involved in synthesis of phaseolotoxin // Journal of Bacteriology. 2007. 189, 7. 2834–2843.

Ahmad, F., Ahmad, I., and Khan, M. S. Screening of free-living rhizospheric bacteria for their multiple plant growth promoting activities // Microbiological Research. 2008. 163, 2. 173–181.

Ahmed, E. and Holmström, S. J. M. Siderophores in environmental research: Roles and applications // Microbial Biotechnology. 2014. 7, 3. 196–208.

Almeida, N. F., Yan, S., Cai, R., Clarke, C. R., Morris, C. E., *et al.* PAMDB, a multilocus sequence typing and analysis database and website for plant-associated microbes. // Phytopathology. 2010. 100, 3. 208–215.

Altschul, S. F. BLAST Algorithm // Encyclopedia of Life Sciences. 2005. 1–4.

Andreatta, M., Nielsen, M., Moller Aarestrup, F., and Lund, O. *In silico* prediction of human pathogenicity in the gamma-proteobacteria // PLoS One. 2010. 5, 10. e13680.

Angiuoli, S. V., Gussman, A., Klimke, W., Cochrane, G., Field, D., *et al.* Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. // Omics : a journal of integrative biology. 2008. 12, 2. 137–141.

Arrebola, E., Carrión, V. J., Cazorla, F. M., Pérez-García, A., Murillo, J., *et al.* Characterisation of the *mgo* operon in *Pseudomonas syringae* pv. syringae UMAF0158 that is required for mangotoxin production // BMC Microbiology. 2012. 12, 1. 10.

Arrebola, E., Cazorla, F. M., Durán, V. E., Rivera, E., Olea, F., *et al.* Mangotoxin: A novel antimetabolite toxin produced by *Pseudomonas syringae* inhibiting ornithine/arginine biosynthesis // Physiological and Molecular Plant Pathology. 2003. 63, 3. 117–127.

Arrebola, E., Cazorla, F. M., Romero, D., Pérez-García, A., and de Vicente, A. A nonribosomal peptide synthetase gene (*mgoA*) of *Pseudomonas syringae* pv. syringae is involved in mangotoxin biosynthesis and is required for full virulence. // Molecular plant-microbe interactions : MPMI. 2007. 20, 5. 500–509.

Auling, G. Pseudomonads // Biotechnology Set. 2001. 401–431.

Ausubel, F. M. Are innate immune signaling pathways in plants and animals conserved? // Nature immunology. 2005. 6, 10. 973–979.

Ayscue, P., Lanzas, C., Ivanek, R., and Gröhn, Y. T. Modeling on-farm *Escherichia coli* O157:H7 population dynamics. // Foodborne pathogens and disease. 2009. 6, 4. 461–470.

Bally, I. S. E. *Mangifera indica* (mango), ver. 3.1 // Species Profiles for Pacific Island Agroforestry. 2006. 25.

Baltrus, D. A., Nishimura, M. T., Romanchuk, A., Chang, J. H., Mukhtar, M. S., *et al.* Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates // PLoS Pathogens. 2011. 7, 7.

Barbosa, E., Rottger, R., Hauschild, A.-C., Azevedo, V., and Baumbach, J. On the limits of computational functional genomics for bacterial lifestyle prediction // Briefings in Functional Genomics. 2014. 13. 398–408.

Baron, C. Antivirulence drugs to target bacterial secretion systems // Current Opinion in Microbiology. 2010. 13, 1. 100–105.

Barras, F., van Gijsegem, F., and Chatterjee, A. K. Extracellular Enzymes and Pathogenesis of Soft-Rot *Erwinia* // Annual Review of Phytopathology. 1994. 32, 1. 201–234.

Barret, M., Egan, F., Fargier, E., Morrissey, J. P., and O'Gara, F. Genomic analysis of the type VI secretion systems in *Pseudomonas* spp.: Novel clusters and putative effectors uncovered // Microbiology. 2011. 157, 6. 1726–1739.

Bartoli, C., Lamichhane, J. R., Berge, O., Guilbaud, C., Varvaro, L., *et al.* A framework to gage the epidemic potential of plant pathogens in environmental reservoirs: the example of kiwifruit canker // Molecular Plant Pathology. 2014.

Bender, C. L., Alarcón-Chaidez, F., and Gross, D. C. *Pseudomonas syringae* phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. // Microbiology and molecular biology reviews : MMBR. 1999. 63, 2. 266–292.

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., *et al.* GenBank // Nucleic Acids Research. 2014. 42, D1.

Berge, O., Monteil, C. L., Bartoli, C., Chandeysson, C., Guilbaud, C., *et al.* A User's Guide to a Data Base of the Diversity of *Pseudomonas syringae* and Its Application to Classifying Strains in This Phylogenetic Complex // PLoS ONE. 2014. 9, 9. e105547.

Bertani, G. Studies on lysogenesis. I. The mode of phage liberation by lysogenic *Escherichia coli*. // Journal of bacteriology. 1951. 62, 3. 293–300.

Berti, A. D. and Thomas, M. G. Analysis of achromobactin biosynthesis by *Pseudomonas syringae* pv. syringae B728a // Journal of Bacteriology. 2009. 191, 14. 4594–4604.

Bi, D., Liu, L., Tai, C., Deng, Z., Rajakumar, K., *et al.* SecReT4: A web-based bacterial type IV secretion system resource // Nucleic Acids Research. 2013. 41, D1.

Bignell, D., Francis, I., and Loria, R. Thaxtomin A production and virulence are controlled by several bld gene global regulators in *Streptomyces scabies* // Molecular Plant Microbe Interactions. 2014. 27, 8. 875–885.

Bingle, L. E., Bailey, C. M., and Pallen, M. J. Type VI secretion: a beginner's guide // Current Opinion in Microbiology. 2008. 11, 1. 3–8.

Block, A., Li, G., Fu, Z. Q., and Alfano, J. R. Phytopathogen type III effector weaponry and their plant targets // Current Opinion in Plant Biology. 2008. 11, 4. 396–403.

Bonfante, P. and Anca, I.-A. Plants, mycorrhizal fungi, and bacteria: a network of interactions. // Annual review of microbiology. 63. 2009. 363–383.

Bose, M. and Barber, R. D. Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. // *In silico* biology. 2006. 6, 3. 223–227.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, a. UniProtKB/Swiss-Prot // Methods in Molecular Biology. 2007. 406. 89–112.

Boyer, F., Fichant, G., Berthod, J., Vandenbrouck, Y., and Attree, I. Dissecting the bacterial type VI secretion system by a genome wide *in silico* analysis: what can be learned from available microbial genomic resources? // BMC genomics. 2009. 10. 104.

Breiman, L. Random forests // Machine learning. 2001. 5–32.

Brooks, D. M., Bender, C. L., and Kunkel, B. N. The *Pseudomonas syringae* phytotoxin coronatine promotes virulence by overcoming salicylic acid-dependent defences in *Arabidopsis thaliana* // Molecular Plant Pathology. 2005. 6, 6. 629–639.

Brossier, F. and Mock, M. Toxins of *Bacillus anthracis* // Toxicon. 2001. 39, 11. 1747–1755.

Buell, C. R., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I. T., *et al.* The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. tomato DC3000. // Proceedings of the National Academy of Sciences of the United States of America. 2003. 100, 18. 10181–10186.

Bull, C. T., de Boer, S. H., Denny, T. P., Firrao, G., Fischer-Le Saux, M., *et al.* Comprehensive list of names of plant pathogenic bacteria, 1980-2007 // Journal of Plant Pathology. 2010. 92, 3. 551–592.

Burrill, T. J. Pear blight // Transactions of the Illinois State Horticultural Society. 1878. 114–116.

Burtnick, M. N., Brett, P. J., Harding, S. V., Ngugi, S. A., Ribot, W. J., *et al.* The cluster 1 type VI secretion system is a major virulence determinant in *Burkholderia pseudomallei* // Infection and Immunity. 2011. 79, 4. 1512–1525.

Buttner, D. Protein Export According to Schedule: Architecture, Assembly, and Regulation of Type III Secretion Systems from Plant- and Animal-Pathogenic Bacteria // Microbiology and Molecular Biology Reviews. 2012. 76, 2. 262–310.

Büttner, D. and Bonas, U. Regulation and secretion of *Xanthomonas* virulence factors // FEMS Microbiology Reviews. 2010. 34, 2. 107–133.

Cabiscol, E., Tamarit, J., and Ros, J. Oxidative stress in bacteria and protein damage by reactive oxygen species // International Microbiology. 2000. 3, 1. 3–8.

Carrión, V. J., Arrebola, E., Cazorla, F. M., Murillo, J., and de Vicente, A. The *mbo* operon is specific and essential for biosynthesis of mangotoxin in *Pseudomonas syringae* // PLoS ONE. 2012. 7, 5.

Carrión, V. J., Gutiérrez-Barranquero, J., Arrebola, E., Bardaji, L., Codina, J. C., *et al.* The mangotoxin biosynthetic operon (*mbo*) is specifically distributed within *Pseudomonas syringae* genomospecies 1 and was acquired only once during evolution // Applied and Environmental Microbiology. 2013. 79, 3. 756–767.

Cazorla, F. M., Codina, J. C., Abad, C., Arrebola, E., Torés, J. A., *et al.* 62-kb plasmids harboring *rulAB* homologues confer UV-tolerance and epiphytic fitness to *Pseudomonas syringae* pv. syringae mango isolates // Microbial Ecology. 2008. 56, 2. 283–291.

Cazorla, F. M., Torés, J. a., Olalla, L., Pérez-García, a., Farré, J. M., *et al.* Bacterial Apical Necrosis of Mango in Southern Spain: A Disease Caused by *Pseudomonas syringae* pv. syringae. // Phytopathology. 1998. 88, 7. 614–620.

Cazorla, F. M., Arrebola, E., Sesma, A., Pérez-García, A., Codina, J. C., *et al.* Copper Resistance in *Pseudomonas syringae* Strains Isolated from Mango Is Encoded Mainly by Plasmids. // Phytopathology. 2002. 92, 8. 909–916.

Chang, Y.-j., Land, M., Hauser, L., Chertkov, O., Del Rio, T. G., *et al.* Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium *Ktedonobacter racemifer* type strain (SOSP1-21T) // Standards in Genomic Sciences. 2011. 5, 1. 97–111.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., *et al.* VFDB: A reference database for bacterial virulence factors // Nucleic Acids Research. 2005. 33, DATABASE ISS.

Chen, X. and Ishwaran, H. Random forests for genomic data analysis // Genomics. 2012. 99, 6. 323–329.

Clarke, C. R., Cai, R., Studholme, D. J., Guttman, D. S., and Vinatzer, B. A. *Pseudomonas syringae* strains naturally lacking the classical *P. syringae hrp/hrc* Locus are common leaf colonizers equipped with an atypical type III secretion system. // Molecular plant-microbe interactions : MPMI. 2010. 23, 2. 198–210.

Coburn, B., Sekirov, I., and Finlay, B. B. Type III secretion systems and disease // Clinical Microbiology Reviews. 2007. 20, 4. 535–549.

Colburn-Clifford, J. M., Scherf, J. M., and Allen, C. *Ralstonia solanacearum* Dps contributes to oxidative stress tolerance and to colonization of and virulence on tomato plants // Applied and Environmental Microbiology. 2010. 76, 22. 7392–7399.

Coll, N. S. and Valls, M. Current knowledge on the *Ralstonia solanacearum* type III secretion system // Microbial Biotechnology. 2013. 6, 6. 614–620.

Colquhoun, J. P. A really useful company. // Australian family physician. 1994. 23, 9. 1799–1801.

Costacurta, A. and Vanderleyden, J. Synthesis of phytohormones by plant-associated bacteria. // Critical reviews in microbiology. 1995. 21, 1. 1–18.

Criminger, J. D., Hazen, T. H., Sobecky, P. A., and Lovell, C. R. Nitrogen fixation by *Vibrio parahaemolyticus* and its implications for a new ecological niche // Applied and Environmental Microbiology. 2007. 73, 18. 5959–5961.

D'Cruze, T., Gong, L., Treerat, P., Ramm, G., Boyce, J. D., *et al.* Role for the *Burkholderia pseudomallei* type three secretion system cluster 1 *bpscn* gene in virulence // Infection and Immunity. 2011. 79, 9. 3659–3664.

Davis, B. M., Lawson, E. H., Sandkvist, M., Ali, A., Sozhamannan, S., *et al.* Convergence of the secretory pathways for cholera toxin and the filamentous phage, CTXphi. // Science (New York, N.Y.). 2000. 288, 5464. 333–335.

Deering, A. J., Mauer, L. J., and Pruitt, R. E. Internalization of *E. coli* O157:H7 and *Salmonella* spp. in plants: A review // Food Research International. 2012. 45, 2. 567–575.

Delanty, N. and Goldstein, D. B. Diagnostic Exome Sequencing: A New Paradigm in Neurology // Neuron. 2013. 80, 4. 841–843.

Delcher, A. L., Bratke, K. a., Powers, E. C., and Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with Glimmer // Bioinformatics. 2007. 23, 6. 673–679.

Denny, T. P. Involvement of bacterial polysaccharides in plant pathogenesis. // Annual review of phytopathology. 1995. 33. 173–197.

Deslandes, L. and Rivas, S. Catch me if you can: Bacterial effectors and plant targets // Trends in Plant Science. 2012. 17, 11. 644–655.

Dong, T. G., Ho, B. T., Yoder-Himes, D. R., and Mekalanos, J. J. Identification of T6SS-dependent effector and immunity proteins by Tn-seq in *Vibrio cholerae*. // Proceedings of the National Academy of Sciences of the United States of America. 2013. 110, 7. 2623–8.

Dudler, R. The role of bacterial phytotoxins in inhibiting the eukaryotic proteasome // Trends in Microbiology. 2014. 22, 1. 28–35.

Durbin, R., Eddy, S., Krogh, a., and Mitchison, G. Biological sequence analysis. 1998. 356.

Eddleston, M., Karalliedde, L., Buckley, N., Fernando, R., Hutchinson, G., *et al.* Pesticide poisoning in the developing world - A minimum pesticides list // Lancet. 2002. 360, 9340. 1163–1167.

Eddy, S. R. What is a hidden Markov model? // Nature biotechnology. 2004. 22, 10. 1315–1316.

Eddy, S. R. Accelerated profile HMM searches // PLoS Computational Biology. 2011. 7, 10.

Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput // Nucleic Acids Research. 2004. 32, 5. 1792–1797.

Egan, F., Barret, M., and O'Gara, F. The SPI-1-like Type III secretion system: more roles than you think. // Frontiers in plant science. 2014. 5. 34.

Emanuelsson, O., Brunak, S. r., von Heijne, G., and Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. // Nature protocols. 2007. 2, 4. 953–971.

Erbs, G. and Newman, M. The role of lipopolysaccharides in induction of plant defence // Molecular Plant Pathology. 2003. 4, 5. 421–425.

Estes, D. M., Dow, S. W., Schweizer, H. P., and Torres, A. G. Present and future therapeutic strategies for melioidosis and glanders // Expert review of anti-infective therapy. 2011. 8, 3. 325–338.

Feil, H., Feil, W. S., Chain, P., Larimer, F., DiBartolo, G., *et al.* Comparison of the complete genome sequences of *Pseudomonas syringae* pv. syringae B728a and pv. tomato DC3000. // Proceedings of the National Academy of Sciences of the United States of America. 2005. 102, 31. 11064–11069.

Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap // Evolution. 1985. 783–791.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., *et al.* The minimum information about a genome sequence (MIGS) specification. // Nature biotechnology. 2008. 26, 5. 541–547.

Figueira, R. and Holden, D. W. Functions of the *Salmonella pathogenicity* island 2 (SPI-2) type III secretion system effectors // Microbiology. 2012. 158, 5. 1147–1161.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., *et al.* Pfam: The protein families database // Nucleic Acids Research. 2014. 42, D1.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. a., Kirkness, E. F., *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. // Science (New York, N.Y.). 1995. 269, 5223. 496–512.

Fletcher, J., Leach, J. E., Eversole, K., and Tauxe, R. Human pathogens on plants: designing a multidisciplinary strategy for research. // Phytopathology. 2013. 103, 4. 306–15.

Fouts, D. E., Abramovitch, R. B., Alfano, J. R., Baldo, A. M., Buell, C. R., *et al.* Genomewide identification of *Pseudomonas syringae* pv. tomato DC3000 promoters controlled by the HrpL alternative sigma factor. // Proceedings of the National Academy of Sciences of the United States of America. 2002. 99, 4. 2275–2280.

Fukada, S. Human Health and the Microbiota: Interactions Between Gut Microbes, Hygiene and The Immune System // McGill Science Undergraduate Research Journal. 2014. 56–60.

Ganguly, N. K., Croft, S., Singh, L., Sinha, S., and Balganesh, T. Biomedicine and Biotechnology: Public Health Impact // BioMed Research International. 2014. 2014. 1–2.

García-Olmedo, F., Rodríguez-Palenzuela, P., Molina, A., Alamillo, J. M., López-Solanilla, E., *et al.* Antibiotic activities of peptides, hydrogen peroxide and peroxynitrite in plant defence // FEBS Letters. 2001. 498, 2-3. 219–222.

Gardan, L., Shafik, H., Belouin, S., Broch, R., Grimont, F., *et al.* DNA relatedness among the pathovars of *Pseudomonas syringae* and description of *Pseudomonas tremae* sp. nov. and *Pseudomonas cannabina* sp. nov. (ex Sutic and Dowson 1959). // International journal of systematic bacteriology. 1999. 49 Pt 2. 469–478.

Gazi, A. D., Sarris, P. F., Fadouloglou, V. E., Charova, S. N., Mathioudakis, N., *et al.* Phylogenetic analysis of a gene cluster encoding an additional, rhizobial-like type III secretion system that is narrowly distributed among *Pseudomonas syringae* strains // BMC Microbiology. 2012. 12, 1. 188.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., *et al.* Bioconductor: open software development for computational biology and bioinformatics. // Genome biology. 2004. 5, 10. R80.

Gerlach, R. G. and Hensel, M. Protein secretion systems and adhesins: The molecular armory of Gram-negative pathogens // International Journal of Medical Microbiology. 2007. 297, 6. 401–415.

Gfeller, A., Liechti, R., and Farmer, E. E. *Arabidopsis* jasmonate signaling pathway. // Science signaling. 2010. 3, 109. cm4.

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., *et al.* Metagenomic analysis of the human distal gut microbiome. // Science (New York, N.Y.). 2006. 312, 5778. 1355–1359.

Gnanamanickam, S. S. Plant-Associated Bacteria. 2006.

Gómez-Lama, C., Schilirò, E., Valverde-Corredor, A., and Mercado-Blanco, J. The biocontrol endophytic bacterium *Pseudomonas fluorescens* PICF7 induces systemic defense responses in aerial tissues upon colonization of olive roots // Frontiers in Microbiology. 2014. 5, September. 1–14.

Gophna, U., Ron, E. Z., and Graur, D. Bacterial type III secretion systems are ancient and evolved by multiple horizontal-transfer events // Gene. 2003. 312, 1-2. 151–163.

Gorunescu, F. Data mining: Concepts, models and techniques // Intelligent Systems Reference Library. 2011. 12, Second Edition. 800.

Gottig, N., Garavaglia, B. S., Daurelio, L. D., Valentine, A., Gehring, C., *et al.* *Xanthomonas axonopodis* pv. citri uses a plant natriuretic peptide-like protein to modify host homeostasis. // Proceedings of the National Academy of Sciences of the United States of America. 2008. 105, 47. 18631–18636.

Gourion, B., Berrabah, F., Ratet, P., and Stacey, G. *Rhizobium*-legume symbioses: the crucial role of plant immunity // Trends in Plant Science. 2015. 20, 3. 186–194.

Gouy, M., Guindon, S., and Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. // Molecular Biology and Evolution. 2010. 27, 2. 221–224.

Graham, J. H., Gottwald, T. R., Cubero, J., and Achor, D. S. *Xanthomonas axonopodis* pv. citri: Factors affecting successful eradication of citrus canker // Molecular Plant Pathology. 2004. 5, 1. 1–15.

Green, S., Studholme, D. J., Laue, B. E., Dorati, F., Lovell, H., *et al.* Comparative genome analysis provides insights into the evolution and adaptation of *Pseudomonas syringae* pv. aesculi on *Aesculus hippocastanum* // PLoS ONE. 2010. 5, 4.

Guerinot, M. L. Microbial iron transport. // Annual review of microbiology. 1994. 48. 743–772.

Guglielmini, J., Quintais, L., Garcillán-Barcia, M. P., de la Cruz, F., and Rocha, E. P. C. The repertoire of ice in prokaryotes underscores the unity, diversity, and ubiquity of conjugation // PLoS Genetics. 2011. 7, 8.

Guo, M., Block, A., Bryan, C. D., Becker, D. F., and Alfano, J. R. *Pseudomonas syringae* catalases are collectively required for plant pathogenesis // Journal of Bacteriology. 2012. 194, 18. 5054–5064.

Gutiérrez-Barranquero, J. A., Carrión, V. J., Murillo, J., Arrebola, E., Arnold, D. L., *et al.* A *Pseudomonas syringae* diversity survey reveals a differentiated phylotype of the pathovar syringae associated with the mango host and mangotoxin production. // Phytopathology. 2013a. 103, 11. 1115–29.

Gutiérrez-Barranquero, J. A., de Vicente, A., Carrión, V. J., Sundin, G. W., and Cazorla, F. M. Recruitment and rearrangement of three different genetic determinants into a conjugative plasmid increase copper resistance in *Pseudomonas syringae* // Applied and Environmental Microbiology. 2013b. 79, 3. 1028–1033.

Guy, L., Kultima, J. R., and Andersson, S. G. E. genoPlotR: Comparative gene and genome visualization in R // Bioinformatics. 2010. 26, 18. 2334–2335.

Haag, A. F., Arnold, M. F. F., Myka, K. K., Kerscher, B., Dall'Angelo, S., *et al.* Molecular insights into bacteroid development during *Rhizobium*-legume symbiosis // FEMS Microbiology Reviews. 2013. 37, 3. 364–383.

Hall, M., National, H., Frank, E., Holmes, G., Pfahringer, B., *et al.* The WEKA Data Mining Software : An Update // SIGKDD Explorations. 2009. 11, 1. 10–18.

Hamilton, J. P., Neeno-Eckwall, E. C., Adhikari, B. N., Perna, N. T., Tisserat, N., *et al.* The Comprehensive Phytopathogen Genomics Resource: A web-based resource for data-mining plant pathogen genomes // Database. 2011. 2011.

Hardoim, P. R., van Overbeek, L. S., and van Elsas, J. D. Properties of bacterial endophytes and their proposed role in plant growth // Trends in Microbiology. 2008. 16, 10. 463–471.

Hassouni, M. E., Chambost, J. P., Expert, D., Van Gijsegem, F., and Barras, F. The minimal gene set member *msrA*, encoding peptide methionine sulfoxide reductase, is a virulence determinant of the plant pathogen *Erwinia chrysanthemi*. // Proceedings of the National Academy of Sciences of the United States of America. 1999. 96, 3. 887–892.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. The elements of statistical learning: data mining, inference and prediction // The Mathematical Intelligencer. 2005. 27, 2. 83–85.

Hatayama, K., Kawai, S., Shoun, H., Ueda, Y., and Nakamura, A. *Pseudomonas azotifigens* sp. nov., a novel nitrogen-fixing bacterium isolated from a compost pile // International Journal of Systematic and Evolutionary Microbiology. 2005. 55, 4. 1539–1544.

He, J., Baldini, R. L., Déziel, E., Saucier, M., Zhang, Q., *et al.* The broad host range pathogen *Pseudomonas aeruginosa* strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes. // Proceedings of the National Academy of Sciences of the United States of America. 2004. 101, 8. 2530–2535.

Henkel, J. S., Baldwin, M. R., and Barbieri, J. T. Toxins from bacteria. // EXS. 2010. 100. 1–29.

Holden, M. T. G., Titball, R. W., Peacock, S. J., Cerdeño Tárraga, A. M., Atkins, T., *et al.* Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. // Proceedings of the National Academy of Sciences of the United States of America. 2004. 101, 39. 14240–14245.

Holden, N., Pritchard, L., and Toth, I. Colonization outwith the colon: Plants as an alternative environmental reservoir for human pathogenic enterobacteria: Review article // FEMS Microbiology Reviews. 2009. 33, 4. 689–703.

Hood, R. D., Singh, P., Hsu, F., Güvener, T., Carl, M. A., *et al.* A Type VI Secretion System of *Pseudomonas aeruginosa* Targets a Toxin to Bacteria // Cell Host and Microbe. 2010. 7, 1. 25–37.

Hornik, K., Buchta, C., and Zeileis, A. Open-source machine learning: R meets Weka // Computational Statistics. 2009. 24, 2. 225–232.

Houben, E. N. G., Korotkov, K. V., and Bitter, W. Take five - Type VII secretion systems of *Mycobacteria* // Biochimica et Biophysica Acta - Molecular Cell Research. 2014. 1843, 8. 1707–1716.

Iniguez, A. L., Dong, Y., and Triplett, E. W. Nitrogen fixation in wheat provided by *Klebsiella pneumoniae* 342. // Molecular plant-microbe interactions : MPMI. 2004. 17, 10. 1078–1085.

Iraola, G., Vazquez, G., Spangenberg, L., and Naya, H. Reduced set of virulence genes allows high accuracy prediction of bacterial pathogenicity in humans // PLoS ONE. 2012. 7, 8.

Izoré, T., Job, V., and Dessen, A. Biogenesis, regulation, and targeting of the type III secretion system // Structure. 2011. 19, 5. 603–612.

Jander, G., Rahme, L. G., and Ausubel, F. M. Positive correlation between virulence of *Pseudomonas aeruginosa* mutants in mice and insects // Journal of Bacteriology. 2000. 182, 13. 3843–3845.

Jani, A. J. and Cotter, P. A. Type VI Secretion: Not just for pathogenesis anymore // Cell Host and Microbe. 2010. 8, 1. 2–6.

Jaradat, Z. W., Al Mousa, W., Elbetieha, A., Al Nabulsi, A., and Tall, B. D. *Cronobacter* spp. - Opportunistic food-borne pathogens. A review of their virulence and environmental-adaptive traits // Journal of Medical Microbiology. 2014. 63, PART 8. 1023–1037.

Joardar, V., Lindeberg, M., Jackson, R. W., Selengut, J., Dodson, R., *et al.* Whole-genome sequence analysis of *Pseudomonas syringae* pv. phaseolicola 1448A reveals divergence among pathovars in genes involved in virulence and transposition // Journal of Bacteriology. 2005. 187, 18. 6488–6498.

Jones, D. T., Taylor, W. R., and Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. // Computer applications in the biosciences : CABIOS. 1992. 8, 3. 275–282.

Juhas, M., Crook, D. W., and Hood, D. W. Type IV secretion systems: Tools of bacterial horizontal gene transfer and virulence // Cellular Microbiology. 2008. 10, 12. 2377–2386.

Kamber, T., Smits, T. H. M., Rezzonico, F., and Duffy, B. Genomics and current genetic understanding of *Erwinia amylovora* and the fire blight antagonist *Pantoea vagans* // Trees - Structure and Function. 2012. 26, 1. 227–238.

Keim, P., Gruendike, J. M., Klevytska, A. M., Schupp, J. M., Challacombe, J., *et al.* The genome and variation of *Bacillus anthracis* // Molecular Aspects of Medicine. 2009. 30, 6. 397–405.

Kennelly, M., Cazorla, F., de Vicente, A., Ramos, C., and Sundin, G. *Pseudomonas syringae* diseases of fruit trees: progress toward understand-ing and control // Plant Disease. 2007. 9. 4–17.

King, E., Ward, M., and Raney, D. Two simple media for the demonstration of pyocyanin and fluorescin. // The Journal of laboratory and clinical medicine. 1954. 44, 2. 301–307.

Kirzinger, M. W. B., Nadarasah, G., and Stavrinides, J. Insights into cross-kingdom plant pathogenic bacteria // Genes. 2011. 2, 4. 980–997.

Kloepper, J. W. Plant Growth-Promoting Rhizobacteria and Plant Growth Under Gnotobiotic Conditions // Phytopathology. 71, 6. 1981. 642.

Koonin, E. V. and Wolf, Y. I. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world // Nucleic Acids Research. 2008. 36, 21. 6688–6719.

Krahn, D., Ottmann, C., and Kaiser, M. The chemistry and biology of syringolins, glidobactins and cepafungins (syrbactins) // Natural Product Reports. 2011. 28, 11. 1854.

Krewulak, K. D. and Vogel, H. J. Structural biology of bacterial iron uptake // Biochimica et Biophysica Acta - Biomembranes. 2008. 1778, 9. 1781–1804.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. // Journal of molecular biology. 2001. 305, 3. 567–580.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., *et al.* Circos: An information aesthetic for comparative genomics // Genome Research. 2009. 19, 9. 1639–1645.

Lacroix, B. and Citovsky, V. The roles of bacterial and host plant factors in *Agrobacterium*-mediated genetic transformation // International Journal of Developmental Biology. 2013. 57, 6-8. 467–481.

Lamelas, A., Gosalbes, M. J., Manzano-Marín, A., Peretó, J., Moya, A., *et al.* *Serratia symbiotica* from the aphid *Cinara cedri*: A missing link from facultative to obligate insect endosymbiont // PLoS Genetics. 2011. 7, 11.

Lamichhane, J. R., Varvaro, L., Parisi, L., Audergon, J.-M., and Morris, C. E. Chapter Four - Disease and Frost Damage of Woody Plants Caused by *Pseudomonas syringae*: Seeing the Forest for the Trees // Advances in Agronomy. 126. 2014. 235–295.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., *et al.* Initial sequencing and analysis of the human genome. // Nature. 2001. 409, 6822. 860–921.

de Lange, O., Schreiber, T., Schandry, N., Radeck, J., Braun, K. H., *et al.* Breaking the DNA-binding code of *Ralstonia solanacearum* TAL effectors provides new possibilities to generate plant resistance genes against bacterial wilt disease // New Phytologist. 2013. 199, 3. 773–786.

Lapidot, A., Romling, U., and Yaron, S. Biofilm formation and the survival of *Salmonella Typhimurium* on parsley // International Journal of Food Microbiology. 2006. 109, 3. 229–233.

Larkin, M. a., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. a., *et al.* Clustal W and Clustal X version 2.0 // Bioinformatics. 2007. 23, 21. 2947–2948.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., *et al.* Machine learning in bioinformatics // Briefings in Bioinformatics. 2006. 7, 1. 86–112.

Lee, J.-H., Choi, Y., Shin, H., Lee, J., and Ryu, S. Complete Genome Sequence of *Cronobacter sakazakii* Temperate Bacteriophage phiES15 // Journal of Virology. 2012. 86, 14. 7713–7714.

Lewis, J. a. and Papavizas, G. Biocontrol of plant diseases: the approach for tomorrow // Crop Protection. 1991. 10, 2. 95–105.

Li, J. and Wang, N. The *gpsX* gene encoding a glycosyltransferase is important for polysaccharide production and required for full virulence in *Xanthomonas citri* subsp. citri // BMC Microbiology. 2012. 12, 1. 31.

Li, P. L., Everhart, D. M., and Farrand, S. K. Genetic and sequence analysis of the pTiC58 *trb* locus, encoding a mating-pair formation system related to members of the type IV secretion family // Journal of Bacteriology. 1998. 180, 23. 6164–6172.

Li, R., Li, Y., Kristiansen, K., and Wang, J. SOAP: Short oligonucleotide alignment program // Bioinformatics. 2008. 24, 5. 713–714.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., *et al.* *De novo* assembly of human genomes with massively parallel short read sequencing // Genome Research. 2010. 20, 2. 265–272.

Liaw, A. and Wiener, M. Classification and Regression by randomForest // R news. 2002. 2. 18–22.

Licandro-Seraut, H., Scornec, H., Pédron, T., Cavin, J.-F., and Sansonetti, P. J. Functional genomics of *Lactobacillus casei* establishment in the gut. // Proceedings of the National Academy of Sciences of the United States of America. 2014. 1–9.

Lim, J., Lee, D. H., and Heu, S. The Interaction of Human Enteric Pathogens with Plants // The Plant Pathology Journal. 2014. 30, 2. 109–116.

Lindeberg, M. Information management of genome enabled data streams for *Pseudomonas syringae* on the *Pseudomonas*-plant interaction (PPI) website // Genes. 2011. 2, 4. 841–852.

Lindeberg, M., Cunnac, S., and Collmer, A. *Pseudomonas syringae* type III effector repertoires: Last words in endless arguments // Trends in Microbiology. 2012. 20, 4. 199–208.

Lindgren, P. B., Peet, R. C., and Panopoulos, N. J. Gene cluster of *Pseudomonas syringae* pv. phaseolicola controls pathogenicity of bean plants and hypersensitivity on nonhost plants // Journal of Bacteriology. 1986. 168, 2. 512–522.

Lindow, S. E. Ecology of *Pseudomonas syringae* relevant to the field use of Ice- deletion mutants constructed *in vitro* for plant frost control // Engineered organisms in the environment: scientific issues. American Society for Microbiology, Washington, DC. 1985.

Liolios, K., Tavernarakis, N., Hugenholtz, P., and Kyrpides, N. C. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. // Nucleic acids research. 2006. 34, Database issue. D332–D334.

van Loon, L. C., Bakker, P. a. H. M., van der Heijdt, W. H. W., Wendehenne, D., and Pugin, A. Early responses of tobacco suspension cells to rhizobacterial elicitors of induced systemic resistance. // Molecular plant-microbe interactions : MPMI. 2008. 21, 12. 1609–1621.

López-Escudero, F. J. and Mercado-Blanco, J. Verticillium wilt of olive: A case study to implement an integrated strategy to control a soil-borne pathogen // Plant and Soil. 2011. 344, 1. 1–50.

López-Madrigal, S., Latorre, A., Porcar, M., Moya, A., and Gil, R. Complete genome sequence of "*Candidatus* Tremblaya princeps"strain PCVAL, an intriguing translational machine below the living-cell status // Journal of Bacteriology. 2011. 193, 19. 5587–5588.

López-Solanilla, E., García-Olmedo, F., and Rodríguez-Palenzuela, P. Inactivation of the *sapA* to *sapF* locus of *Erwinia chrysanthemi* reveals common features in plant and animal bacterial pathogenesis. // The Plant cell. 1998. 10, 6. 917–924.

Loria, R., Kers, J., and Joshi, M. Evolution of plant pathogenicity in *Streptomyces*. // Annual review of phytopathology. 2006. 44. 469–487.

Love, B. C. Comparing supervised and unsupervised category learning. // Psychonomic bulletin & review. 2002. 9, 4. 829–835.

Lugtenberg, B. and Kamilova, F. Plant-growth-promoting rhizobacteria. // Annual review of microbiology. 2009. 63. 541–556.

Luscombe, N. M., Greenbaum, D., and Gerstein, M. What is bioinformatics? A proposed definition and overview of the field // Methods of Information in Medicine. 2001. 40, 4. 346–358.

Macdonald, E. M., Powell, G. K., Regier, D. a., Glass, N. L., Roberto, F., *et al.* Secretion of Zeatin, Ribosylzeatin, and Ribosyl-1" -Methylzeatin by *Pseudomonas savastanoi*: Plasmid-Coded Cytokinin Biosynthesis. // Plant physiology. 1986. 82, 3. 742–747.

Maggiorani Valecillos, A., Rodríguez Palenzuela, P., and López-Solanilla, E. The role of several multidrug resistance systems in *Erwinia chrysanthemi* pathogenesis. // Molecular plant-microbe interactions : MPMI. 2006. 19, 6. 607–613.

Mahajan-Miklos, S., Tan, M. W., Rahme, L. G., and Ausubel, F. M. Molecular mechanisms of bacterial virulence elucidated using a *Pseudomonas aeruginosa-Caenorhabditis elegans* pathogenesis model // Cell. 1999. 96, 1. 47–56.

Maldonado-González, M. M., Prieto, P., Ramos, C., and Mercado-Blanco, J. From the root to the stem: Interaction between the biocontrol root endophyte *Pseudomonas fluorescens* PICF7 and the pathogen *Pseudomonas savastanoi* NCPPB 3335 in olive knots // Microbial Biotechnology. 2013. 6, 3. 275–287.

Manaia, C. M. and Moore, E. R. B. *Pseudomonas thermotolerans* sp. nov., a thermotolerant species of the genus *Pseudomonas sensu stricto* // International Journal of Systematic and Evolutionary Microbiology. 2002. 52, 6. 2203–2209.

Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., *et al.* Top 10 plant pathogenic bacteria in molecular plant pathology // Molecular Plant Pathology. 2012. 13, 6. 614–629.

Marchler-Bauer, a., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., *et al.* CDD: NCBI's conserved domain database // Nucleic Acids Research. 2014. 43, D1. D222–D226.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., *et al.* Genome sequencing in microfabricated high-density picolitre reactors. // Nature. 2005. 437, 7057. 376–380.

Martínez-García, P. M., Ruano-Rosa, D., Schilirò, E., Prieto, P., Ramos, C., *et al.* Complete genome sequence of *Pseudomonas fluorescens* strain PICF7, an indigenous root endophyte from olive (*Olea europaea* L.) and effective biocontrol agent against *Verticillium dahliae* // Standards in Genomic Sciences. 2015a. 10, 1. 10.

Martínez-García, P. M., Ramos, C., and Rodríguez-Palenzuela, P. T346Hunter: A Novel Web-Based Tool for the Prediction of Type III, Type IV and Type VI Secretion Systems in Bacterial Genomes // Plos One. 2015b. 10, 4. e0119317.

Martínez, J. L., Sánchez, M. B., Martínez-Solano, L., Hernandez, A., Garmendia, L., *et al.* Functional role of bacterial multidrug efflux pumps in microbial natural ecosystems // FEMS Microbiology Reviews. 2009. 33, 2. 430–449.

Martínez-Vaz, B. M., Fink, R. C., Diez-Gonzalez, F., and Sadowsky, M. J. Enteric Pathogen-Plant Interactions: Molecular Connections Leading to Colonization and Growth and Implications for Food Safety // Microbes and Environments. 2014. 29, 2. 123–135.

Massa, S., Cesaroni, D., Poda, G., and Trovatelli, L. D. Isolation of *Yersinia enterocolitica* and related species from river water. // Zentralblatt fur Mikrobiologie. 1988. 143, 8. 575–581.

Matas, I. M., Lambertsen, L., Rodríguez-Moreno, L., and Ramos, C. Identification of novel virulence genes and metabolic pathways required for full fitness of *Pseudomonas*

savastanoi pv. savastanoi in olive (*Olea europaea*) knots // New Phytologist. 2012. 196, 4. 1182–1196.

Matas, I. M., Castañeda Ojeda, M. P., Aragón, I. M., Antúnez-Lamas, M., Murillo, J., *et al.* Translocation and functional analysis of *Pseudomonas savastanoi* pv. savastanoi NCPPB 3335 type III secretion system effectors reveals two novel effector families of the *Pseudomonas syringae* complex. // Molecular Plant-Microbe interactions : MPMI. 2014. 27, 5. In press.

Matas, I. M., Pérez-Martínez, I., Quesada, J. M., Rodríguez-Herva, J. J., Penyalver, R., *et al.* *Pseudomonas savastanoi* pv. savastanoi contains two *iaaL* paralogs, one of which exhibits a variable number of a trinucleotide (TAC) tandem repeat // Applied and Environmental Microbiology. 2009. 75, 4. 1030–1035.

Matthysse, A. G., Jaeckel, P., and Jeter, C. *attG* and *attC* mutations of *Agrobacterium tumefaciens* are dominant negative mutations that block attachment and virulence. // Canadian journal of microbiology. 2008. 54, 4. 241–247.

Maxam, a. M. and Gilbert, W. A new method for sequencing DNA. // Proceedings of the National Academy of Sciences of the United States of America. 1977. 74, 2. 560–564.

McNear, D. H. The rhizosphere - Rottes, soil and everything in between // Nature reviews. Microbiology. 2012. 4, 3. 1.

Mead, P. S., Slutsker, L., Dietz, V., McCaig, L. F., Bresee, J. S., *et al.* Food-related illness and death in the United States // Emerging Infectious Diseases. 1999. 5, 5. 607–625.

Mercado-Blanco, J., Rodríguez-Jurado, D., Hervás, A., and Jiménez-Diaz, R. M. Suppression of Verticillium wilt in olive planting stocks by root-associated fluorescent *Pseudomonas* spp. // Biological Control. 2004. 30, 2. 474–486.

Merhej, V., Adékambi, T., Pagnier, I., Raoult, D., and Drancourt, M. *Yersinia massiliensis* sp. nov., isolated from fresh water // International Journal of Systematic and Evolutionary Microbiology. 2008. 58, 4. 779–784.

Méric, G., Kemsley, E. K., Falush, D., Saggers, E. J., and Lucchini, S. Phylogenetic distribution of traits associated with plant colonization in *Escherichia coli* // Environmental Microbiology. 2013. 15, 2. 487–501.

Meyer, J. and Abdallah, M. The fluorescent pigment of *Pseudomonas fluorescens*: biosynthesis, purification and physicochemical properties // J Gen Microbiol. 1978. 107, 2. 319–328.

Mhedbi-Hajri, N., Darrasse, A., Pigné, S., Durand, K., Fouteau, S., *et al.* Sensing and adhesion are adaptive functions in the plant pathogenic xanthomonads. // BMC evolutionary biology. 2011. 11. 67.

Miguel, E., Poza-Carrión, C., López-Solanilla, E., Aguilar, I., Llama-Palacios, A., *et al.* Evidence against a direct antimicrobial role of H2O2 in the infection of plants by *Erwinia chrysanthemi*. // Molecular plant-microbe interactions : MPMI. 2000. 13, 4. 421–429.

Molicotti, P., Bua, A., and Zanetti, S. Cost-effectiveness in the diagnosis of tuberculosis: Choices in developing countries // Journal of Infection in Developing Countries. 2014. 8, 1. 24–38.

Montesinos, E. and Vilardell, P. Relationships among population levels of *Pseudomonas syringae*, amount of ice nuclei, and incidence of blast of dormant flower buds in commercial pear orchards in Catalunya, Spain // Phytopathology. 1991. 81. 113–119.

Moragrega, C., Matias, J., Aletà, N., Montesinos, E., and Rovira, M. Apical Necrosis and Premature Drop of Persian (English) Walnut Fruit Caused by *Xanthomonas arboricola* pv. juglandis // Plant Disease. 2011. 95, 12. 1565–1570.

Morris, C. E., Glaux, C., Latour, X., Gardan, L., Samson, R., *et al.* The Relationship of Host Range, Physiology, and Genotype to Virulence on Cantaloupe in *Pseudomonas syringae* from Cantaloupe Blight Epidemics in France. // Phytopathology. 2000. 90, 6. 636–646.

Mougous, J. D., Cuff, M. E., Raunser, S., Shen, A., Zhou, M., *et al.* A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus. // Science (New York, N.Y.). 2006. 312, 5779. 1526–1530.

Münzinger, M., Budzikiewicz, H., Expert, D., Enard, C., and Meyer, J. M. Achromobactin, a new citrate siderophore of *Erwinia chrysanthemi* // Zeitschrift fur Naturforschung - Section C Journal of Biosciences. 2000. 55, 5-6. 328–332.

Mur, L. A. J., Kenton, P., Lloyd, A. J., Ougham, H., and Prats, E. The hypersensitive response; The centenary is upon us but how much do we know? // Journal of Experimental Botany. 59, 3. 2008. 501–520.

Mutka, A. M., Fawley, S., Tsao, T., and Kunkel, B. N. Auxin promotes susceptibility to *Pseudomonas syringae* via a mechanism independent of suppression of salicylic acid-mediated defenses // Plant Journal. 2013. 74, 5. 746–754.

Myeni, S., Child, R., Ng, T. W., Kupko, J. J., Wehrly, T. D., *et al.* *Brucella* Modulates Secretory Trafficking via Multiple Type IV Secretion Effector Proteins // PLoS Pathogens. 2013. 9, 8.

Nakamura, Y., Cochrane, G., and Karsch-Mizrachi, I. The international nucleotide sequence database collaboration // Nucleic Acids Research. 2013. 41, D1.

Nandi, T., Holden, M. T. G., Didelot, X., Mehershahi, K., Boddey, J. A., *et al.* *Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory and epigenetic profiles // Genome Research. 2014. 141.

Navarro, L., Dunoyer, P., Jay, F., Arnold, B., Dharmasiri, N., *et al.* A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. // Science (New York, N.Y.). 2006. 312, 5772. 436–439.

Notredame, C., Higgins, D. G., and Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. // Journal of molecular biology. 2000. 302, 1. 205–217.

O'Brien, H. E., Thakur, S., and Guttman, D. S. Evolution of Plant Pathogenesis in *Pseudomonas syringae*: A Genomics Perspective. // Annual review of phytopathology. 2011. 49. 269–289.

Ongena, M., Jourdan, E., Adam, A., Paquot, M., Brans, A., *et al.* Surfactin and fengycin lipopeptides of *Bacillus subtilis* as elicitors of induced systemic resistance in plants // Environmental Microbiology. 2007. 9, 4. 1084–1090.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST) // Nucleic Acids Research. 2014. 42, D1.

Pappas, S. Your Body Is a habitat ... for Bacteria // Science Now Daily News. 2009.

Paradis, E., Claude, J., and Strimmer, K. APE: Analyses of phylogenetics and evolution in R language // Bioinformatics. 2004. 20, 2. 289–290.

Pea, J. M. and Vityaev, E. Knowledge discovery in bioinformatics // Intelligent Data Analysis. 2010. 14, 2. 157–158.

Peek, N. and Swift, S. Intelligent Data Analysis for Knowledge Discovery, Patient Monitoring and Quality Assessment // (B1) Methods of Information in Medicine. 2012. 318–322.

van Peer, R., Niemann, G. J., and Schippers, B. Induced Resistance and Phytoalexin Accumulation in Biological Control of Fusarium Wilt of Carnation by *Pseudomonas* sp. Strain WCS417r.PDF // Phytopathology. 1991. 81, 7. 728–734.

Peeters, N., Carrère, S., Anisimova, M., Plener, L., Cazalé, A.-C., *et al.* Repertoire, unified nomenclature and evolution of the Type III effector gene set in the Ralstonia solanacearum species complex. // BMC genomics. 2013. 14, 1. 859.

Pegg, G. and Brady, B. Verticillium Wilts // Physiological and Molecular Plant Pathology. 2003.

Pérez-Martínez, I., Rodríguez-Moreno, L., Lambertsen, L., Matas, I. M., Murillo, J., *et al.* Fate of a *Pseudomonas savastanoi* pv. savastanoi type III secretion system mutant in olive plants (*Olea europaea* L.) // Applied and Environmental Microbiology. 2010. 76, 11. 3611–3619.

Philippot, L., Raaijmakers, J. M., Lemanceau, P., and van der Putten, W. H. Going back to the roots: the microbial ecology of the rhizosphere. // Nature reviews. Microbiology. 2013. 11, 11. 789–99.

Pitzschke, A. *Agrobacterium* infection and plant defense-transformation success hangs by a thread. // Frontiers in plant science. 2013. 4, December. 519.

Pizarro-Cerdá, J. and Cossart, P. Bacterial adhesion and entry into host cells // Cell. 2006. 124, 4. 715–727.

Pradhan, B. B., Ranjan, M., and Chatterjee, S. XadM, a Novel Adhesin of *Xanthomonas oryzae* pv. oryzae, Exhibits Similarity to Rhs Family Proteins and Is Required for Optimum Attachment, Biofilm Formation, and Virulence // Molecular Plant-Microbe Interactions. 2012. 25, 9. 1157–1170.

Prieto, P. and Mercado-Blanco, J. Endophytic colonization of olive roots by the biocontrol strain *Pseudomonas fluorescens* PICF7 // FEMS Microbiology Ecology. 2008. 64, 2. 297–306.

Prieto, P., Navarro-Raya, C., Valverde-Corredor, A., Amyotte, S. G., Dobinson, K. F., *et al.* Colonization process of olive tissues by *Verticillium dahliae* and its *in planta* interaction with the biocontrol root endophyte *Pseudomonas fluorescens* PICF7 // Microbial Biotechnology. 2009. 2, 4. 499–511.

Prieto, P., Schilirò, E., Maldonado-González, M. M., Valderrama, R., Barroso-Albarracín, J. B., *et al.* Root Hairs Play a Key Role in the Endophytic Colonization of Olive Roots by *Pseudomonas* spp. with Biocontrol Activity // Microbial Ecology. 2011. 62, 2. 435–445.

Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy // Nucleic Acids Research. 2012. 40, D1.

Pukatzki, S., Ma, A. T., Revel, A. T., Sturtevant, D., and Mekalanos, J. J. Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin. // Proceedings of the National Academy of Sciences of the United States of America. 2007. 104, 39. 15508–15513.

Pukatzki, S., Ma, A. T., Sturtevant, D., Krastins, B., Sarracino, D., *et al.* Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system. // Proceedings of the National Academy of Sciences of the United States of America. 2006. 103, 5. 1528–1533.

Pundhir, S. and Kumar, A. SSPred: A prediction server based on SVM for the identification and classification of proteins involved in bacterial secretion systems. // Bioinformation. 2011. 6, 10. 380–382.

Quesada, J. M., Penyalver, R., Pérez-Panadés, J., Salcedo, C. I., Carbonell, E. a., *et al.* Comparison of chemical treatments for reducing epiphytic *Pseudomonas savastanoi* pv. savastanoi populations and for improving subsequent control of olive knot disease // Crop Protection. 2010. 29, 12. 1413–1420.

Rahme, L. G., Stevens, E. J., Wolfort, S. F., Shao, J., Tompkins, R. G., *et al.* Common virulence factors for bacterial pathogenicity in plants and animals. // Science (New York, N.Y.). 1995. 268, 5219. 1899–1902.

Ramos, C., Matas, I. M., Bardaji, L., Aragón, I. M., and Murillo, J. *Pseudomonas savastanoi* pv. savastanoi: Some like it knot // Molecular Plant Pathology. 2012. 13, 9. 998–1009.

Ran, L. X., Li, Z. N., Wu, G. J., Van Loon, L. C., and Bakker, P. a. H. M. Induction of systemic resistance against bacterial wilt in *Eucalyptus urophylla* by fluorescent *Pseudomonas* spp // European Journal of Plant Pathology. 2005. 113, 1. 59–70.

Reinhardt, J. A., Baltrus, D. A., Nishimura, M. T., Jeck, W. R., Jones, C. D., *et al.* *De novo assembly* using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. oryzae // Genome Research. 2009. 19, 2. 294–305.

Ren, Y., Strobel, G., Sears, J., and Park, M. *Geobacillus* sp., a thermophilic soil bacterium producing volatile antibiotics // Microbial Ecology. 2010. 60, 1. 130–136.

Rigano, L. A., Siciliano, F., Enrique, R., Sendín, L., Filippone, P., *et al.* Biofilm formation, epiphytic fitness, and canker development in *Xanthomonas axonopodis* pv. citri. // Molecular plant-microbe interactions : MPMI. 2007. 20, 10. 1222–1230.

Río-Álvarez, I., Rodríguez-Herva, J. J., Martínez, P. M., González-Melendi, P., García-Casado, G., *et al.* Light regulates motility, attachment and virulence in the plant pathogen *Pseudomonas syringae* pv tomato DC3000 // Environmental Microbiology. 2013. 16, 7. 2072–2085.

Río-Álvarez, I., Muñoz Gómez, C., Navas-Vásquez, M., García-Martínez, P. M., Antúnez-Lamas, M., *et al.* Role of Dickeya dadantii 3937 chemoreceptors in the entry to Arabidopsis leaves through wounds // Molecular Plant Pathology. 2014. n/a–n/a.

Rodrigues, E. P., Rodrigues, L. S., Oliveira, A. L. M., Baldani, V. L. D., Teixeira, K. R. D. S., *et al.* *Azospirillum amazonense* inoculation: effects on growth, yield and N2 fixation of rice (*Oryza sativa* L.) // Plant and Soil. 2008. 302, 1-2. 249–261.

Rodríguez-Palenzuela, P., Matas, I. M., Murillo, J., López-Solanilla, E., Bardaji, L., *et al.* Annotation and overview of the *Pseudomonas savastanoi* pv. savastanoi NCPPB 3335 draft genome reveals the virulence gene complement of a tumour-inducing pathogen of woody hosts // Environmental Microbiology. 2010. 12, 6. 1604–1620.

Romero, D., De Vicente, a., Olmos, J. L., Dávila, J. C., and Pérez-García, a. Effect of lipopeptides of antagonistic strains of *Bacillus subtilis* on the morphology and ultrastructure of the cucurbit fungal pathogen *Podosphaera fusca* // Journal of Applied Microbiology. 2007. 103, 4. 969–976.

Rosenblueth, M., Martínez, L., Silva, J., and Martínez-Romero, E. *Klebsiella variicola*, a novel species with clinical and plant-associated isolates. // Systematic and applied microbiology. 2004. 27, 1. 27–35.

Rothberg, J. M. and Leamon, J. H. The development and impact of 454 sequencing. // Nature biotechnology. 2008. 26, 10. 1117–1124.

Russell, A. B., Peterson, S. B., and Mougous, J. D. Type VI secretion system effectors: poisons with a purpose. // Nature reviews. Microbiology. 2014. 12, 2. 137–48.

Ryu, C.-M., Farag, M. A., Hu, C.-H., Reddy, M. S., Wei, H.-X., *et al.* Bacterial volatiles promote growth in *Arabidopsis*. // Proceedings of the National Academy of Sciences of the United States of America. 2003. 100, 8. 4927–4932.

Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., *et al.* Genome sequence of the plant pathogen *Ralstonia solanacearum*. // Nature. 2002. 415, 6871. 497–502.

Salto-Tellez, M. and Gonzalez De Castro, D. Next-generation sequencing: A change of paradigm in molecular diagnostic validation // Journal of Pathology. 2014. 234, 1. 5–10.

Sanger, F., Nicklen, S., and Coulson, a. R. DNA sequencing with chain-terminating inhibitors. // Proceedings of the National Academy of Sciences of the United States of America. 1977. 74, 12. 5463–5467.

Santoro, F., Vianna, M. E., and Roberts, A. P. Variation on a theme; an overview of the Tn*916*/Tn*1545* family of mobile genetic elements in the oral and nasopharyngeal streptococci // Frontiers in Microbiology. 2014. 5, October. 1–10.

Santos, M., Diánez, F., de Cara, M., Camacho, F., and Tello, J. El control biológico de plagas y enfermedades. Un encuadre crítico // Cuadernos de Estudios Agroalimentarios. 2010. 1. 61–72.

Sarkar, S. F., Gordon, J. S., Martin, G. B., and Guttman, D. S. Comparative genomics of host-specific virulence in *Pseudomonas syringae* // Genetics. 2006. 174, 2. 1041–1056.

Sarris, P. F., Skandalis, N., Kokkinidis, M., and Panopoulos, N. J. *In silico* analysis reveals multiple putative type VI secretion systems and effector proteins in *Pseudomonas syringae* pathovars // Molecular Plant Pathology. 2010. 11, 6. 795–804.

Saxonov, S., Berg, P., and Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. // Proceedings of the National Academy of Sciences of the United States of America. 2006. 103, 5. 1412–1417.

Schalk, I. J. and Guillon, L. Pyoverdine biosynthesis and secretion in *Pseudomonas aeruginosa*: Implications for metal homeostasis // Environmental Microbiology. 2013. 15, 6. 1661–1673.

Schell, M. A., Ulrich, R. L., Ribot, W. J., Brueggemann, E. E., Hines, H. B., *et al.* Type VI secretion is a major virulence determinant in *Burkholderia mallei* // Molecular Microbiology. 2007. 64, 6. 1466–1485.

Schellenberg, B., Ramel, C., and Dudler, R. *Pseudomonas syringae* virulence factor syringolin A counteracts stomatal immunity by proteasome inhibition. // Molecular plant-microbe interactions : MPMI. 2010. 23, 10. 1287–1293.

Schikora, A., Garcia, A. V., and Hirt, H. Plants as alternative hosts for *Salmonella* // Trends in Plant Science. 2012. 17, 5. 245–249.

Schilirò, E., Ferrara, M., Nigro, F., and Mercado-Blanco, J. Genetic Responses Induced in Olive Roots upon Colonization by the Biocontrol Endophytic Bacterium *Pseudomonas fluorescens* PICF7 // PLoS ONE. 2012. 7, 11.

Schippers, B., Bakker, a. W., Bakker, P. a. H. M., and Van Peer, R. Beneficial and deleterious effects of HCN-producing pseudomonads on rhizosphere interactions // Plant and Soil. 1990. 129, 1. 75–83.

Schmid, M., Iversen, C., Gontia, I., Stephan, R., Hofmann, A., *et al.* Evidence for a plant-associated natural habitat for *Cronobacter* spp. // Research in Microbiology. 2009. 160, 8. 608–614.

Scholz-Schroeder, B. K., Hutchison, M. L., Grgurina, I., and Gross, D. C. The contribution of syringopeptin and syringomycin to virulence of *Pseudomonas syringae* pv. syringae strain B301D on the basis of *sypA* and *syrB1* biosynthesis mutant analysis. // Molecular plant-microbe interactions : MPMI. 2001. 14, 3. 336–348.

Scholze, H. and Boch, J. TAL effectors are remote controls for gene activation // Current Opinion in Microbiology. 2011. 14, 1. 47–53.

Scortichini, M. and Marcelletti, S. Definition of plant pathogenic *Pseudomonas* genomospecies of the *Pseudomonas syringae* complex through a phylogenomic approach. // Phytopathology. 2014.

Shelburne, S. a., Davenport, M. T., Keith, D. B., and Musser, J. M. The role of complex carbohydrate catabolism in the pathogenesis of invasive streptococci // Trends in Microbiology. 2008. 16, 7. 318–325.

Shin, S. C., Kim, S.-H., You, H., Kim, B., Kim, a. C., *et al.* *Drosophila* Microbiome Modulates Host Developmental and Metabolic Homeostasis via Insulin Signaling // Science. 2011. 334, 6056. 670–674.

Shrivastava, S. and Mande, S. S. Identification and functional characterization of gene components of type VI secretion system in bacterial genomes // PLoS ONE. 2008. 3, 8.

Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B., and Jackson, R. W. *Pseudomonas* genomes: Diverse and adaptable // FEMS Microbiology Reviews. 2011. 35, 4. 652–680.

Silverman, J. M., Brunet, Y. R., Cascales, E., and Mougous, J. D. Structure and Regulation of the Type VI Secretion System // Annual Review of Microbiology. 2012. 66, 1. 453–472.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. ROCR: Visualizing classifier performance in R // Bioinformatics. 2005. 21, 20. 3940–3941.

Sokolova, M. and Lapalme, G. A systematic analysis of performance measures for classification tasks // Information Processing & Management. 2009. 45, 4. 427–437.

Souza, R. C., Quispe Saji, G. d. R., Costa, M. O. C., Netto, D. S., Lima, N. C. B., *et al.* AtlasT4SS: A curated database for type IV secretion systems // BMC Microbiology. 2012. 12, 1. 172.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., *et al.* The Bioperl toolkit: Perl modules for the life sciences // Genome Research. 2002. 12, 10. 1611–1618.

Stephan, R., Lehner, A., Tischler, P., and Rattei, T. Complete genome sequence of *Cronobacter turicensis* LMG 23827, a food-borne pathogen causing deaths in neonates // Journal of Bacteriology. 2011. 193, 1. 309–310.

Stevens, M. P., Haque, A., Atkins, T., Hill, J., Wood, M. W., *et al.* Attenuated virulence and protective efficacy of a *Burkholderia pseudomallei bsa* type III secretion mutant in murine models of melioidosis // Microbiology. 2004. 150, 8. 2669–2676.

Stockman, J. German Outbreak of *Escherichia coli* O104:H4 Associated with Sprouts // Yearbook of Pediatrics. 2013. 2013. 287–289.

Studholme, D. J., Ibanez, S. G., MacLean, D., Dangl, J. L., Chang, J. H., *et al.* A draft genome sequence and functional screen reveals the repertoire of type III secreted proteins of *Pseudomonas syringae* pathovar tabaci 11528. // BMC genomics. 2009. 10. 395.

Suarez, G., Sierra, J. C., Erova, T. E., Sha, J., Horneman, A. J., *et al.* A type VI secretion system effector protein, VgrG1, from *Aeromonas hydrophila* that induces host cell toxicity by ADP ribosylation of actin // Journal of Bacteriology. 2010. 192, 1. 155–168.

Sudo, T., Shinohara, K., Dohmae, N., Takio, K., Usami, R., *et al.* Isolation and characterization of the gene encoding an aminopeptidase involved in the selective toxicity of ascamycin toward *Xanthomonas campestris* pv. citri. // The Biochemical journal. 1996. 319 (Pt 1). 99–102.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., *et al.* MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods // Molecular Biology and Evolution. 2011. 28, 10. 2731–2739.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. A genomic perspective on protein families. // Science (New York, N.Y.). 1997. 278, 5338. 631–637.

von Tils, D., Blädel, I., Schmidt, M. A., and Heusipp, G. Type II secretion in *Yersinia*-a secretion system for pathogenicity and environmental fitness. // Frontiers in cellular and infection microbiology. 2012. 2, December. 160.

Todar, K. The Genus *Bacillus* // Todar's online text book of bacteriology. 2005. 20.

Toth, I. K., Bell, K. S., Holeva, M. C., and Birch, P. R. J. Soft rot erwiniae: From genes to genomes // Molecular Plant Pathology. 2003. 4, 1. 17–30.

Tseng, T.-T., Tyler, B. M., and Setubal, J. a. C. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. // BMC microbiology. 2009. 9 Suppl 1. S2.

Uppalapati, S. R., Ayoubi, P., Weng, H., Palmer, D. a., Mitchell, R. E., *et al.* The phytotoxin coronatine and methyl jasmonate impact multiple phytohormone pathways in tomato // Plant Journal. 2005. 42, 2. 201–217.

Urban, M., Pant, R., Raghunath, a., Irvine, a. G., Pedro, H., *et al.* The Pathogen-Host Interactions database (PHI-base): additions and future developments // Nucleic Acids Research. 2014. 43, D1. D645–D655.

Van Loon, L. C. Plant responses to plant growth-promoting rhizobacteria // European Journal of Plant Pathology. 2007. 119, 3. 243–254.

Peñaloza Vázquez, A., Kidambi, S. P., Chakrabarty, A. M., and Bender, C. L. Characterization of the alginate biosynthetic gene cluster in *Pseudomonas syringae* pv. syringae. // Journal of bacteriology. 1997. 179, 14. 4464–4472.

Vergunst, A. C., Schrammeijer, B., den Dulk-Ras, A., de Vlaam, C. M., Regensburg-Tuïnk, T. J., *et al.* VirB/D4-dependent protein translocation from *Agrobacterium* into plant cells. // Science (New York, N.Y.). 2000. 290, 5493. 979–982.

Vernikos, G. S. and Parkhill, J. Interpolated variable order motifs for identification of horizontally acquired DNA: Revisiting the *Salmonella* pathogenicity islands // Bioinformatics. 2006. 22, 18. 2196–2203.

Vidaver, A. and Lambrecht, P. Bacteria as plant pathogens // The Plant Health Instructor. 2004.

Vilches, S., Wilhelms, M., Yu, H. B., Leung, K. Y., Tomás, J. M., *et al. Aeromonas hydrophila* AH-3 AexT is an ADP-ribosylating toxin secreted through the type III secretion system // Microbial Pathogenesis. 2008. 44, 1. 1–12.

Vinatzer, B. A., Teitzel, G. M., Lee, M. W., Jelenska, J., Hotton, S., *et al.* The type III effector repertoire of *Pseudomonas syringae* pv. syringae B728a and its role in survival and disease on host and non-host plants // Molecular Microbiology. 2006. 62, 1. 26–44.

Voisard, C., Keel, C., Haas, D., and Dèfago, G. Cyanide production by *Pseudomonas fluorescens* helps suppress black root rot of tobacco under gnotobiotic conditions. // The EMBO journal. 1989. 8, 2. 351–358.

Vojnov, A. A., Slater, H., Daniels, M. J., and Dow, J. M. Expression of the gum operon directing xanthan biosynthesis in *Xanthomonas campestris* and its regulation *in planta*. // Molecular plant-microbe interactions : MPMI. 2001. 14, 6. 768–774.

Völksch, B. and Weingart, H. Toxin production by pathovars of *Pseudomonas syringae* and their antagonistic activities against epiphytic microorganisms. // Journal of basic microbiology. 1998. 38, 2. 135–145.

Voth, D. E., Broederdorf, L. J., and Graham, J. G. Bacterial Type IV Secretion Systems: Versatile Virulence Machines // Future Microbiology. 2012. 7, 2. 241–257.

Walker, T. S., Bais, H. P., Déziel, E., Schweizer, H. P., Rahme, L. G., *et al.* *Pseudomonas aeruginosa*-plant root interactions. Pathogenicity, biofilm formation, and root exudation. // Plant physiology. 2004. 134, 1. 320–331.

Wandersman, C. Concluding remarks on the special issue dedicated to Bacterial secretion systems: Function and structural biology // Research in Microbiology. 2013. 164, 6. 683–687.

Wang, G. and Raaijmakers, J. M. Antibiotics production by bacterial agents and its role in biological control // The journal of applied ecology. 2004. 15, 6. 1100–1104.

Wang, Y., Huang, H., Sun, M., Zhang, Q., and Guo, D. T3DB: an integrated database for bacterial type III secretion system // BMC Bioinformatics. 2012. 13, 1. 66.

Watanabe, Y., Ozasa, K., Mermin, J. H., Griffin, P. M., Masuda, K., *et al.* Factory outbreak of *Escherichia coli* O157:H7 infection in Japan // Emerging Infectious Diseases. 1999. 5, 3. 424–428.

Wei, G. by Select Strains of Plant Growth-Promoting Rhizobacteria // Phytopathology. 1991. 81, 12. 1508.

Wensing, A., Braun, S. D., Büttner, P., Expert, D., Völksch, B., *et al.* Impact of siderophore production by *Pseudomonas syringae* pv. syringae 22d/93 on epiphytic fitness and biocontrol Activity against *Pseudomonas syringae* pv. glycinea 1a/96 // Applied and Environmental Microbiology. 2010. 76, 9. 2704–2711.

Wesseling, C., McConnell, R., Partanen, T., and Hogstedt, C. Agricultural Pesticide use in Developing Countries: Health Effects and Research Needs // International Journal of Health Services. 1997. 27, 2. 273–308.

Wiedemann, A. A., Virlogeux-Payant, I., Chausse, A. M., Schikora, A., and Velge, P. Interactions of Salmonella with animals and plants // Frontiers in Microbiology. 2015. 5, January. 1–18.

Winsor, G. L., Lam, D. K. W., Fleming, L., Lo, R., Whiteside, M. D., *et al.* *Pseudomonas* Genome Database: Improved comparative analysis and population genomics capability for *Pseudomonas* genomes // Nucleic Acids Research. 2011. 39, SUPPL. 1.

Yang, J., Kloepper, J. W., and Ryu, C. M. Rhizosphere bacteria help plants tolerate abiotic stress // Trends in Plant Science. 2009. 14, 1. 1–4.

Yang, K. Q., Qu, W. W., Liu, X., Liu, H. X., and Hou, L. Q. Causing Brown Apical Necrosis of Walnut in China // Plant Disease. 2011. 95, 6. 773–773.

Young, J. Olive knot and its pathogens // Australasian Plant Pathology. 2004. 33. 33–39.

Yuan, J., Li, B., Zhang, N., Waseem, R., Shen, Q., *et al.* Production of bacillomycin- and macrolactin-type antibiotics by *Bacillus amyloliquefaciens* NJN-6 for suppressing soilborne plant pathogens // Journal of Agricultural and Food Chemistry. 2012. 60, 12. 2976–2981.

Zechner, E. L., Lang, S., and Schildbach, J. F. Assembly and mechanisms of bacterial type IV secretion machines. // Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 2012. 367, 1592. 1073–87.

Zhang, L., Jia, Y., Wang, L., and Fang, R. A proline iminopeptidase gene upregulated *in planta* by a LuxR homologue is essential for pathogenicity of *Xanthomonas campestris* pv. campestris // Molecular Microbiology. 2007. 65, 1. 121–136.

Zhang, X. X. and Rainey, P. B. Exploring the sociobiology of pyoverdin-producing *Pseudomonas* // Evolution. 2013. 67, 11. 3161–3174.

Zhang, Z. and Wood, W. I. A profile hidden Markov model for signal peptides generated by HMMER // Bioinformatics. 2003. 19, 2. 307–308.

Zhao, J. and Grant, S. F. a. Advances in whole genome sequencing technology. // Current pharmaceutical biotechnology. 2011. 12, 2. 293–305.

Zheng, X. Y., Spivey, N. W., Zeng, W., Liu, P. P., Fu, Z. Q., *et al.* Coronatine promotes *Pseudomonas syringae* virulence in plants by activating a signaling cascade that inhibits salicylic acid accumulation // Cell Host and Microbe. 2012. 11, 6. 587–596.

Zimaro, T., Thomas, L., Marondedze, C., Garavaglia, B. S., Gehring, C., *et al.* Insights into *Xanthomonas axonopodis* pv. citri biofilm through proteomics. // BMC microbiology. 2013. 13, 1. 186.

Zimaro, T., Thomas, L., Marondedze, C., Sgro, G. G., Garofalo, C. G., *et al.* The type III protein secretion system contributes to *Xanthomonas citri* subsp. citri biofilm formation. // BMC microbiology. 2014. 14. 96.

ANEXO:

OTRAS PUBLICACIONES

environmental microbiology

Environmental Microbiology (2013) doi:10.1111/1462-2920.12240

Light regulates motility, attachment and virulence in the plant pathogen *Pseudomonas syringae* pv tomato DC3000

Isabel Río-Álvarez,[1,2] José Juan Rodríguez-Herva,[1,2] Pedro Manuel Martínez,[1,3] Pablo González-Melendi,[1,2] Gloria García-Casado,[4] Pablo Rodríguez-Palenzuela[1,2] and Emilia López-Solanilla[1,2]*

[1]*Centro de Biotecnología y Genómica de Plantas (CBGP), Universidad Politécnica de Madrid-Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Parque Científico y Tecnológico de la UPM. Campus de Montegancedo, Pozuelo de Alarcón, Madrid 28223, Spain.*

[2]*Departamento de Biotecnología, Escuela Técnica Superior de Ingenieros Agrónomos, UPM. Avda. Complutense S/N. 28040 Madrid, Spain.*

[3]*Área de Genética, Facultad de Ciencias, Universidad de Málaga, Instituto de Hortofruticultura Subtropical y Mediterránea 'La Mayora' (IHSM-UMA-CSIC), Málaga E-29071, Spain.*

[4]*Unidad de Genómica, Centro Nacional de Biotecnología-CSIC, Campus Universidad Autónoma, Madrid 28049, Spain.*

Summary

***Pseudomonas syringae* pv tomato DC3000 (Pto) is the causal agent of the bacterial speck of tomato, which leads to significant economic losses in this crop. Pto inhabits the tomato phyllosphere, where the pathogen is highly exposed to light, among other environmental factors. Light represents a stressful condition and acts as a source of information associated with different plant defence levels. Here, we analysed the presence of both blue and red light photoreceptors in a group of *Pseudomonas*. In addition, we studied the effect of white, blue and red light on Pto features related to epiphytic fitness. While white and blue light inhibit motility, bacterial attachment to plant leaves is promoted. Moreover, these phenotypes are altered in a blue-light receptor mutant. These light-controlled changes during the epiphytic stage cause a reduction in virulence, highlighting the relevance of motility during the entry process to the plant apoplast. This study demonstrated the key role of light perception in the Pto phenotype switching and its effect on virulence.**

Received 25 June, 2013; revised 30 July, 2013; accepted 30 July, 2013. *For correspondence. E-mail emilia.lopez@upm.es; Tel. (+34) 913364556; Fax (+34) 913365757.

Introduction

Light and circadian cycle regulate plant defence mechanisms (Roden and Ingle, 2009; Bhardwaj *et al*., 2011; Wang *et al*., 2011; Kangasjarvi *et al*., 2012). Plant pathogens might have evolved to sense light conditions associated with different levels of plant resistance. Indeed, bacteria possess mechanisms for the detection of light, which changes the motile behaviour (phototaxis) and other physiological processes for adaptation to the environment (Armitage and Hellingwerf, 2003; Purcell and Crosson, 2008). The effect of light has been extensively studied in phototrophic bacteria and only recently examined in non-phototrophic bacteria (Elias-Arnanz *et al*., 2011), suggesting that the use of light as a source of information might be an evolutionary advantage. One of the reasonable explanations to consider is that this ability has an advantage in the development of a protective program against certain harmful light wavelengths, but recent reviews have shown that in many bacteria, light governs other important lifestyle decisions (Gomelsky and Hoff, 2011).

The bacterial photosensory protein modules for visible light include: (i) the blue light sensing light, oxygen and voltage LOV (light, oxygen and voltage), which belongs to the PAS (Per-Arnt-Sim) domain family, BLUF (blue-light sensing using FAD), PYP (photoactive yellow protein), and the cryptochrome/photolyase family; (ii) proteorhodopsins; and (iii) the red light sensing bacteriophytochromes. The light-sensing capacity of these proteins is mediated through associations with specific light-absorbing molecules or chromophores. Most of these photosensory proteins have a modular architecture, where the light-sensing input domain (e.g. LOV) can be combined with diverse output domains, such as histidine kinases (HKs), HATPase, EAL from phosphodiesterases, sulfate transporters anti-sigma factor antagonist, or response regulator (RR) and helix-turn-helix. The light-induced control of gene expression involves a wide gamut

of regulatory mechanisms associated with other stress responses (Elias-Arnanz *et al*., 2011).

Although the photochemical properties of different types of photoreceptors have been analysed in some animal and plant pathogens (Lamparter *et al*., 2002; Swartz *et al*., 2007; Barkovits *et al*., 2008; 2011; Rottwinkel *et al*., 2010; Ondrusch and Kreft, 2011), the putative role of these proteins in the regulation of virulence remains poorly understood. Recent works show the involvement of photosensory proteins in the control of adaptive responses in *Agrobacterium tumefaciens* (Atu) (Oberpichler *et al*., 2008), *Rhizobium leguminosarum* (Rle) (Bonomi *et al*., 2012) and *Xanthomonas axonopodis* pv citri (Xax) (Kraiselburd *et al*., 2012). These recent and interesting reports revealed a key role for blue light perception in the control of bacterial lifestyles.

Pseudomonas syringae pv tomato DC3000 (Pto) is a hemibiotrophic pathogen that multiplies in the apoplastic space exploiting live host cells. Pto causes bacterial speck in tomatoes and other plants, which is characterized by the development of necrotic symptoms in leaves, stems and fruits. Pto epiphytically and endophytically grows on plant foliage without causing disease symptoms (Hirano and Upper, 2000). The epiphytic lifestyle represents the initial phase of foliar colonization by Pto. During this stage, Pto is exposed to the atmosphere and subjected to the diurnal cycle. Eventually, Pto enters the plant leaf apoplast where it multiplies exploiting live host cells assisted by the type three secretion system that injects effector proteins into plant cells (Alfano and Collmer, 1997; Preston, 2000; Gohre and Robatzek, 2008; Block and Alfano, 2011).

During the life in the phyllosphere, bacteria feature specific adaptations to an environment characterized by nutrient limitation, fluctuating water availability, exposition to ultraviolet radiation, or the presence of antimicrobials from plants or other microorganisms (Vorholt, 2012). Bacteria actively move to favorable sites on leaf surfaces. *Pseudomonas syringae* has been shown to regulate motility by quorum signal mechanisms during the epiphytic stage (Quiñones *et al*., 2005). Motility during this phase might be assisted by chemotaxis towards nutrients or plant molecules. Actually, Pto has been shown to be strongly attracted by plant origin molecules present in the tomato host plant (Cuppels, 1988). On the other hand, commensal bacteria found in the phyllosphere have developed specific adaptive mechanisms mainly devoted to adhere to the leaf surface and to produce EPS (Exopolysacharide) protective structures (Danhorn and Fuqua, 2007; Vorholt, 2012). It might be hypothesized that epiphytic fitness in a phytopathogenic bacteria would be determined by the ability to move looking for a favourable environment together with the ability to adhere to plant surface and to resist over the diurnal cycle the stressful situations imposed by the environment. The switch to a pathogenic lifestyle would require the entry to the plant apoplast, which is a motility dependent process. This switch might be dependent on light conditions. Interestingly, homologues to light receptors have been identified in Pto (van der Horst *et al*., 2007). A Pto LOV protein exhibits light-regulated HK activity (Swartz *et al*., 2007) and light-regulated communication between its various domains (Cao *et al*., 2008). Moreover, the two Pto PHY proteins are active under red light *in vitro* (Shah *et al*., 2012). However, no data concerning the putative role of these photosensory proteins in the pathogenesis have been reported.

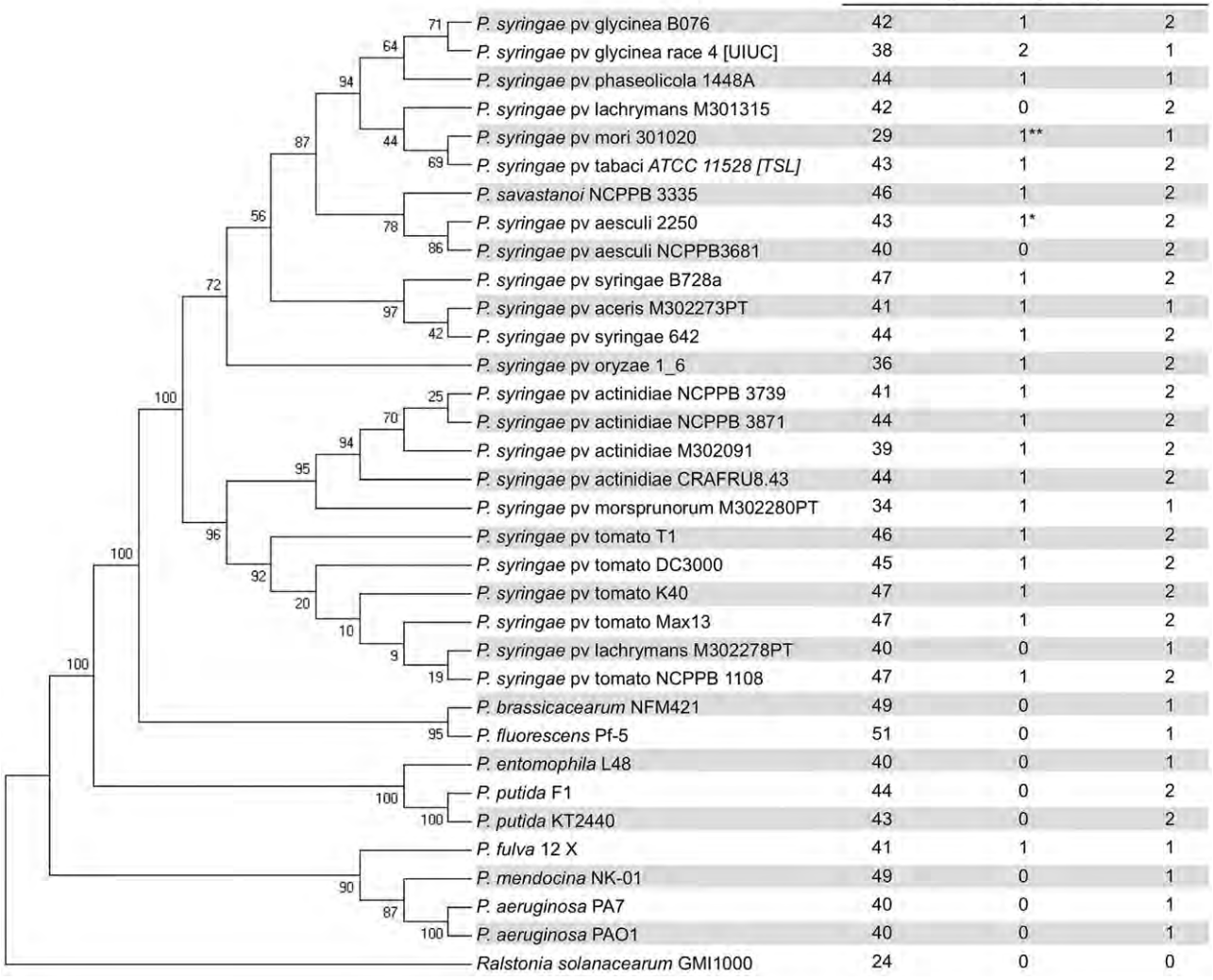
In this study, we conducted a comparative analysis of the available *Pseudomonas* genomes to identify proteins containing LOV or PHY domains. In addition, we studied the effect of light on Pto phenotypes like motility and adhesion, and how this affects virulence. We observed that the blue light component is primarily responsible for the observed phenotypes. This conclusion was confirmed through an analysis of the phenotype resulting from a mutation in the LOV protein involved in blue light perception.

Taken together, these results show the specific control of white light over traits associated with Pto virulence.

Results

Plant-associated Pseudomonas *possess LOV and PHY photoreceptors*

Among the bacterial photosensory protein modules for visible light, blue light-sensing LOV proteins and red light-sensing bacteriophytochromes have been identified in many bacterial species. We have screened 33 *Pseudomonas* genomes for the presence of genes encoding these proteins, with respect to the phylogeny of the strains analysed, using a bioinformatics approach for the identification of the domains involved in this perception (Fig. 1, Tables S1 and S2). It has been previously suggested that the PAS family proteins, with the structure LOV-HK-RR, are predominately found in plant pathogenic bacteria (Losi and Gartner, 2008; 2011) and that these domains share important functional amino acid residues for photochemistry and signalling. Most of the *P. syringae* genomes analysed here possessed at least one protein with this structure, whereas other non-pathogenic *Pseudomonas*, even plant-associated bacteria, such as *P. fluorescens*, did not contain these proteins. The tomato pathogen Pto genome encodes 48 proteins containing PAS domains. Among these, only one protein presented the LOV-HK-RR structure. Other strains of this pathovar also showed the presence of one LOV-HK-RR protein. Regarding the presence of PHY-containing proteins, which might be involved in red light perception, the



	PAS	LOV (HK-HATPase-RR)	PHY
P. syringae pv glycinea B076	42	1	2
P. syringae pv glycinea race 4 [UIUC]	38	2	1
P. syringae pv phaseolicola 1448A	44	1	1
P. syringae pv lachrymans M301315	42	0	2
P. syringae pv mori 301020	29	1**	1
P. syringae pv tabaci *ATCC 11528 [TSL]*	43	1	2
P. savastanoi NCPPB 3335	46	1	2
P. syringae pv aesculi 2250	43	1*	2
P. syringae pv aesculi NCPPB3681	40	0	2
P. syringae pv syringae B728a	47	1	2
P. syringae pv aceris M302273PT	41	1	1
P. syringae pv syringae 642	44	1	2
P. syringae pv oryzae 1_6	36	1	2
P. syringae pv actinidiae NCPPB 3739	41	1	2
P. syringae pv actinidiae NCPPB 3871	44	1	2
P. syringae pv actinidiae M302091	39	1	2
P. syringae pv actinidiae CRAFRU8.43	44	1	2
P. syringae pv morsprunorum M302280PT	34	1	1
P. syringae pv tomato T1	46	1	2
P. syringae pv tomato DC3000	45	1	2
P. syringae pv tomato K40	47	1	2
P. syringae pv tomato Max13	47	1	2
P. syringae pv lachrymans M302278PT	40	0	1
P. syringae pv tomato NCPPB 1108	47	1	2
P. brassicacearum NFM421	49	0	1
P. fluorescens Pf-5	51	0	1
P. entomophila L48	40	0	1
P. putida F1	44	0	2
P. putida KT2440	43	0	2
P. fulva 12 X	41	1	1
P. mendocina NK-01	49	0	1
P. aeruginosa PA7	40	0	1
P. aeruginosa PAO1	40	0	1
Ralstonia solanacearum GMI1000	24	0	0

Fig. 1. Presence of photosensory proteins in *Pseudomonas* strains in a phylogenetic context. The evolutionary history was inferred using the Maximum Likelihood method based on the JTT matrix-based model. The percentage of trees in which the associated taxa clustered in the bootstrap test (500 replicates) is shown next to the branches. PAS, LOV, HK-HATPase, RR and PHY stand for per (period circadian protein) arnt (aryl hydrocarbon receptor nuclear translocator protein) sim (single minded protein); LOV; HK containing ATPase domain; response regulator; phytochrome; respectively. *The protein contains the module LOV-HK-HATPase but lacks the RR domain. **The protein contains the module LOV-HK but lacks the HATPase and RR domains. The *Ralstonia solanacearum* genome was included as an outgroup for the phylogenetic analysis.

number of putative bacteriophytochromes identified varied between one and two for most of the *P. syringae* phytopathogenic strains analysed and for soil plant-associated bacteria. Pto has two bacteriophytochromes.

White light and particularly the blue component inhibit bacterial motility and promote plant attachment

Pto exhibits a swarming phenotype on 0.3% agar plates, where flagellated bacteria typically develop surface-associated movement (Berti *et al.*, 2007). To determine the effect of white light on Pto swarming, the bacteria were cultured for 16 h under 15, 50, 55, 60 and 70 μE m^{-2} s^{-1} white light or dark conditions (Fig. S1). Notably, light intensities generated through artificial indoor lighting are typically below 20 μE m^{-2} s^{-1} in standard lab rooms, while the intensities are higher outdoors and in plant growth chambers (100–150 μE m^{-2} s^{-1}). A clear inhibition of motility was observed under white light intensities above 60 μE m^{-2} s^{-1}, whereas the characteristic-swarming behaviour described for Pto was observed at lower intensities and under dark conditions (Fig. S1). To reveal the particular contribution of the two major types of monochromatic light (i.e. blue and red light) to the inhibition of Pto motility, bacteria were

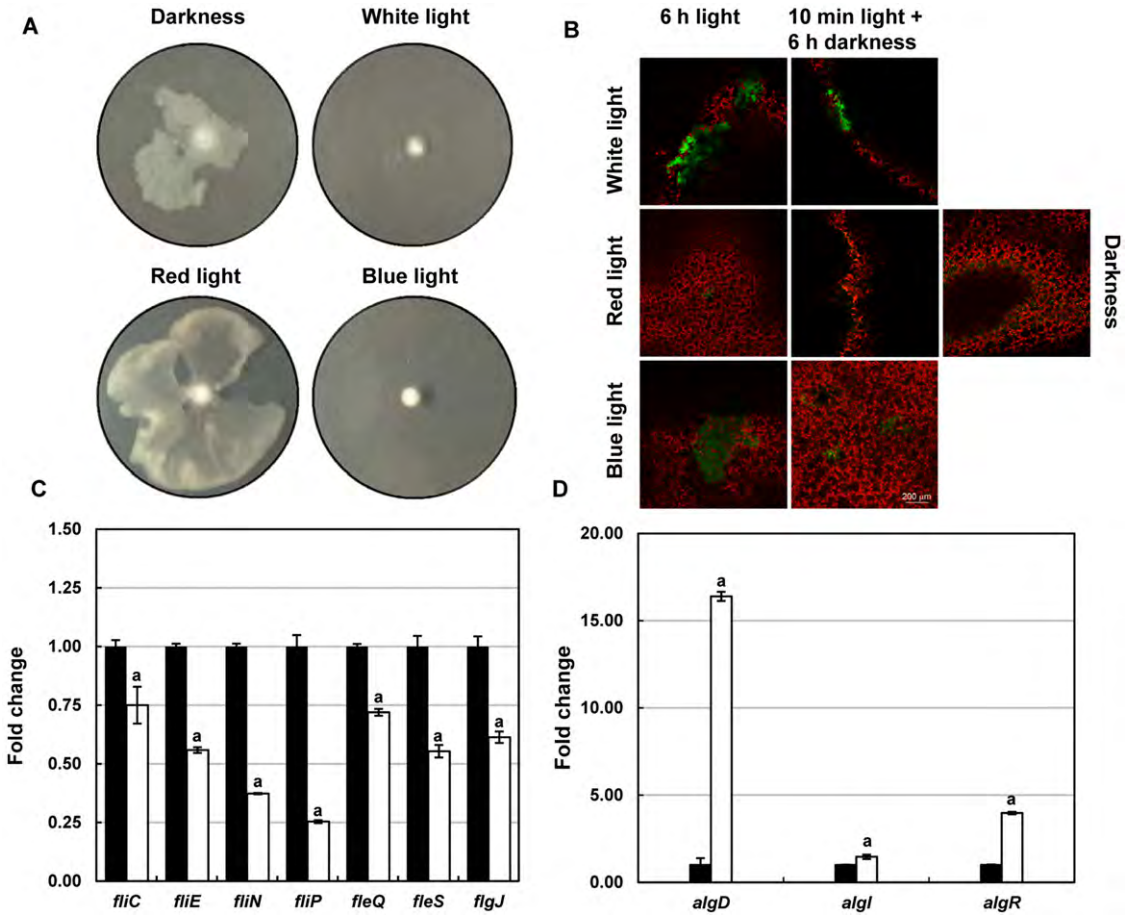


Fig. 2. Light conditions affect Pto swarming motility and attachment to plant leaves.
A. KB agar plates (0.3%) were inoculated with the WT strain using a sterile toothpick and incubated for 16 h at 28°C. The light conditions included: darkness, white light (70 µE m^{-2} s^{-1}), red light (20 µE m^{-2} s^{-1}) and blue light (20 µE m^{-2} s^{-1}).
B. 2×10^7 cfu ml^{-1} was placed on the underside of *A. thaliana* leaves. The leaves were incubated for 6 h under constant light (white, red or blue), 10 min under light (white, red or blue), followed by 6 h of dark conditions or for 6 h in the darkness. Overlay of the GFP signal and chlorophyll autofluorescent confocal images is shown.
C. Differential expression of flagella and (D) alginate biosynthesis genes was evaluated by qRT-PCR after a 10 min white light treatment (white bars) with respect to the darkness (black bars). The means and standard errors of three replicates are shown. Error bars represent SEM (Standard Error Mean). (a) Significant differences between light and dark treatments were determined according to Student's *t* test ($P < 0.05$). The results shown in (A–D) are representative of at least three independent experiments.

grown under white (70 µE m^{-2} s^{-1}), red (20 µE m^{-2} s^{-1}) or blue (20 µE m^{-2} s^{-1}) light conditions. The data showed that the blue light component of white light is sufficient to inhibit Pto motility, and red light exerts the same effects as dark conditions facilitating bacterial motility (Fig. 2A).

Because light causes oxidative stress, which could have a toxic effect on bacteria, Pto growth was monitored for 24 h in both liquid and solid medium (Fig. S2) under the same lighting conditions described earlier. The results confirm that the photo-inhibition of Pto motility is not the consequence of an indirect effect on bacterial growth.

On the other hand, biofilm development is incompatible with the single cell motile stage, and light influences the switch between both stages in bacteria (Gomelsky and Hoff, 2011). As blue light exerts a potent inhibitory effect on motility, we hypothesized that blue light could signal the switch to a non-motile attached stage. To test this hypothesis, we assessed the attachment of a Pto-GFP (green fluorescent protein) strain to *Arabidopsis thaliana* leaves under different light conditions (Fig. 2B). We observed that after 6 h under white and specifically under blue light, the bacteria remained attached, forming organized structures, which could be considered as biofilms (although biofilm and attachment are not necessarily equivalent). Under dark conditions and red light, the remaining bacteria were not attached. Moreover, a 10-min light treatment before plant challenge was sufficient to induce the observed phenotypes regarding bacterial attachment (Fig. 2B).

White light controls expression of alginate biosynthesis genes and flagellar biogenesis

To get more insight about the effect of light on the observed phenotypes, we carried out qRT-PCR (quantitative real time-PCR) analyses of the expression of genes involved in motility and polysaccharide production. To analyse the expression of genes involved in flagellum biosynthesis, we selected *fliP* and *flgJ* that codify elements related with the biosynthesis, *fliC* that codifies flagellin, *fliN* and *fliE* that codify a flagellar motor switch protein, and the flagellar hook basal body, respectively, and finally *fleS* and *fleQ* involved in signal transduction and regulation of flagellar functions. The selection was based both on the differential functions reflected as well as on their genetic location in different transcriptional units. The data showed the downregulation by a 10-min white light treatment of these genes associated with the observed motility phenotype (Fig. 2C). Exopolysaccharides have been described to be involved in adhesion and resistance mechanisms during the epiphytic stage. Alginate is a Pto exopolysaccharide that contributes to virulence and epiphytic fitness (Yu *et al*., 1999; Keith *et al*., 2003). We chose *algI*, *algR* and *algD* for the expression analysis. AlgI and AlgD are involved in the alginate biosynthesis and were selected for being located in different operons. AlgR is a regulatory protein controlling alginate biosynthesis. We found that these genes were upregulated by a 10-min white light treatment (Fig. 2D).

Virulence is differentially regulated by blue and red light

As light controls motility and surface attachment in Pto, we examined whether the light also alters the virulence of this pathogen. Tomato plants, the host plants of Pto infection, were challenged with bacterial cells, pretreated or not with white light for 10 min before inoculation. Then, plants were transferred to a climate-controlled chamber under standard light cycle conditions (12 h light/12 h darkness), starting in the darkness. After 6 days, the symptoms and bacterial population were recorded. Figure 3A shows that both the symptoms (evaluated as number of chlorotic-necrotic spots) and bacterial population were reduced in plants challenged with bacterial cells pretreated with light.

Arabidopsis thaliana is also a host plant for Pto, so we performed the same experiment using this plant and observed the same differences after white light treatment as observed in tomato (Fig. 3B). To differentiate the effects of blue and red light on bacterial virulence, *Arabidopsis* plants were inoculated with bacterial cells subjected to the 10-min treatment under blue or red monochromatic lights. The results showed that blue light is responsible for the reduction of the symptoms, whereas red light generates the opposite effect, enhancing the appearance of symptoms compared with the darkness treatment.

To determine the biological significance of the light treatment in the control of the virulence, we designed a specific plant inoculation experiment. As the LOV photoreceptor dark recovery occurs within 5650 s (94.16 min) (Cao *et al*., 2008), we performed a plant challenge experiment in which bacterial inoculum was subjected to a 10 min light/100 min darkness/10 min light cycle. After each treatment an aliquot of the bacterial suspension was used to inoculate *Arabidopsis* plants. Figure 3C shows the bacterial population at 6 days post-inoculation (dpi) in plants challenged with the different suspensions. The results show that after treatment in darkness for 100 min, virulence was restored compared with the inhibition attained after light treatment; a subsequent 10-min light treatment reduced the bacterial virulence again.

Light controls the entry to the plant apoplast

As both the motility phenotype and virulence are altered after white light treatment and it has been previously suggested the requirement of an active motility to enter the plant apoplast, we decided to further study the relevance of flagellar motility in the context of the entry process.

We first examined whether the amount of FliC, the flagellar filament protein, was altered. The presence of this protein in samples from bacterial cultures grown 6 h under white light and dark conditions was assessed through Western blot analyses (Fig. 4A and B). The apparent molecular weight of the lower band corresponds to the predicted size for the Pto flagellin (29 kDa). The upper band corresponds to the expected size of the glycosylated flagellin (32 kDa), as previously described (Takeuchi *et al*., 2003). The results clearly showed that the amount of the non-glycosylated flagellin was reduced under light conditions. In addition, we stained the bacterial samples obtained from the border regions of the migration zones of swarming plates for flagella visualization. Figure 4C shows that under 16 h white light conditions, Pto cells practically do not present flagellar structures.

To test whether the repression of motility by white light is affecting the entry process and therefore the onset of the infection, we infiltrated *A. thaliana* leaves with Pto cells pretreated or not with white light for 10 min before inoculation. We compared two types of inoculation methods: spray inoculation versus infiltration under the premise that sprayed cells would need active motility to reach entry sites. The results showed that when infiltrated with Pto cells pretreated with white light, the differences regarding both symptoms and bacterial population in plant are smaller than those when using spray inoculation

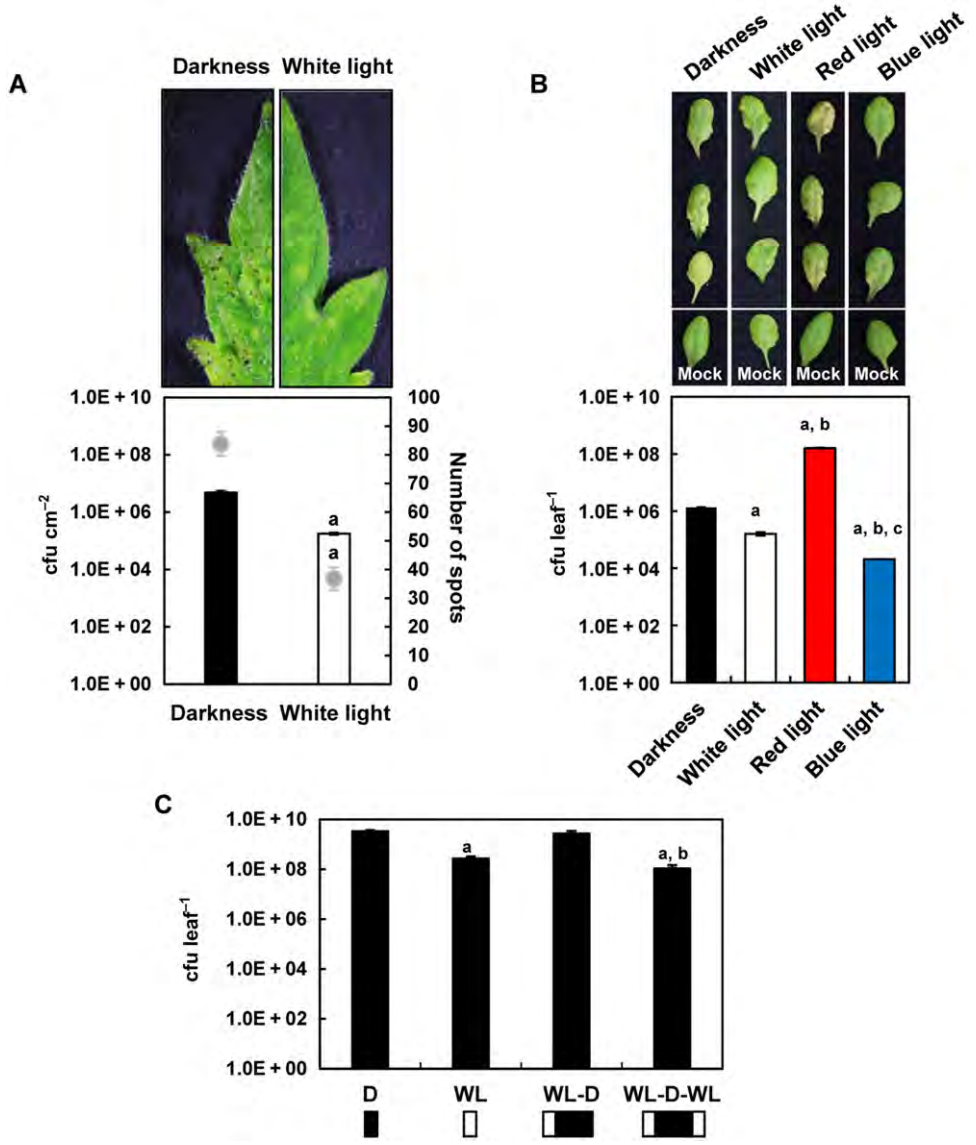


Fig. 3. Light treatment of Pto before infection reduces disease symptoms and bacterial population.
A. Bacterial population per unit area (bars) and number of lesions (grey points) on tomato plants were determined at 6 days post-inoculation (dpi) by dipping with suspensions of 10^8 cfu ml^{-1} cells pretreated for 10 min with white light or maintained under darkness. The means and standard errors of five replicates are shown. (a) Significant differences between light and dark treatments were determined according to Student's *t* test ($P < 0.05$).
B. Disease symptoms and bacterial population (bars) were observed on *A. thaliana* Col-0 leaves at 6 dpi by spraying Pto cells suspensions containing 3×10^8 cfu ml^{-1} pretreated for 10 min either with white light (70 μE m^{-2} s^{-1}), red light (20 μE m^{-2} s^{-1}), blue light (20 μE m^{-2} s^{-1}) or maintained under darkness. The means and standard errors of three replicates are shown. Significant differences with respect to the dark (a), white light (b) and red light (c) treatments were determined according to Student's *t* test ($P < 0.05$).
C. Pto population in *A. thaliana* Col-0 leaves at 6 dpi by spraying Pto cells suspensions containing 3×10^8 cfu ml^{-1}. The bacterial suspension was pretreated for 10 min with white light (70 μE m^{-2} s^{-1}) (WL) or maintained under darkness (D). The bacterial suspension treated with light was subsequently subjected to a 100-min dark treatment (WL-D) and treated again with 10 min of white light (WL-D-WL). After each treatment, an aliquot of the inoculum was used to challenge plants. The means and standard errors of three replicates (three leaves each) are shown. Significant differences with respect to the dark treatment (D) (a), and with respect to the second dark treatment (WL-D) (b) were determined according to Student's *t* test ($P < 0.05$). The results shown in (A–C) are representative of at least three independent experiments.

(Fig. 4D and results in Fig. 3B). A Pto *fliC* mutant strain was used to check the requirement of the flagellum during the entry process. The virulence of this strain when sprayed is significantly less virulent than the wild-type (WT) strain (Fig. 4E), while when using infiltration, the *fliC* mutant virulence is at the same level of the WT virulence (Fig. 4F). This reveals that an active movement is essential to enter the plant apoplast and to initiate the infection.

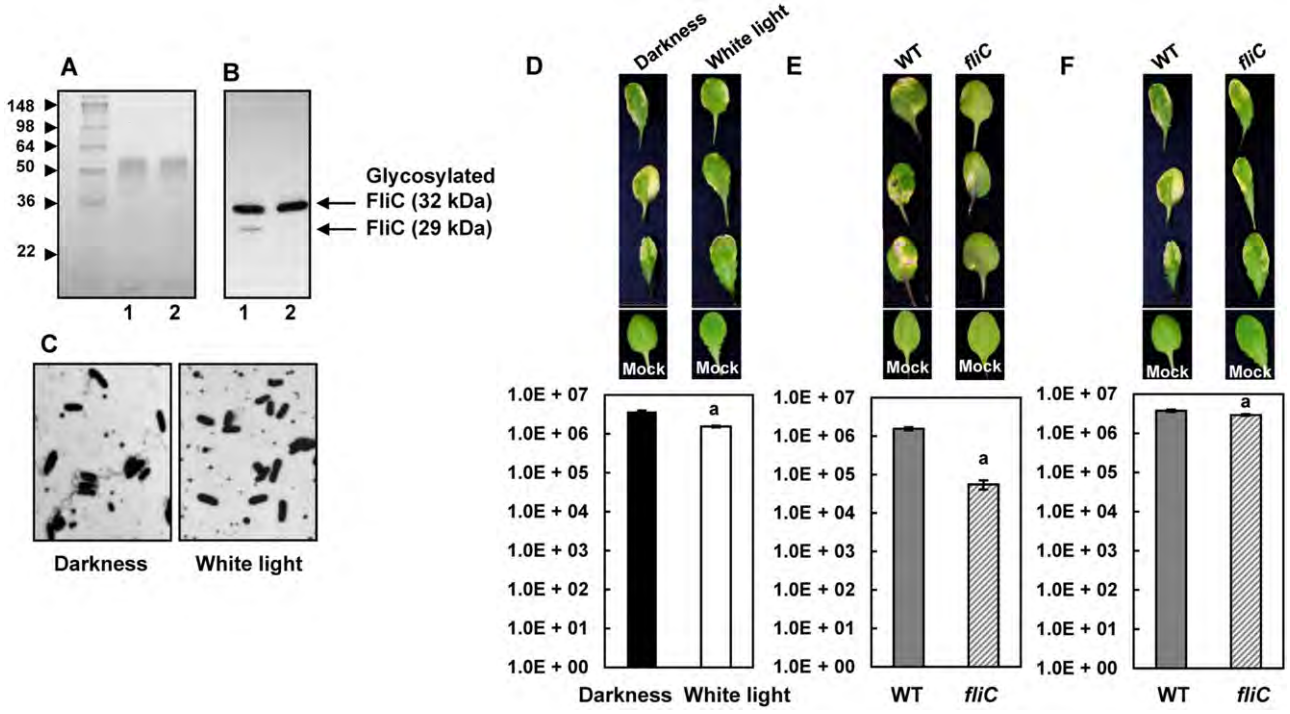


Fig. 4. White light controls the entry to plant apoplast mediated by an active motility.
A. Extracellular and membrane proteins from bacterial cultures exposed to darkness (lane 1) or white light (lane 2) were electrophoresed in 15% acrylamide gels and stained with Coomassie Brilliant Blue.
B. Western blot analysis of these proteins reacted with a flagellin antiserum. The observed bands running at approximately 29 and 32 kDa correspond to FliC and glycosylated FliC, respectively.
C. Images (630×) of bacteria stained with simplified Leifson method are shown.
D. Disease symptoms and bacterial population (bars) were observed on *A. thaliana* Col-0 leaves at 2 dpi by infiltrating Pto cells suspensions containing 10^5 cfu ml^{-1} pretreated for 10 min with white light (70 μE m^{-2} s^{-1}) or maintained under dark conditions. The means and standard errors of three replicates are shown. (a) Significant differences between light and dark treatments were determined according to Student's *t* test ($P < 0.05$).
E. Disease symptoms and bacterial population (bars) were observed on *A. thaliana* Col-0 leaves at 6 dpi by spraying Pto WT and *fliC* cells suspensions containing 3×10^8 cfu ml^{-1}.
F. Disease symptoms and bacterial population were observed on *A. thaliana* Col-0 leaves at 2 dpi by infiltrating Pto WT and *fliC* cells suspensions containing 10^5 cfu ml^{-1}. For (E) and (F), the means and standard errors of three replicates are shown. (a) Significant differences between Pto WT and Pto *fliC* treatments were determined according to Student's *t* test ($P < 0.05$). The results shown in (A)–(F) are representative of at least three independent experiments.

All together, these results suggest that the effect of light over Pto motility is the main responsible of the reduction in virulence observed in light-pretreated cells.

The ability to regulate light-related phenotypes is compromised in the LOV-Pto mutant

To determine the contribution of the LOV-Pto photoreceptor to the light-dependent phenotypes observed, a mutant in this gene was constructed. The swarming motility of this strain was assayed under different light conditions (Fig. 5A), and the results showed that neither white nor blue light completely inhibited the motility of the LOV mutant compared with the inhibition observed in the WT strain. However, the behaviour of this strain under red light conditions was not altered. The growth rate of LOV mutant in culture media was similar to that of the WT strain both under darkness and light conditions (data not shown). The phenotype was restored to WT levels in the complemented strain. These results suggest that this photoreceptor is partially responsible for the bacterial response under both white and blue light conditions. Moreover, the analysis of the expression of the *fliC* gene through qRT-PCR in the LOV-Pto mutant showed that there were no differences in the expression of this gene under light treatment or darkness (Fig. 5B).

The attachment of the LOV-Pto mutant strain to the plant surface was also evaluated in *Arabidopsis* leaves (Fig. 5C). While the WT strain attached leaves under light conditions, the mutant strain did not remain attached to the plant surfaces. The results suggest that this phenotype was also under the control of light perception through this photoreceptor. Moreover, through qRT-PCR analyses, we observed that under

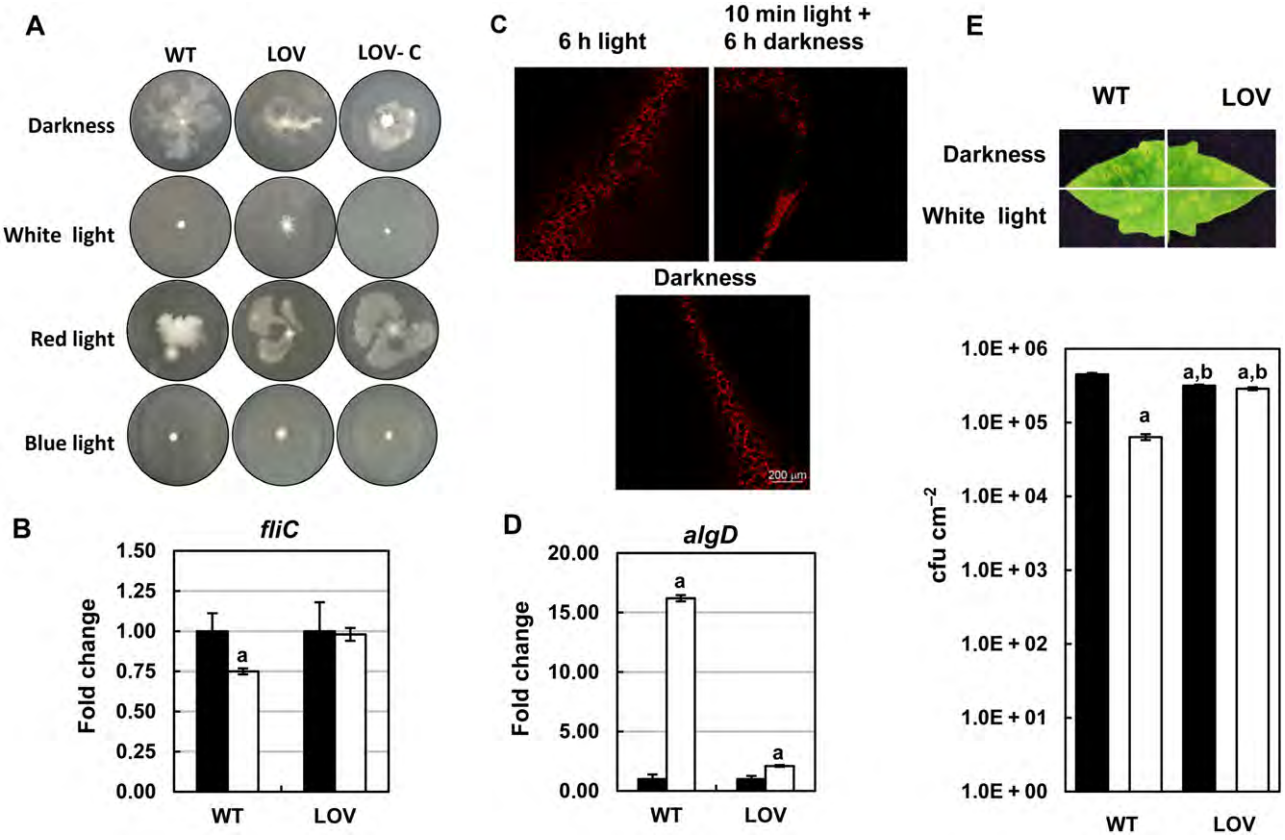


Fig. 5. LOV-Pto mutant response to light is altered.
A. Semi-solid KB agar (0.3%) plates were inoculated with Pto WT, LOV-Pto mutant and a LOV-C complement strain using a sterile toothpick. The plates were incubated for 16 h at 28°C under dark, white light (70 µE m^{-2} s^{-1}), red light (20 µE m^{-2} s^{-1}) or blue light (20 µE m^{-2} s^{-1}) conditions.
B. 2×10^7 cfu ml^{-1} of the LOV-Pto mutant was placed on the underside of *A. thaliana* leaves. The leaves were incubated for 6 h under constant white, red or blue light or for 10 min with white, red or blue light, followed by 6 h of dark conditions or for 6 h in the darkness. Overlay of the GFP signal and chlorophyll autofluorescent confocal images is shown.
C. Differential expression of *fliC* and (D) *algD* genes was evaluated by qRT-PCR after a 10-min white light treatment (white bars) or dark treatment (black bars) in a LOV-Pto mutant with respect to the Pto WT strain. The means and standard errors of three replicates are shown. (a) Significant differences between dark and light treatments were determined according to Student's *t* test ($P < 0.05$).
E. Disease symptoms and bacterial population (bars) were determined on tomato plants at 6 dpi by dipping with suspensions containing 10^8 cfu ml^{-1} of Pto WT and LOV-Pto mutant cells pretreated for 10 min with white light (70 µE m^{-2} s^{-1}) or maintained under darkness. The means and standard errors of five replicates are shown. Significant differences with respect to the dark treatment (a) and with respect to the white light treatment (b) of the WT strain were determined according to Student's *t* test ($P < 0.05$). The results shown in (A)–(E) are representative of at least three independent experiments.

light conditions, the expression of *algD* in the LOV-Pto mutant was much lower than the expression in the WT strain (Fig. 5D).

To analyse the virulence of LOV-Pto mutant, we challenged tomato plants with the WT and mutant strains after a 10-min treatment under dark or white light conditions. We observed that regardless of the light treatment, the mutant strain was less virulent than the WT strain for both symptom production and bacterial population at 6 dpi. Moreover, the virulence in the mutant strain was not regulated by white light, and the symptoms and bacterial population recorded *in planta* was not diminished after the white light treatment, as observed for the WT strain (Fig. 5E).

Discussion

Proteins with photo-reactive LOV or PHY domains involved in blue and red light perception, respectively, have been identified in phytopathogenic bacteria (Rottwinkel *et al.*, 2010; Bonomi *et al.*, 2012; Kraiselburd *et al.*, 2012). Our analysis of the distribution of the canonical LOV-HK-RR structure revealed the presence of at least one member of this family in almost all the plant-pathogenic *P. syringae* genomes analysed. Interestingly, the canonical blue light receptor is not present in root-associated *Pseudomonas* species (*P. putida* and *P. fluorescens*), suggesting that the evolutionary incentive to use blue light as a source of information primarily exists

among bacteria exposed to day light cycling conditions, as observed for most of the *P. syringae* strains analysed. These plant pathogens typically exhibit an epiphytic phase prior to the pathogenic phase (Hirano and Upper, 2000). In contrast with blue light receptors, phytochromes have been identified in many different strains, regardless of their niche. Indeed, these proteins (and specifically those classified as bathy phytochromes) have been described to dominate in *Rhizobiales* and are most frequently represented in soil bacteria (Rottwinkel *et al*., 2010). The light regime in the soil is rich in long-wavelength light. Therefore, this is congruent with the idea that it might have an evolutionary incentive for the use of light as a source of information (van der Horst *et al*., 2007), being red light more important for soil bacteria and white light more important for bacteria in the phyllosphere.

Therefore, the aim of the work is the study of the light effect on Pto during the stages previous to the infection, when this bacterium inhabits the phyllosphere (epiphytic stage) and how that influences the entry process, and subsequently the success of the infection. Of course, we do not discard a light effect during the interaction, but probably that effect is influenced by the plant response to the light.

During the course of this research, it became clear that the type, duration and intensity of the light treatments are critical to the phenotype regulation of Pto. Previous analyses of the light influence on different heterotrophic bacteria have been conducted using light treatments over long time periods (6–48 h), i.e. the experiments with Rle or Xax (Bonomi *et al*., 2012; Kraiselburd *et al*., 2012), or over shorter time periods (10 min to 1 h), i.e. experiments with non-plant pathogenic bacteria, such as *L. monocytogenes* (Ondrusch and Kreft, 2011) and *Brucella* spp. (Swartz *et al*., 2007). The selection of the duration of the light treatment considered both the kinetics of the photocycle for the blue light photoreceptor present in Pto and the stress response associated with light conditions (Ziegelhoffer and Donohue, 2009). The time required for light activation of the LOV-photoreceptor and photoadduct formation, which initiates signal transduction, is short, at 20 ns and 1.5 μs, respectively. Whereas the recovery of the photoadduct to the dark state is slow, yielding an average lifetime of 5650 s (Cao *et al*., 2008). Based on these considerations, we used a short light treatment of 10 min to further analyse the effects of light on different phenotypes or the gene expression that were not associated with long-term adaptive responses to light stress (oxidative stress, osmotic stress, etc.). The intensity of the white light used in our experiments (70 $\mu E\ m^{-2}\ s^{-1}$) is close to the illumination level at twilight on a day with clear sky conditions (Thorington, 1985). In contrast, the light intensity conferred by standard artificial indoor lighting is approximately 15 $\mu E\ m^{-2}\ s^{-1}$, while at midday on a sunny day at the coordinates 40°26′10″N 3°48′50″W, the intensity of the light outdoors is 1000 and 140 $\mu E\ m^{-2}\ s^{-1}$ in a greenhouse.

Motility has been associated with bacterial infection of the host plant, as the entry process into plant apoplast depends on it (Panopoulos and Schroth, 1974; Ichinose *et al*., 2003; Yao and Allen, 2006; Antunez-Lamas *et al*., 2009). The results obtained in the present study showed that the blue, but not the red, component of white light is responsible for the inhibition of swarming motility in Pto. This feature is altered by light treatment in other plant-associated bacteria (Oberpichler *et al*., 2008; Bonomi *et al*., 2012; Kraiselburd *et al*., 2012). For Atu, white light also inhibits motility, whereas for Rle and Xax, motility is not affected. Notably, the light intensities for Atu and in the present study are higher than those for Rle and Xax. Moreover, the gene expression analysis after the 10 min of treatment under these light conditions showed the downregulation of the genes involved in flagellar synthesis and function. These data were confirmed through Western blot experiments to detect the flagellar filament protein (FliC) and microscopic analyses of cells. The flagellar proteins are inhibited in the other models analysed (Oberpichler *et al*., 2008; Bonomi *et al*., 2012; Kraiselburd *et al*., 2012). The analysis of motility and the *fliC* gene expression upon light treatment in the LOV-Pto mutant suggests that motility regulation depends on this photoreceptor. These results support the hypothesis that the blue component of the light inhibits flagella-dependent motility in Pto. Moreover, there are clear differences in the results obtained using an inoculation method that requires motility (spray) versus an inoculation method that does not (infiltration). The results using infiltration still reveals differences between the white light treatment and the dark treatment suggesting that motility is the main factor affected by the treatment, although other virulence determinants might be regulated by light at this stage. Actually, a Pto *fliC* mutant is severely affected in virulence when the inoculation method makes an active movement essential to enter the plant apoplast and initiate the infection. The enhanced virulence of a Pto *fliC* mutant observed when infiltrated could be due to the absence of flagellin recognition by plant. Actually, it has been described that plant immunity triggered by flagellin limits pathogen fitness (Cai *et al*., 2011). Flagellar motility contributes to epiphytic fitness in *Pseudomonas* spp. (Haefele and Lindow, 1987), and very recently, Yu and colleagues (2013) have found that flagellar motility is an adaptive feature of the epiphytic phase in *P. syringae* pv syringae B728a. Our results suggest that flagellar motility, which is essential for spreading on the leaf and for the entry to the plant apoplast, is controlled by light. In this context, the active motility of bacteria on plant surfaces would be favoured under low-intensity light conditions.

On the other hand, biofilm formation and attachment to plant surfaces requires the inhibition of bacterial motility and, therefore, the switch from a motile state to a sessile state (Verstraeten *et al*., 2008). The experiments using GFP-labelled bacteria showed that a light treatment as short as 10 min was sufficient to promote bacterial attachment to *Arabidopsis* leaves. Consistently, the qRT-PCR data showed that the genes involved in alginate synthesis were upregulated after light treatment. Moreover, the LOV-Pto mutant did not attach to *Arabidopsis* leaves and showed an altered expression of *algD* in response to light, suggesting that the blue component of white light is responsible for this phenotype. For *Rhizobiales*, including Atu and Rle, the adherence to biotic surfaces is inhibited by light, whereas for Xax, similar to Pto, blue light induces adherence to plant surfaces. Light might inhibit the transition from one state to another, but the particular regulation depends on the environmental significance of light for the bacteria in a given niche (Gomelsky and Hoff, 2011). These results are in line with our hypothesis that the switch from a motile to a sessile state is regulated by light in Pto.

The virulence of Pto (measured as symptom development and bacterial population) was diminished after the 10 min white and blue light treatments in both tomato and *Arabidopsis*. The reduced virulence could be the consequence of the downregulation of flagellar genes, which in turn causes a reduction in the motility-dependent entry process. Therefore, the results obtained in the present study suggest that light absence is an important factor for the transition from the pre-infection to the infection phase. A parallel phenomenon has been reported in Atu and Xax (Oberpichler *et al*., 2008; Kraiselburd *et al*., 2012). However, in the present study, we treated the cells prior to plant inoculation for 10 min with the desired light, while in previous Atu and Xax experiments, the plants were subjected to the light treatment after infection. Surprisingly, for Pto, red light treatment increases the bacterial virulence in *Arabidopsis* plants. It cannot be ruled out that the downstream regulation through white light perception reflects the balance between red and blue light. Further investigations regarding the signal transduction process under both types of light will be informative to unveil this aspect.

The LOV-Pto mutant is less virulent than the WT regardless of the light treatment. A similar result has been reported for the mutant in the Xax homologue protein (Kraiselburd *et al*., 2012). However, we cannot rule out that this photoreceptor might be regulating other aspects of the bacterial physiology. In any case, these results show that the mutant loses the ability to regulate the response to the changing light conditions.

In this study, we conducted a comparative analysis of the available *Pseudomonas* genomes to identify proteins containing LOV and PHY domains. In addition, we studied the effect of light on Pto phenotypes like motility and adhesion, which in turn are related to virulence. We observed that the blue light component is primarily responsible for the observed phenotypes. This conclusion was confirmed through an analysis of the phenotype resulting from a mutation in the LOV protein involved in blue light perception.

In summary, this work demonstrated the key role of light perception in Pto phenotype switching. Several questions arise from these results: What is the relative importance of this environmental signal in the context of the plant–pathogen interactions? Does Pto use light as the master signal to regulate virulence? It is reasonable to propose that Pto uses this environmental signal not only as a direct warning of desiccation and oxidative stress but also as an indirect warning of the plant defence level. The knowledge of light-sensing mechanisms and the pathways involved in the control of phenotype switch will open new avenues for designing strategies to interfere with this phenomenon to prevent the establishment of bacterial infections.

Experimental procedures

Identification of protein domains

The genome sequences (including plasmids sequences when present) of 33 bacterial species belonging to the genus *Pseudomonas* were downloaded from the ASAP (A Systematic Annotation Package) website (http://asap.ahabs.wisc.edu/asap/home.php) and NCBI (National Center for Biotechnology Information) (http://www.ncbi.nlm.nih.gov/). The identification of PHY and PAS protein domains was performed using an internal pipeline. First, PHY and PAS domains full alignments were downloaded from the PFAM website of the Sanger Centre (http://pfam.sanger.ac.uk/) using the PFAM identifiers PF00989 and PF00360, respectively. These alignments were used to build Hidden Markov Models (HMMs) using HMMER v2.3.2 (Eddy, 1998). Finally, a customized BioPerl (Stajich *et al*., 2002) script combined with HMMER searched for HMM hits using an *E*-value of 0.005 (Tables S1 and S2). For the LOV domains, only PAS-containing proteins containing the amino acid sequence GXNCRFLQ (Briggs, 2007) within the PAS domain were selected. HisKA, HATPase and Response_reg domains were identified using the PFAM identifiers PF00512, PF02518 and PF00072, respectively. The *Ralstonia solanacearum* GMI1000 genome was included as an outgroup.

Phylogenetic tree

The phylogenetic analysis of the bacterial species was performed through multilocus sequence analysis using a concatenated data set of *gltA*, *recA*, *rpoA*, *gyrA* and *gyrB* genes. The phylogenetic tree was obtained using the Maximum Likelihood method based on the JTT (Jones-Taylor-Thornton) matrix-based model (Jones *et al*., 1992). The percentage of trees in which the associated taxa clustered in the bootstrap test (500 replicates) is shown next to the branches in Fig. 1 (Felsenstein, 1985). Initial tree(s) for the heuristic search were obtained automatically using Neighbor-Join and BioNJ

algorithms to a matrix of pairwise distances estimated using a JTT model and then selecting the topology with superior log likelihood value. The analysis involved 34 amino acid sequences. All positions containing gaps and missing data were eliminated. A total of 2739 positions were included in the final data set. Multiple alignments and evolutionary analyses were conducted using MEGA5 software (Tamura *et al.*, 2011).

Microbiological methods and illumination conditions

The strains used in this study are described in Table S3. The Pto strains were grown at 28°C in King's B (KB) medium (King *et al.*, 1954). *Escherichia coli* strains were grown in Luria broth at 37°C. The antibiotics were used at the following final concentrations (μg ml^{-1}): ampicillin, 100; kanamycin, 20; gentamicin, 5; streptomycin, 20; rifampicin, 50. The bacterial growth was monitored at OD_{600} (optical density at 600 nm), using a Jenway 6300 spectrophotometer. White light from fluorescent tubes (cool daylight) was used at intensities of 15, 50, 55, 60 and 70 μE m^{-2} s^{-1}. Red and blue light from GreenPower LED HF red (660 nm) and GreenPower LED HF blue (470 nm) (PHILIPS, Amsterdam, the Netherlands), respectively, were used at an intensity of 30 μE m^{-2} s^{-1}. The effects of different colour lights on bacterial growth were assessed both in solid and liquid media. One millilitre of an exponential-phase (0.6 OD_{600}) culture was spread on the surface of 0.3% (wt vol^{-1}) KB agar plates. The plates were incubated at 28°C overnight under white (70 μE m^{-2} s^{-1}), red (20 μE m^{-2} s^{-1}) or blue light (20 μE m^{-2} s^{-1}), or maintained in darkness. Subsequently, the plates were carefully washed with 10 mM $MgCl_2$, and the bacterial growth was determined through serial dilution and plating. Pto was grown in liquid KB medium at 150 r.p.m. for 24 h at 28°C under the different light conditions as described earlier. Samples were collected every 90 min. For the Pto strains used in this study, an OD_{600} of 1.0 corresponded to 10^9 colony-forming units (cfu). Darkness conditions were examined using two different procedures with similar results: (i) plates and flasks were maintained in darkness covered with two layers of aluminum foil, and (ii) the plates and flasks were maintained in darkness inside a dark box. The temperature was monitored at all location for all experiments to ensure that the test conditions did not affect the temperature selected in the chamber.

DNA manipulation

The plasmids used in this study are described in Table S3. Briefly, to generate the LOV-Pto mutant, an internal 828 bp sequence of PSPTO_2896 (corresponding to nucleotides 40–868 of the coding sequence) from the Pto genome was amplified using Pfu DNA polymerase (Biotools, Madrid, Spain) and cloned into pKNG101 (Kaniga *et al.*, 1991). The resulting plasmid was transferred into Pto via triparental mating (de Lorenzo and Timmis, 1994). As pKNG101 cannot replicate in Pto, single crossover integrants were selected through streptomycin resistance. Plasmid integration was confirmed through PCR and Southern blot analyses.

For mutant complementation, the PSPTO_2896 gene was amplified from the Pto genome using primers that included the four basepair sequence (CACC) at the 5′ end of the forward primer necessary for the directional cloning into pENTR™/SD/D TOPO (Invitrogen, Carlsbad, CA, USA). The expression clones were generated through the recombination of the pENTR plasmid with the pCPP5040 expression vector using LR clonase (Invitrogen). The recombination reactions were performed according to the manufacturer's instructions. All expression constructs were confirmed through sequence analysis. The resulting plasmid was transferred into Pto via triparental mating (de Lorenzo and Timmis, 1994).

Swarming motility assays

The Pto strains were grown in KB medium to the exponential phase (OD_{600} = 0.5) under dark conditions. Subsequently, the cultures were centrifuged for 10 min at 6000 *g* at room temperature, and the cell pellets were homogenized with a sterile toothpick. Soft KB agar plates (0.3% wt vol^{-1}) were inoculated using a toothpick and incubated for 16 h at 28°C under different white light illumination conditions (15, 50, 55, 60 and 70 μE m^{-2} s^{-1}) or covered with two layers of aluminum foil under dark conditions. For the light-colour phototaxis assays, the plates were inoculated and incubated, as described earlier, under white (70 μE m^{-2} s^{-1}), red (20 μE m^{-2} s^{-1}) or blue light (20 μE m^{-2} s^{-1}) or covered with two layers of aluminum foil for dark conditions. Six plates for each treatment were incubated in three independent experiments.

Leaf attachment and confocal microscopy

For the leaf attachment assays, the Pto strains were grown for 24 h in KB agar plates under dark conditions. The cells were scraped from the plates and suspended in 10 mM $MgCl_2$ at a final concentration of 10^8 cfu ml^{-1}. A 200 μl drop was placed on the underside of the detached leaves from 3 week old *A. thaliana* plants. This process was conducted in dark conditions. Subsequently, three different experiments were set up: three replicates were incubated for 6 h under constant light (white, blue or red), three replicates were incubated for 10 min under light (white, blue or red), followed by 6 h under dark conditions, and three replicates were incubated for 6 h in the darkness. Before confocal microscopy, the leaves were washed twice through immersion in 10 mM $MgCl_2$. The GFP signal and chlorophyll autofluorescence were collected on a Leica TCS SP8 confocal microscope (Leica Microsystems, Wetzlar, Germany) using laser lines of 488 and 633 nm, respectively. The same gain and offset settings were used for the different treatments. The images were processed using the LAS AF Lite 3.1.0 (Leica Microsystems). Three leaves were inoculated for each treatment in three independent experiments.

Plant bioassays

For plant inoculations, the Pto strains were grown for 24 h on KB agar plates under dark conditions. The cells were scraped from the plates and suspended in 10 mM $MgCl_2$ at the desired final concentration for each assay. This process was conducted under dark conditions (less than 0.05 μE m^{-2} s^{-1}). Prior to inoculation, the bacterial suspensions were either subjected to light treatment (white, red or blue light) for

10 min or maintained in the darkness. For tomato infection assays, Silwet L-77 at a final concentration of 0.02% vol/vol was added to the bacterial cell suspensions. All plants were incubated in a growth chamber at 25°C with 12 h light (100 μE m^{-2} s^{-1}) and a relative humidity of 60%.

For the virulence assays in tomato, 10 3 week old tomato plants (*Solanum lycopersicum* cv Moneymaker) were inoculated by dipping for 30 s into a bacterial suspension containing 10^8 cfu ml^{-1}. The symptoms were recorded at 6 dpi. The numbers of lesions on five equivalent tomato leaflets (with regard to age and location on plants) were counted. Bacterial growth in tomato leaves was measured by determining the average number of cfu isolated from five infected leaves. The plant material was homogenized in 10 mM $MgCl_2$, and serial dilutions were plated on KB solid medium containing rifampicin.

For virulence assays in *Arabidopsis*, 4 week old plants of *A. thaliana* Col-0 were inoculated. For the spraying assays, 25 plants were inoculated for each treatment with a suspension containing 3×10^8 cfu ml^{-1} and three plants were mock-sprayed with 10 mM $MgCl_2$. The symptoms were recorded at 6 dpi. For the infiltration assays, 25 leaves were infiltrated with a suspension containing 10^5 cfu ml^{-1}, and three leaves were mock-infiltrated with 10 mM $MgCl_2$. The symptoms were recorded at 2 dpi. In both kinds of experiments, bacterial growth on the leaves was determined by measuring the average number of cfu isolated from three leaves. The plant material was processed as described earlier. To assess the dark recovery of virulence, the same bacterial suspension was subjected to a 10 min white light treatment, followed by a 100 min darkness treatment and subsequently treated again with 10 min of white light. After each treatment, an aliquot of the inoculum was used to challenge the plants as described earlier. All plant bioassays were conducted three times.

RNA isolation and qRT-PCR analysis

Flasks with 40 ml of KB were inoculated with Pto cells to $OD_{600} = 0.05$ and grown with shaking at 200 r.p.m. at 28°C under dark conditions to $OD_{600} = 0.5$. The sample was split into two. Twenty millilitres were incubated for 10 min under white light with an illuminance of 70 μE m^{-2} s^{-1}, and the other 20 ml were covered with two layers of aluminum foil for dark treatment. RNA from three biological replicates from each treatment were obtained and cleaned with RNAeasy kit (QIAGEN, Venlo, the Netherlands). RNA was quantified using a NanoDrop spectrophotometer (NanoDrop, Wilmington, DE, USA) and quality checked using an Agilent Bioanalyser 2100 (Agilent Technologies, Santa Clara, CA, USA). Bacterial RNA (1 μg) was used for cDNA synthesis using random hexamers with the high-capacity cDNA reverse transcription kit, according to the manufacturer's instructions (Applied Biosystems, Foster City, CA, USA). Specific primers were designed to amplify fragments of approximately 100 bp for all genes (Table S4) using Primer3Plus. The *rpoD* gene, which has been previously described as a housekeeping gene for *P. syringae* (Sawada *et al.*, 1999), was chosen as an internal standard. The qRT-PCR was performed in an ABI PRISM 7300 Real-Time PCR System (Applied Biosystems) using a SYBR Green PCR master mix (Applied Biosystems) according to the manufacturer's instructions. The following thermocycling conditions were used: 1 cycle at 95°C for 10 min; 50 cycles at 95°C for 15 s and 60°C for 1 min; and 1 cycle at 95°C for 15 s, 60°C for 30 s and 95°C for 15 s. The relative expression ratio was calculated as the differences between the cycle threshold (Benjamini *et al.*, 2001) values using the equation $2^{-\Delta\Delta Ct}$ as previously described (Pfaffl, 2001; Rotenberg *et al.*, 2006). A melting curve was performed at the end of each assay to certify the absence of primer-dimers and the presence of a single PCR product.

Protein extraction and Western blots

Flagellin from Pto cultures incubated 6 h under white light or dark conditions was purified, as previously described (Capdevila *et al.*, 2004). The preparations were separated on 15% SDS/PAGE (sodium dodecyl sulfate/Polyacrylamide gel electrophoresis) and stained with Coomassie Brilliant Blue R-250. The same electrophoretic conditions were used for the Western blot analysis. Flagellin antiserum (anti-FliC) was used at 1: 20 000. The serum is an anti-FliC polyclonal antibody raised against the *P. fluorescens* F113 FliC protein.

Bacterial flagella stain

Flagella of bacteria taken directly from swarming plates incubated at 28°C overnight under white light (70 μE m^{-2} s^{-1}) or dark conditions were stained as previously described (Clark, 1976). Samples were examined with the oil immersion objective in an optical microscope (Axiophot, Zeiss, Jena, Germany).

Acknowledgements

This research has been supported by project grant AGL2009-12757 (from the Ministerio de Ciencia e Innovación, Spain) and AGL2012-32516 (from the Ministerio de Economía y Competitividad, Spain). J. J. Rodríguez-Herva was funded by the 'Ramón y Cajal' program from the Ministerio de Ciencia e Innovación, Spain (RYC-2007-01045). I. Río-Álvarez was supported by the FPI program (Ministerio de Ciencia e Innovación, Spain). P. M. Martínez was supported by the Campus de Excelencia Internacional Andalucía Tech and AGL2011-30343-C02-01. We acknowledge Rafael Rivilla for kindly providing the anti-FliC serum and Alan Collmer for kindly providing the Pto *fliC* mutant strain. We thank Matilde Barón and Robert Jackson for critical reading and helpful comments.

References

Alfano, J.R., and Collmer, A. (1997) The type III (Hrp) secretion pathway of plant pathogenic bacteria: trafficking harpins, Avr proteins, and death. *J Bacteriol* **179:** 5655–5662.

Antunez-Lamas, M., Cabrera-Ordonez, E., Lopez-Solanilla, E., Raposo, R., Trelles-Salazar, O., Rodriguez-Moreno, A., *et al.* (2009) Role of motility and chemotaxis in the pathogenesis of *Dickeya dadantii* 3937 (ex *Erwinia chrysanthemi* 3937). *Microbiology* **155:** 434–442.

Armitage, J.P., and Hellingwerf, K.J. (2003) Light-induced behavioral responses ('phototaxis') in prokaryotes. *Photosynth Res* **76:** 145–155.

Barkovits, K., Harms, A., Benkartek, C., Smart, J.L., and Frankenberg-Dinkel, N. (2008) Expression of the phytochrome operon in *Pseudomonas aeruginosa* is dependent on the alternative sigma factor RpoS. *FEMS Microbiol Lett* **280:** 160–168.

Barkovits, K., Schubert, B., Heine, S., Scheer, M., and Frankenberg-Dinkel, N. (2011) Function of the bacteriophytochrome BphP in the RpoS/Las quorum-sensing network of *Pseudomonas aeruginosa*. *Microbiology* **157:** 1651–1664.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001) Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **125:** 279–284.

Berti, A.D., Greve, N.J., Christensen, Q.H., and Thomas, M.G. (2007) Identification of a biosynthetic gene cluster and the six associated lipopeptides involved in swarming motility of *Pseudomonas syringae* pv. tomato DC3000. *J Bacteriol* **189:** 6312–6323.

Bhardwaj, V., Meier, S., Petersen, L.N., Ingle, R.A., and Roden, L.C. (2011) Defence responses of Arabidopsis thaliana to infection by *Pseudomonas syringae* are regulated by the circadian clock. *PLoS ONE* **6:** e26968.

Block, A., and Alfano, J.R. (2011) Plant targets for *Pseudomonas syringae* type III effectors: virulence targets or guarded decoys? *Curr Opin Microbiol* **14:** 39–46.

Bonomi, H.R., Posadas, D.M., Paris, G., Carrica Mdel, C., Frederickson, M., Pietrasanta, L.I., *et al*. (2012) Light regulates attachment, exopolysaccharide production, and nodulation in *Rhizobium leguminosarum* through a LOV-histidine kinase photoreceptor. *Proc Natl Acad Sci USA* **109:** 12135–12140.

Briggs, W.R. (2007) The LOV domain: a chromophore module servicing multiple photoreceptors. *J Biomed Sci* **14:** 499–504.

Cai, R., Lewis, J., Yan, S., Liu, H., Clarke, C.R., Campanile, F., *et al*. (2011) The plant pathogen *Pseudomonas syringae* pv. tomato is genetically monomorphic and under strong selection to evade tomato immunity. *PLoS Pathog* **8:** e1002130.

Cao, Z., Buttani, V., Losi, A., and Gartner, W. (2008) A blue light inducible two-component signal transduction system in the plant pathogen *Pseudomonas syringae* pv. tomato. *Biophys J* **94:** 897–905.

Capdevila, S., Martinez-Granero, F.M., Sanchez-Contreras, M., Rivilla, R., and Martin, M. (2004) Analysis of *Pseudomonas fluorescens* F113 genes implicated in flagellar filament synthesis and their role in competitive root colonization. *Microbiology* **150:** 3889–3897.

Clark, W.A. (1976) A simplified Leifson flagella stain. *J Clin Microbiol* **3:** 632–634.

Cuppels, D. (1988) Chemotaxis by *Pseudomonas syringae* pv. tomato. *Appl Environ Microbiol* **54:** 629–632.

Danhorn, T., and Fuqua, C. (2007) Biofilm formation by plant-associated bacteria. *Annu Rev Microbiol* **61:** 401–422.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* **14:** 755–763.

Elias-Arnanz, M., Padmanabhan, S., and Murillo, F.J. (2011) Light-dependent gene regulation in nonphototrophic bacteria. *Curr Opin Microbiol* **14:** 128–135.

Felsenstein, J. (1985) Confidence limits on phylogenies – an approach using the bootstrap. *Evolution* **39:** 783–791.

Gohre, V., and Robatzek, S. (2008) Breaking the barriers: microbial effector molecules subvert plant immunity. *Annu Rev Phytopathol* **46:** 189–215.

Gomelsky, M., and Hoff, W.D. (2011) Light helps bacteria make important lifestyle decisions. *Trends Microbiol* **19:** 441–448.

Haefele, D.M., and Lindow, S.E. (1987) Flagellar motility confers epiphytic fitness advantages upon *Pseudomonas syringae*. *Appl Environ Microbiol* **53:** 2528–2533.

Hirano, S.S., and Upper, C.D. (2000) Bacteria in the leaf ecosystem with emphasis on *Pseudomonas syringae* – a pathogen, ice nucleus, and epiphyte. *Microbiol Mol Biol Rev* **64:** 624–653.

van der Horst, M.A., Key, J., and Hellingwerf, K.J. (2007) Photosensing in chemotrophic, non-phototrophic bacteria: let there be light sensing too. *Trends Microbiol* **15:** 554–562.

Ichinose, Y., Shimizu, R., Ikeda, J., Taguchi, F., Marutani, M., Mukaihara, T., *et al*. (2003) Need for flagella for complete virulence of *Pseudomonas syringae* pv. tabaci: genetic analysis with flagella-defective mutants ?fliC and ?fliD in host tobacco plants. *J Gen Plant Pathol* **69:** 244–249.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8:** 275–282.

Kangasjarvi, S., Neukermans, J., Li, S., Aro, E.M., and Noctor, G. (2012) Photosynthesis, photorespiration, and light signalling in defence responses. *J Exp Bot* **63:** 1619–1636.

Kaniga, K., Delor, I., and Cornelis, G.R. (1991) A wide-host-range suicide vector for improving reverse genetics in gram-negative bacteria: inactivation of the *blaA* gene of *Yersinia enterocolitica*. *Gene* **109:** 137–141.

Keith, R.C., Keith, L.M., Hernandez-Guzman, G., Uppalapati, S.R., and Bender, C.L. (2003) Alginate gene expression by *Pseudomonas syringae* pv. tomato DC3000 in host and non-host plants. *Microbiology* **149:** 1127–1138.

King, E.O., Ward, M.K., and Raney, D.E. (1954) Two simple media for the demonstration of pyocyanin and fluorescein. *J Lab Clin Med* **44:** 301–307.

Kraiselburd, I., Alet, A.I., Tondo, M.L., Petrocelli, S., Daurelio, L.D., Monzon, J., *et al*. (2012) A LOV protein modulates the physiological attributes of *Xanthomonas axonopodis* pv. citri relevant for host plant colonization. *PLoS ONE* **7:** e38226.

Lamparter, T., Michael, N., Mittmann, F., and Esteban, B. (2002) Phytochrome from *Agrobacterium tumefaciens* has unusual spectral properties and reveals an N-terminal chromophore attachment site. *Proc Natl Acad Sci USA* **99:** 11628–11633.

de Lorenzo, V., and Timmis, K.N. (1994) Analysis and construction of stable phenotypes in gram-negative bacteria with Tn*5*- and Tn*10*-derived minitransposons. *Methods Enzymol* **235:** 386–405.

Losi, A., and Gartner, W. (2008) Bacterial bilin- and flavin-binding photoreceptors. *Photochem Photobiol Sci* **7:** 1168–1178.

Losi, A., and Gartner, W. (2011) Old chromophores, new photoactivation paradigms, trendy applications: flavins in blue light-sensing photoreceptors. *Photochem Photobiol* **87:** 491–510.

Oberpichler, I., Rosen, R., Rasouly, A., Vugman, M., Ron, E.Z., and Lamparter, T. (2008) Light affects motility and infectivity of *Agrobacterium tumefaciens*. *Environ Microbiol* **10:** 2020–2029.

Ondrusch, N., and Kreft, J. (2011) Blue and red light modulates SigB-dependent gene transcription, swimming motility and invasiveness in *Listeria monocytogenes*. *PLoS ONE* **6:** e16151.

Panopoulos, N., and Schroth, M.N. (1974) Role of flagellar motility in the invasion of bean leaves by *Pseudomonas phaseolicola*. *Phytopathology* **64:** 1389–1397.

Pfaffl, M.W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* **29:** e45.

Preston, G.M. (2000) *Pseudomonas syringae* pv. tomato: the right pathogen, of the right plant, at the right time. *Mol Plant Pathol* **1:** 263–275.

Purcell, E.B., and Crosson, S. (2008) Photoregulation in prokaryotes. *Curr Opin Microbiol* **11:** 168–178.

Quiñones, B., Dulla, G., and Lindow, S.E. (2005) Quorum sensing regulates exopolysaccharide production, motility, and virulence in *Pseudomonas syringae*. *Mol Plant Microbe Interact* **18:** 682–693.

Roden, L.C., and Ingle, R.A. (2009) Lights, rhythms, infection: the role of light and the circadian clock in determining the outcome of plant-pathogen interactions. *Plant Cell* **21:** 2546–2552.

Rotenberg, D., Thompson, T.S., German, T.L., and Willis, D.K. (2006) Methods for effective real-time RT-PCR analysis of virus-induced gene silencing. *J Virol Methods* **138:** 49–59.

Rottwinkel, G., Oberpichler, I., and Lamparter, T. (2010) Bathy phytochromes in rhizobial soil bacteria. *J Bacteriol* **192:** 5124–5133.

Sawada, H., Suzuki, F., Matsuda, I., and Saitou, N. (1999) Phylogenetic analysis of *Pseudomonas syringae* pathovars suggests the horizontal gene transfer of *argK* and the evolutionary stability of hrp gene cluster. *J Mol Evol* **49:** 627–644.

Shah, R., Schwach, J., Frankenberg-Dinkel, N., and Gartner, W. (2012) Complex formation between heme oxygenase and phytochrome during biosynthesis in *Pseudomonas syringae* pv. tomato. *Photochem Photobiol Sci* **11:** 1026–1031.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12:** 1611–1618.

Swartz, T.E., Tseng, T.S., Frederickson, M.A., Paris, G., Comerci, D.J., Rajashekara, G., *et al*. (2007) Blue-light-activated histidine kinases: two-component sensors in bacteria. *Science* **317:** 1090–1093.

Takeuchi, K., Taguchi, F., Inagaki, Y., Toyoda, K., Shiraishi, T., and Ichinose, Y. (2003) Flagellin glycosylation island in *Pseudomonas syringae* pv. glycinea and its role in host specificity. *J Bacteriol* **185:** 6658–6665.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28:** 2731–2739.

Thorington, L. (1985) Spectral, irradiance, and temporal aspects of natural and artificial light. *Ann N Y Acad Sci* **453:** 28–54.

Verstraeten, N., Braeken, K., Debkumari, B., Fauvart, M., Fransaer, J., Vermant, J., *et al*. (2008) Living on a surface: swarming and biofilm formation. *Trends Microbiol* **10:** 496–506.

Vorholt, J.A. (2012) Microbial life in the phyllosphere. *Microbiology* **10:** 828–840. Nature Reviews

Wang, W., Barnaby, J.Y., Tada, Y., Li, H., Tor, M., Caldelari, D., *et al*. (2011) Timing of plant immune responses by a central circadian regulator. *Nature* **470:** 110–114.

Yao, J., and Allen, C. (2006) Chemotaxis is required for virulence and competitive fitness of the bacterial wilt pathogen *Ralstonia solanacearum*. *J Bacteriol* **188:** 3697–3708.

Yu, J., Penaloza-Vazquez, A., Chakrabarty, A.M., and Bender, C.L. (1999) Involvement of the exopolysaccharide alginate in the virulence and epiphytic fitness of *Pseudomonas syringae* pv. syringae. *Mol Microbiol* **33:** 712–720.

Yu, X., Lund, S.P., Scott, R.A., Greenwald, J.W., Records, A.H., Nettleton, D., *et al*. (2013) Transcriptional responses of *Pseudomonas syringae* to growth in epiphytic versus apoplastic leaf sites. *Proc Natl Acad Sci USA* **110:** E425–E434.

Ziegelhoffer, E.C., and Donohue, T.J. (2009) Bacterial responses to photo-oxidative stress. *Nat Rev Microbiol* **7:** 856–863.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Pto swarming motility depends on light intensity. KB agar plates (0.3%) were inoculated with the WT strain using a sterile toothpick. The plates were incubated for 16 h at 28°C under different light conditions. Similar results were obtained in three independent experiments.

Fig. S2. Different light conditions do not affect Pto bacterial growth. The light conditions included are: darkness, white light (70 μE m^{-2} s^{-1}), red light (20 μE m^{-2} s^{-1}) and blue light (20 μE m^{-2} s^{-1}). (A) Bacterial growth curves in KB liquid medium. (B) Bacterial population on KB soft-agar plates recorded after an overnight incubation. The values are the means and standard errors of three replicates. Similar results were obtained in three independent experiments.

Table S1. Bioinformatics search for photosensory proteins among *Pseudomonas*. Bioinformatics search for LOV, HK, HATPase and RR domains among *Pseudomonas* genomes using Hidden Markov Models. Gene ID from the ASAP database except for (*) which corresponds to the NCBI database.

Table S2. Bioinformatics search for bacteriophytochromes among *Pseudomonas*. Bioinformatics search for PHY domains among *Pseudomonas* genomes using Hidden Markov Models. Gene ID from the ASAP database except for (*) which correspond to the NCBI database.

Table S3. Bacterial strains and plasmids used in this study.

Table S4. qRT-PCR primers used in this study.

Molecular Plant Pathology

MOLECULAR PLANT PATHOLOGY DOI: 10.1111/mpp.12227

Role of *Dickeya dadantii* 3937 chemoreceptors in the entry to Arabidopsis leaves through wounds

ISABEL RÍO-ÁLVAREZ[1,2], CRISTINA MUÑOZ-GÓMEZ[1,2], MARIELA NAVAS-VÁSQUEZ[1,2], PEDRO M. MARTÍNEZ-GARCÍA[3], MARÍA ANTÚNEZ-LAMAS[2], PABLO RODRÍGUEZ-PALENZUELA[1,2] AND EMILIA LÓPEZ-SOLANILLA[1,2,*]

[1]*Centro de Biotecnología y Genómica de Plantas (CBGP), Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Parque Científico y Tecnológico de la UPM, Universidad Politécnica de Madrid, Campus de Montegancedo, 28223 Pozuelo de Alarcón, Madrid, Spain*

[2]*Departamento de Biotecnología, Escuela Técnica Superior de Ingenieros Agrónomos, UPM, Avda, Complutense S/N, 28040 Madrid, Spain*

[3]*Área de Genética, Facultad de Ciencias, Instituto de Hortofruticultura Subtropical y Mediterránea 'La Mayora' (IHSM-UMA-CSIC), Universidad de Málaga, E-29071 Málaga, Spain*

SUMMARY

Chemotaxis enables bacteria to move towards an optimal environment in response to chemical signals. In the case of plant-pathogenic bacteria, chemotaxis allows pathogens to explore the plant surface for potential entry sites with the ultimate aim to prosper inside plant tissues and to cause disease. Chemoreceptors, which constitute the sensory core of the chemotaxis system, are usually transmembrane proteins which change their conformation when sensing chemicals in the periplasm and transduce the signal through a kinase pathway to the flagellar motor. In the particular case of the soft-rot pathogen *Dickeya dadantii* 3937, jasmonic acid released in a plant wound has been found to be a strong chemoattractant which drives pathogen entry into the plant apoplast. In order to identify candidate chemoreceptors sensing wound-derived plant compounds, we carried out a bioinformatics search of candidate chemoreceptors in the genome of *Dickeya dadantii* 3937. The study of the chemotactic response to several compounds and the analysis of the entry process to Arabidopsis leaves of 10 selected mutants in chemoreceptors allowed us to determine the implications of at least two of them (ABF-0020167 and ABF-0046680) in the chemotaxis-driven entry process through plant wounds. Our data suggest that ABF-0020167 and ABF-0046680 may be candidate receptors of jasmonic acid and xylose, respectively.

Keywords: chemoreceptors, *Dickeya dadantii* 3937, entry, jasmonic acid, xylose.

INTRODUCTION

Dickeya dadantii 3937 (*Dd*3937) is an enterobacterium included in the Top 10 List of phytopathogenic bacteria. This catalogue is based on the scientific/economic importance of the plant bacterial pathogens included (Mansfield *et al*., 2012). *Dickeya dadantii* causes soft rot in a wide range of plant species, including important crops, such as potato (Charkowski *et al*., 2012). Several mechanisms are needed for the adaptation and virulence of *D. dadantii* inside the plant. The most important virulence factor is the production of a large set of enzymes and isoenzymes (pectate lyases, polygalacturonases, pectin methylesterases, etc.) that disassemble the plant cell wall (Barras *et al*., 1994). Other processes involved in the adaptation of *D. dadantii* to the hostile plant conditions are as follows: (i) iron uptake, for which *D. dadantii* synthesizes two siderophores: chrysobactin and achromobactin (Munzinger *et al*., 2000; Persmark *et al*., 1989); (ii) mechanisms to counteract the presence of plant toxic substances, such as multidrug resistance (MDR) efflux pumps (Barabote *et al*., 2003; Maggiorani Valecillos *et al*., 2006); (iii) other mechanisms to overcome the presence of antimicrobial compounds (Costechareyre *et al*., 2013; Llama-Palacios *et al*., 2003, 2005; López-Solanilla *et al*., 1998; 2001; Rio-Alvarez *et al*., 2012). Moreover, the *hrp*-encoded type III secretion system (T3SS) is required for the full virulence of *D. dadantii* (Bauer *et al*., 1994; Yang *et al*., 2002; Yap *et al*., 2005).

All the above-cited mechanisms are needed for the adaptation and virulence of *D. dadantii* inside the plant. Nevertheless, bacteria spend the majority of their life outside the plant and bacterial entry into the plant is critical. The life cycle of *D. dadantii* during the infection of potatoes requires adhesion to the plant surface and penetration into the plant tissues, either via wound sites or through natural openings, such as stomata (Reverchon and Nasser, 2013). For this, an active bacterial movement towards entry sites is required. Chemotaxis, which enables bacterial cells to move towards certain stimuli and away from others (Porter *et al*., 2011), is crucial for the colonization of the host and for the establishment of a successful infection (Moens and Vanderleyden, 1996).

The bacterial chemotaxis system is the best-studied biological gradient sensor and has been analysed in detail in *Escherichia coli*

**Correspondence*: Email: emilia.lopez@upm.es

(Sourjik and Wingreen, 2012). The system is composed of bacterial chemoreceptors or methyl-accepting chemotaxis proteins (MCPs). Bacterial chemoreceptors are hexagonally packed trimers of receptor dimers (Briegel *et al*., 2012) that recognize specific chemicals (even redox state) at the periplasm and regulate the flagellar function, assisted by a signal transduction pathway composed of the so-called Che proteins (Hazelbauer and Lai, 2010). The structure of a typical bacterial chemoreceptor is as follows: (i) a sensor domain or ligand-binding region (LBR), which can be either periplasmic or cytoplasmic, and which detects chemical signals and induces a conformational change; (ii) a HAMP-type adaptation domain (cytosolic), which transmits the molecular stimulus created by ligand binding at the LBR to the: (iii) MCP domain or signalling domain (cytosolic), which undergoes reversible methylation at multiple sites and transduces the signal to the downstream signalling cascade of Che proteins (Falke and Hazelbauer, 2001). Despite the fact that the chemotaxis system is conserved among bacteria, chemoreceptors differ largely in their protein topology. A large-scale comparative genomic analysis of the MCP signalling and adaptation domain family (from 312 prokaryotic genomes) allowed the identification of seven major chemoreceptor classes and three distinct structural regions within the cytoplasmic domain: signalling, methylation and flexible bundle subdomains (Alexander and Zhulin, 2007). Moreover, although most LBRs are periplasmic, others are cytoplasmic, and sometimes the LBR is lacking in some chemoreceptors (Lacal *et al*., 2010a). Much of the knowledge about the ligands for bacterial chemoreceptors comes from studies in *E. coli*, which has four chemoreceptors (Tar, Tsr, Trg and Tap) and an aeroreceptor (Aer). The ligand profile of each of the four chemoreceptors has been determined: the Tar receptor is primarily for aspartate and maltose, the Tsr receptor for serine and leucine, the Trg receptor for ribose and galactose, and the Tap receptor for dipeptides and pyrimidines (Lacal *et al*., 2010a). The Aer receptor, which lacks the periplasmic domain, mediates aerotaxis in response to oxygen-related cellular redox changes (Bibikov *et al*., 2004). Although there is little information on the ligands for other bacterial chemoreceptors, a considerable number of studies have focused on this issue, not only in animal-pathogenic bacteria (Nishiyama *et al*., 2012; Rahman *et al*., 2014; Sweeney *et al*., 2013), but also in environmental bacteria (Lacal *et al*., 2010b; Nichols *et al*., 2012; Oku *et al*., 2012).

In the case of plant-associated bacteria, active chemotaxis may represent an opportunity for the establishment of an interaction with the plant, either symbiotic or pathogenic. Nevertheless, fewer studies exist on chemotaxis and chemoreceptors in plant-pathogenic bacteria. Chemotaxis is needed for virulence, biofilm formation and competitive fitness in *Ralstonia solanacearum* (Yao and Allen, 2006, 2007) and for pathogenicity in *Agrobacterium tumefaciens* (Hawes and Smith, 1989). The chemotactic behaviour towards several compounds has been described in *Pseudomonas syringae* (Cuppels, 1988) and *Xanthomonas campestris* (Kamoun and Kado, 1990), although the role of chemotaxis in the interaction with the host has not been elucidated to date. However, a particular case is *Dd*3937. The contribution of motility and chemotaxis to the pathogenicity of *Dd*3937 has been studied (Antúnez-Lamas *et al*., 2009). Genes involved in the chemotaxis transduction system (*cheW*, *B*, *Y* and *Z*) and in the structure of the flagellar motor (*motA*) are required for swimming and entry into Arabidopsis leaves. Its capacity to mediate chemoattraction and chemorepellance in response to compounds such as sugars, amino acids and plant hormones, such as jasmonic acid (JA) (Antunez-Lamas *et al*., 2009), has been assessed. JA, which participates in wound signalling in plants (León *et al*., 2001), is a strong chemoattractant for *Dd*3937. Furthermore, this perception seems to drive the ingress of this bacterium inside plant tissues through wounds. These results suggest that *Dd*3937 may have at least one chemoreceptor responsible for the perception of JA. Moreover, jasmonate-dependent modifications of the pectin matrix of the plant cell wall during potato development function as a defence mechanism targeted by *Dd*3937 virulence factors (Taurino *et al*., 2014). This finding links the plant hormone, JA, with the plant cell wall modifications. In dicot secondary cell walls, the major hemicellulose is a polymer of β-(1,4)-linked xylose units, called xylan (Rennie and Scheller, 2014), and xylose is released when the plant cell wall is altered. Therefore, it can also be hypothesized that *Dd*3937 may possess chemoreceptors to detect plant cell wall-derived compounds.

In this study, we carried out a bioinformatics analysis followed by an experimental approach to identify putative chemoreceptors involved in the perception of compounds released in plant wounds. Moreover, we analysed the role of these chemoreceptors in the entry to the plant apoplast and the colonization ability in Arabidopsis plants.

RESULTS

*Dd*3937 possesses an unusually high number of chemoreceptors among Enterobacteria

We screened the *Dd*3937 genome (http://asap.ahabs.wisc.edu/asap/home.php) for the presence of proteins containing the highly conserved MCP signalling domain. The *Dd*3937 genome encodes 47 proteins containing the MCP domain. Despite the fact that *Dd*3937 belongs to Enterobacteria, this result contrasts with the average number of MCP genes found in other Enterobacteria: 29 MCPs per genome (Lacal *et al*., 2010a). The amino acid sequences of these 47 chemoreceptors were searched for transmembrane regions (TMs) using the DAS server (Table 1). According to the presence of TMs, *Dd*3937 chemoreceptors were classified into six different topologies (Fig. 1A), as described previously (Lacal *et al*., 2010a). In the case of *Dd*3937, all topologies, including class Ia,

Table 1 Amino acid coordinates of methyl-accepting chemotaxis protein (MCP) domains and transmembrane regions (TMs) determine the topology of the *Dd*3937 chemoreceptor.

ID		TM						MCP		Topology class‡
ASAP*	UNIPROT†	i_1	j_1	i_2	j_2	i_3	j_3	i	j	
ABF-0014536	E0SHS4	10	32	181	203			334	501	Ia
ABF-0014618	E0SM30	36	58	214	232			367	531	Ia
ABF-0014722	E0SF47	7	29	190	209			344	511	Ia
ABF-0014824	E0SDP4	10	32	183	205			340	507	Ia
ABF-0015168	E0SH41	5	24	177	199			335	501	Ia
ABF-0015513	E0SBF1	20	42	200	222			360	523	Ia
ABF-0015600	E0SE76	12	34	278	295			431	598	Ia
ABF-0015603	E0SE77	38	60	300	317			451	618	Ia
ABF-0016115	E0SN86	13	35	269	285			421	588	Ia
ABF-0016585	E0SIA2	10	32	189	211			343	510	Ia
ABF-0016979	E0SEW8	10	32	183	205			343	510	Ia
ABF-0017097	E0SCJ7	17	39	272	290			427	593	Ia
ABF-0017662	E0SFJ0	5	27	187	204			342	508	Ia
ABF-0017665	E0SFJ2	36	58	213	235			372	538	Ia
ABF-0017668	E0SFJ3	10	32	184	206			344	511	Ia
ABF-0017672	E0SFJ4	10	32	184	206			347	514	Ia
ABF-0017674	E0SFJ5	20	39	201	223			359	526	Ia
ABF-0017863	E0SI11	10	32	189	211			346	510	Ia
ABF-0017896	E0SHY9	11	33	178	200			337	503	Ia
ABF-0018502	E0SEF0	10	32	187	209			346	512	Ia
ABF-0018541	E0SFF6	10	32	182	204			336	505	Ia
ABF-0018585	E0SJX6	20	42	199	221			357	524	Ia
ABF-0018754	E0SIM4	10	32	179	201			336	503	Ia
ABF-0018765	E0SIN3	20	42	204	226			356	518	Ia
ABF-0018892	E0SJ51	20	42	194	216			355	522	Ia
ABF-0019050	E0SAP6	15	37	270	292			427	591	Ia
ABF-0019306	E0SF78	13	35	282	299			435	601	Ia
ABF-0019718	E0SHT9	7	29	180	202			336	502	Ia
ABF-0019790	E0SF98	10	32	187	209			344	510	Ia
ABF-0019851	E0SM37	7	29	179	194			333	499	Ia
ABF-0019852	E0SM36	20	42	195	214			348	515	Ia
ABF-0019855	E0SM35	13	35	185	204			340	507	Ia
ABF-0019858	E0SM34	31	53	202	222			357	522	Ia
ABF-0020167	E0SDB0	10	29	178	200			340	507	Ia
ABF-0020252	E0SF13	25	47	288	314			439	604	Ia
ABF-0020431	E0SDX8	5	27	168	190			327	489	Ia
ABF-0046680	E0SMA1	58	77	348	370			504	670	Ia
ABF-0017537	E0SD21	10	32	142	147			421	586	Ia
ABF-0017419	E0SJN1	138	160					295	462	Ib
ABF-0014726	E0SF49	159	178	188	200			335	497	II
ABF-0019309	E0SF75	164	181	186	205			341	506	II
ABF-0014843	E0SDM9	147	151	165	180	185	196	337	499	III
ABF-0016380	E0SEN9	30	52	183	205	336	350	490	656	III
ABF-0018511	E0SEE2	13	30	196	203	315	330	468	633	III
ABF-0016436	E0SMM7							297	417	IVa
ABF-0017090	E0SCK2							474	615	IVa
ABF-0017824	E0SI40							17	131	IVa

*Nomenclature from the ASAP database (https://asap.ahabs.wisc.edu/asap/logon/php).
†Nomenclature from the UNIPROT database (http://www.uniprot.org/).
‡Topology class according to Lacal *et al*. (2010a).

show similar results to those reported by Lacal *et al*. (2010a). Furthermore, in the most abundant topology Ia, we grouped the LBRs according to their length (Fig. 1B). *Dd*3937 LBRs (Ia) can be clustered into two groups as described previously by Lacal *et al*. (2010a). Most LBRs (60%) are characterized by sizes corresponding to cluster I, the majority of the LBRs being those with sizes between 150 and 159 amino acids (Fig. 1B).

LBR sequences were analysed by CLUSTAL W2 (Larkin *et al*., 2007), searching for diversity among the sequences (Table S1, see Supporting Information). The results showed that the level of sequence conservation was generally low and no duplications were observed. Moreover, LBRs of each of the 47 MCP proteins in *Dd*3937 were searched for Pfam components (Finn *et al*., 2014), using a customized BioPerl (Stajich *et al*., 2002) script. Only hits

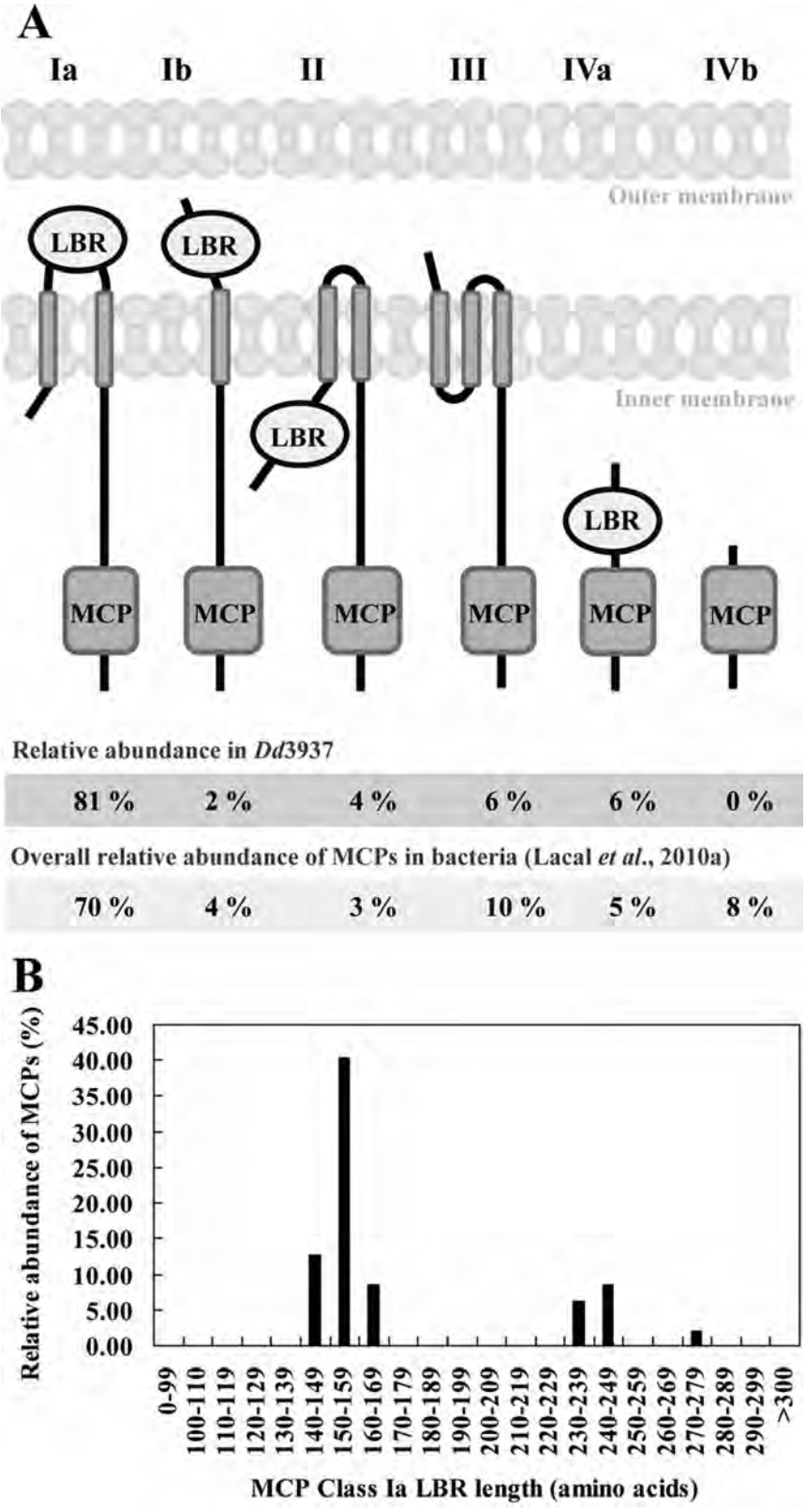


Fig. 1 (A) Classification of *Dd*3937 chemoreceptors into six different topologies (roman numerals); 47 MCP sequences from *Dd*3937 were analysed. Classification is based on clustering annotated by Lacal *et al*. (2010a). The prediction of transmembrane regions (TMs) was performed using the DAS server (Cserzo *et al*., 1997). Arabic numbers indicate the relative abundance of receptors with a given topology in *Dd*3937 and in bacteria according to Lacal *et al*. (2010a). MCP, methyl-accepting chemotaxis protein; LBR, ligand-binding region. (B) Relative abundance of *Dd*3937 MCP sequences (belonging to class Ia) according to the size of the LBR (amino acids).

with an E-value of less than or equal to 0.005 were retained (Table S2, see Supporting Information). The results showed that 23 of the LBRs fall into the annotated domain 4HB_MCP_1 (four helix bundle sensory module for signal transduction) (Ulrich and Zhulin, 2005), which is a ubiquitous sensory module in prokaryotic signal transduction. Five are homologues to the annotated domain TarH, a member of a family of transmembrane receptors that mediate the chemotactic response in certain enteric bacteria, such as *Salmonella typhimurium* and *E. coli* (Kim *et al*., 1996). Homology with other annotated domains involved in sensory processes can be found.

Global BLAST of LBRs from *Dd*3937 chemoreceptors suggests putative receptors for plant compounds

In order to identify *Dd*3937 chemoreceptors involved in the sensing of plant-specific compounds, LBRs were further analysed by bioinformatics methods. LBR sequences were extracted from the protein sequence as described in Experimental procedures. LBR sequences were subjected to local BLAST against a bacterial genome database (this database contains complete bacterial genomes available as of 14 February 2014 from GeneBank Ref Seq) using an *ad hoc* PERL script (for more details, see Experimental procedures). Table S3 (see Supporting Information) shows similar LBR sequences (E-value $\leq 10^{-30}$) to those found in other bacteria. Figure 2A shows the distribution of *Dd*3937 LBRs among the following groups of bacteria: (i) other *Dickeya* strains; (ii) plant-pathogenic bacteria; (iii) animal-pathogenic bacteria (considering the occurrence of a high number of similar LBR sequences in animal pathogens); and (iv) non-pathogenic bacteria (including bacteria present in plants, soil and water). Among the 47 LBRs of *Dd*3937, nine were widely distributed among bacteria regardless of their niche (animals, plants, soils and water). Another seven LBRs, apart from being present in plant-pathogenic bacteria, were also found in environmental (non-pathogenic) bacteria. Moreover, another four LBRs, apart from being present in *Dickeya* strains, were also found in environmental (non-pathogenic) bacteria. Only for one LBR were similar sequences found exclusively in animal and environmental bacteria. The largest group, 13 LBRs, was found in other plant-pathogenic bacteria besides the *Dickeya* genus, and another large group containing 12 members included LBRs exclusively present in the *Dickeya* genus. This means that 56% of *Dd*3937 LBRs show similarity only with other LBRs from phytopathogenic bacteria(Fig. 2B). Interestingly, one *Dd*3937 LBR

A

Topology class	ID[a]	In other *Dickeya* strains	In plant-pathogenic bacteria	In animal-pathogenic bacteria	In non-pathogenic bacteria
Ib	ABF-0014726	■	■	■	■
III	ABF-0014843	■	■	■	■
Ia	ABF-0016115	■	■	■	■
IVa	ABF-0016436	■	■	■	■
Ia	ABF-0017537	■	■	■	■
Ia	ABF-0018585	■	■	■	■
Ia	**ABF-0018754**	■	■	■	■
Ia	ABF-0018765	■	■	■	■
II	ABF-0019309	■	■	■	■
III	ABF-0016380	■	■		■
IVa	ABF-0017090	■	■		■
III	ABF-0018511	■	■		■
Ia	ABF-0018892	■	■		■
Ia	ABF-0019852	■	■		■
Ia	ABF-0019855	■	■		■
Ia	**ABF-0046680**	■	■		■
Ia	ABF-0014722	■		■	■
Ia	**ABF-0014536**	■	■		
Ia	ABF-0014824	■	■		
Ia	ABF-0015168	■	■		
Ia	ABF-0015513	■	■		
Ia	ABF-0016585	■	■		
Ia	ABF-0016979	■	■		
Ia	ABF-0017097	■	■		
Ia	**ABF-0017668**	■	■		
Ia	ABF-0017674	■	■		
Ia	ABF-0019050	■	■		
Ia	ABF-0019790	■	■		
Ia	ABF-0019851	■	■		
Ia	ABF-0018502	■	■		
Ia	**ABF-0018541**	■			■
IVa	**ABF-0017824**	■			■
Ia	ABF-0017896	■			■
Ia	ABF-0020431	■			■
Ia	ABF-0014618	■			
Ia	ABF-0015600	■			
Ia	ABF-0015603	■			
Ib	ABF-0017419	■			
Ia	**ABF-0017662**	■			
Ia	**ABF-0017665**	■			
Ia	**ABF-0017863**	■			
Ia	ABF-0017672	■			
Ia	ABF-0019306	■			
Ia	ABF-0019718	■			
Ia	ABF-0019858	■			
Ia	**ABF-0020252**	■			
Ia	**ABF-0020167**	■			

B

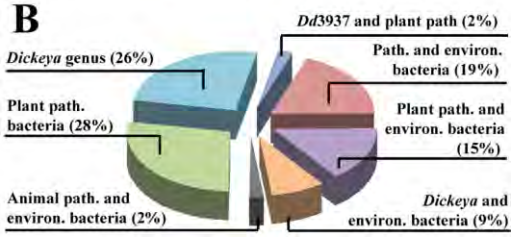


Fig. 2 (A) Distribution of *Dd*3937 ligand-binding regions (LBRs) among bacteria. [a]Nomenclature from the ASAP database (https://asap.ahabs.wisc.edu/asap/logon/php). In bold, genes selected for further analyses. Grey boxes indicate that similar sequences to a given *Dd*3937 LBR were found in the global BLAST analysis. (B) Schematic representation of the distribution of *Dd*3937 LBRs among bacteria according to their ecological niche. path., pathogenic; environ., environmental.

was not found in other *Dickeya* strains, but was found in other plant-pathogenic bacteria.

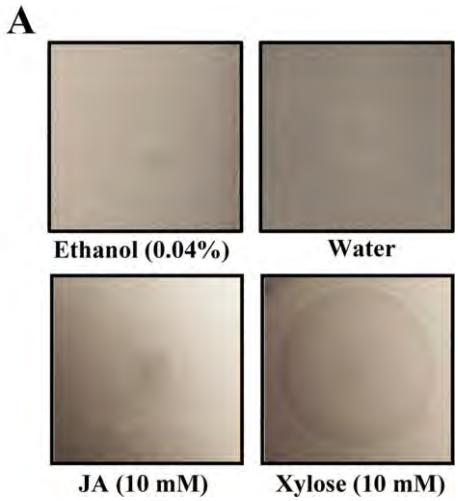
Mutants in individual chemoreceptors of *Dd*3937 are affected in the perception of JA and xylose

In order to unveil the role of the perception of plant molecules during the interaction of *Dd*3937 with plants, chemoreceptors from the different groups shown in Fig. 1 were selected. Thus, ABF-0018754 was selected as a control for chemotaxis assays as it is the candidate for serine perception in *Dd*3937. ABF-0046680 was chosen from the group of chemoreceptors whose LBRs showed similarity with those of plant-pathogenic bacteria, but also environmental bacteria. ABF-0014536 and ABF-0017668 were selected from the group of chemoreceptors whose LBRs showed similarity only with those of plant-pathogenic bacteria; ABF-0018541 and ABF-0017824 were selected from the group of chemoreceptors whose LBRs showed similarity with those of *Dickeya* strains (not with other plant-pathogenic bacteria) and environmental bacteria. ABF-0017662, ABF-0017665, ABF-0017863 and ABF-0020252 were selected from the group of chemoreceptors whose LBRs showed similarity only with those of *Dickeya* strains. Finally, ABF-0020167 was chosen because of its particularity of not showing similarity with other *Dickeya* strains, but with other plant-pathogenic bacteria. These chemoreceptors were mutagenized by different approaches (for more details, see Experimental procedures). ABF-0018754, chosen as a control for chemotaxis assays, restored serine perception when heterologously expressed in the non-chemotactic *E. coli* UU1250 strain (Bibikov *et al*., 2004) (Fig. S1, see Supporting Information).

In order to elucidate how the mutations in the chemoreceptors cited above affected the movement of *Dd*3937 towards putative entry sites, such as wounds, the chemotactic response of wild-type (WT) and mutant strains towards JA and xylose was analysed. As mentioned previously, JA secreted in plant wounds is a strong chemoattractant for *Dd*3937 (Antunez-Lamas *et al*., 2009) and xylose is released when the plant cell wall is altered.

The analysis of the chemotaxis towards these plant compounds was performed using the classical 'in-plug' assay (Tso and Adler, 1974), which is based on the ability of bacteria to swim in soft agar and move towards/away from a given chemical. The results were recorded from soft agar–bacteria plates based on either the observation of attraction haloes or of the absence of a chemotactic response (Fig. 3). Independent of the intensity of the haloes, these were considered as a positive chemoattraction when an apparent ring of accumulated cells was found at a certain distance from the plug.

From the results of the 'in-plug' assays, we selected mutants *14536*, *17824*, *18541*, *20167* and *46680* for quantitative chemotaxis assays towards JA, and mutants *18541*, *20167*, *20252* and *46680* for quantitative chemotaxis assays towards xylose. For this purpose, a Palleroni-like assay was performed using 5-μL tips instead of glass capillaries, as handling allows a high-throughput screening in comparison with the classic Palleroni assay (Palleroni, 1976). This assay allows the calculation of the chemotaxis index



B

	JA	Xylose
WT	+	+
14536	Ø	+
17662	+	+
17665	+	+
17668	+	+
17824	Ø	+
17863	+	+
18541	Ø	Ø
18754	+	+
20167	Ø	Ø
20252	+	Ø
46680	Ø	Ø

Fig. 3 (A) Typical attraction haloes of *Dd*3937 wild-type (WT) towards jasmonic acid (JA) and xylose in the 'in-plug' assay. A drop containing a given chemical, either JA (10 mM) or xylose (10 mM), or their corresponding controls, ethanol (0.04% v/v) and water, was deposited on the (0.25%) soft-agar bacteria, and haloes were recorded after 20 min. (B) Chemotactic response of *Dd*3937 strains towards JA and xylose. Chemotactic response of *Dd*3937 WT and mutant strains was evaluated in the presence of a concentration gradient of JA (10 mM) and xylose (10 mM). The chemotactic response was considered as attraction (+) or non-chemotactic response (Ø) in comparison with that of the WT strain for each chemical. Similar results were obtained in three independent experiments.

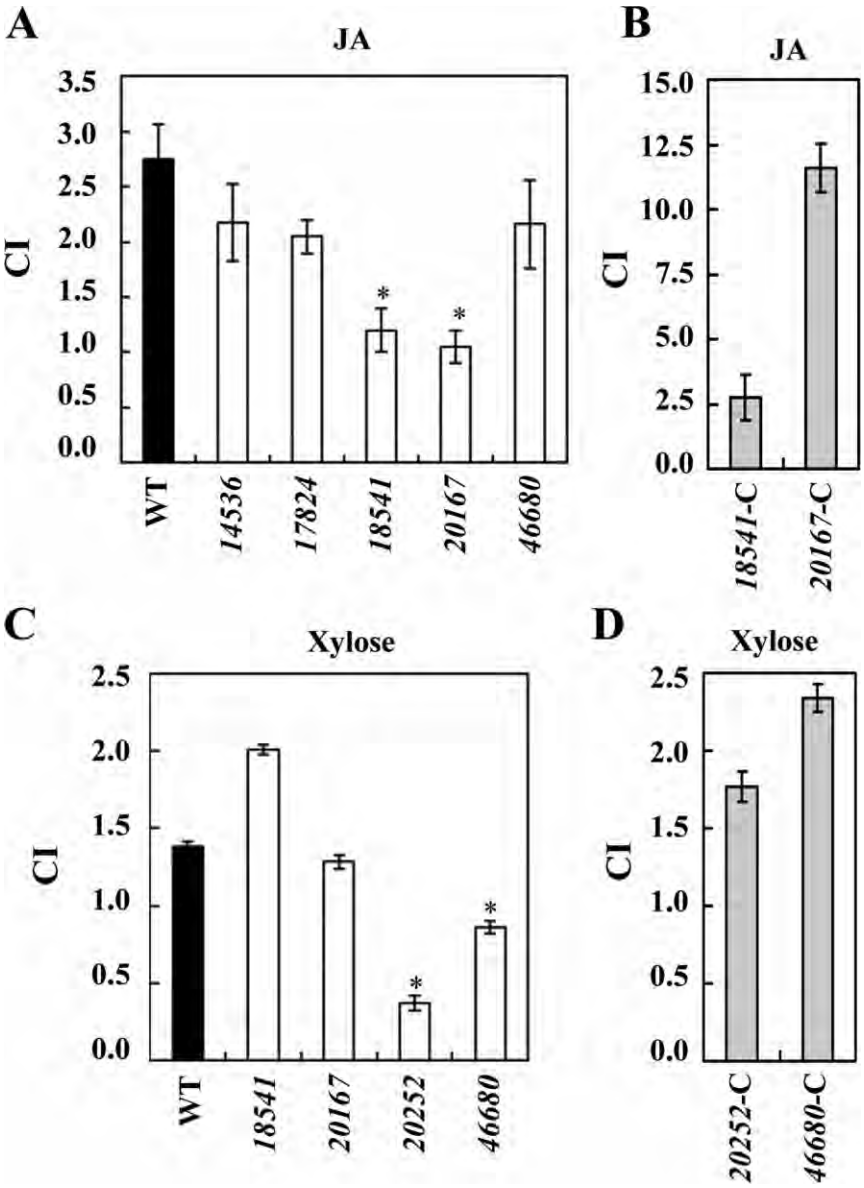


Fig. 4 Quantitative chemotaxis assays of *Dd*3937 strains towards jasmonic acid (JA) and xylose. Bars represent the chemotaxis index (CI) for wild-type (WT) and mutant strains towards 1 mM JA (A) and 1 mM xylose (C). Those mutants showing a different chemotaxis towards JA or xylose were complemented and the chemotaxis towards 1 mM JA (B) or xylose (D) was tested using the same procedure. CI is defined as the number of cells that accumulate after 20 min within a tip containing a given compound divided by the number of cells accumulated within a tip containing the solvent. The means and standard errors are shown. Error bars represent the standard error of the mean (SEM). Asterisk indicates significant differences between WT and mutant CI determined according to Student's *t*-test ($P < 0.05$). Similar results were obtained in three independent experiments.

(CI). As bacterial cells can move actively towards/away from a given chemical contained in the capillary, values of CI greater than 1.0 are considered as chemoattraction and values less than 1.0 as chemorepulsion (Fig. 4). The results in Fig. 4 indicate that, in the case of JA, only *18541* and *20167* mutants showed a reduced chemoattraction towards this plant hormone. In the case of xylose, *20252* and *46680* mutants were affected in the perception of this sugar. As a control for the quantitative (capillary) assays, chemotaxis towards serine of the WT and mutant *18754* was assayed in each experiment (data not shown). Strains overexpressing the WT chemoreceptors restored the ability to develop a chemoattraction response in the presence of the respective compound (Fig. 4B, D).

To check that these results were not the product of an altered chemotactic ability, *Dd*3937 mutants were tested for chemotaxis towards serine. All mutants, except *18754* (candidate serine receptor of *Dd*3937), showed the same chemotaxis towards serine as did the WT strain (Fig. S2, see Supporting Information). Moreover, to demonstrate that these mutants do not have a general defect in motility, we carried out a swimming assay in soft agar-rich medium. This assay, which reflects the sum of taxis towards compounds present in the medium, showed that the spreading of these strains was not altered with respect to that of the WT strain (Fig. S3, see Supporting Information).

*Dd*3937 chemoreceptors contribute in a different manner to entry into Arabidopsis leaves

As stated above, how pathogens enter plant tissues is critical for pathogenesis. To ascertain the particular relevance of the selected

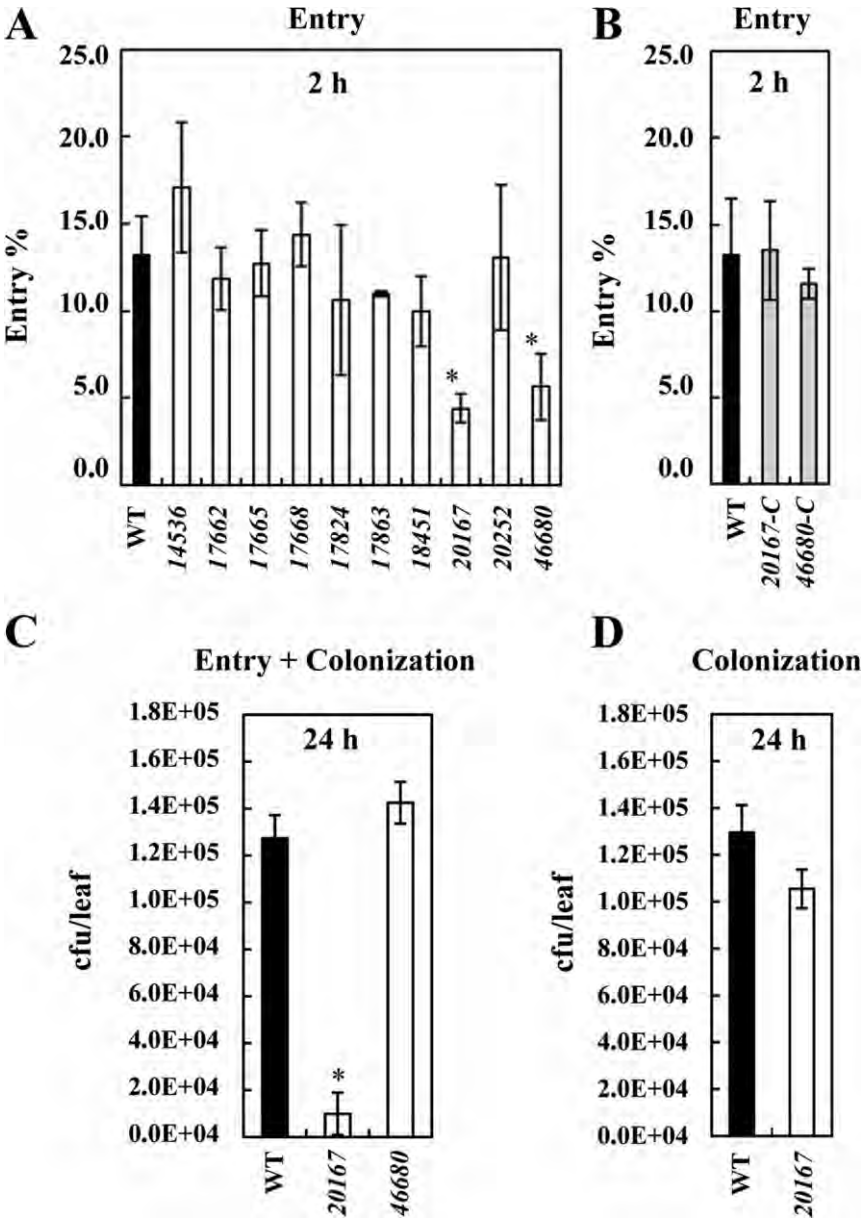


Fig. 5 (A) Entry of *Dd*3937 wild-type (WT) and mutant strains in *Arabidopsis thaliana* leaves. For each strain, 10^5 cells were placed on a wound. This process was repeated in 15 leaves from at least five different plants. Bacterial populations inside the leaf were estimated after 2 h. (B) Entry of *Dd*3937 WT and complemented strains in *A. thaliana* leaves. For each strain, 10^5 cells were placed on a wound. This process was repeated in 15 leaves from at least five different plants. Bacterial populations inside the leaf were estimated after 2 h. (C) Entry-dependent colonization assay in *A. thaliana* leaves. 10^5 cells of WT or mutant (*20167* and *46680*) strains were placed on a wound. This process was repeated in 15 leaves from at least five different plants. Bacterial populations inside the leaf were estimated after 24 h. (D) Non-entry-dependent pathogenicity assay in *A. thaliana* leaves. 2×10^6 cells of WT or *20167* strains were infiltrated into the abaxial side of leaves. This process was repeated in 15 leaves from at least five different plants. The bacterial population inside the leaf was estimated after 24 h. cfu, colony-forming unit. For (A–D), means and standard errors are shown. Error bars represent standard error of the mean (SEM). Asterisk indicates significant differences between WT and mutant population (estimated either as percentage entry or total population inside the leaves) determined according to Student's *t*-test ($P < 0.05$). Similar results were obtained in three independent experiments.

*Dd*3937 chemoreceptors in this process, and how this could affect the subsequent development of symptoms, WT and mutant strains were challenged to entry and colonization assays in Arabidopsis leaves. *Arabidopsis thaliana* has been described and characterized previously as a susceptible host of *Dd*3937 (formerly *Erwinia chrysanthemi* 3937) (Dellagi *et al*., 2005; Fagard *et al*., 2007).

To compare the ability to detect plant compounds released in a wound, we first performed entry assays in *A. thaliana* Col-0 leaves, as described previously (Antunez-Lamas *et al*., 2009). For each particular strain, the cells which entered the plant apoplast were measured relative to the initial inoculum of cells (cells recovered from the leaf/cells put on the wound; expressed as a percentage), and these ratios were used for comparison with the WT strain ratio. Figure 5A shows the differences in entry ability of the 10 MCP mutants tested in comparison with the WT strain. Two mutants, *20167* and *46680*, were significantly affected in their ability to enter the plant apoplast in contrast with the WT strain. Entry assays in *A. thaliana* leaves were performed as described previously. *20167*-C and *46680*-C complemented strains showed entry percentages close to those of the WT levels (Fig. 5B).

To ascertain whether the delay in entry into the plant apoplast exhibited by *20167* and *46680* mutants could affect the subsequent establishment of infection, the same assay described above was performed, changing the incubation time from 2 to 24 h (Fig. 5C). Figure 5C shows that the delay in entry of *20167* observed at 2 h affected the population inside the leaves at 24 h,

whereas *46680* reached WT population levels at 24 h. To check that the decrease in the *20167* population inside the leaves at 24 h was not a result of impairment in colonization ability, but a direct consequence of reduced entry, a non-entry-dependent colonization assay was performed. For this, WT and *20167* cell suspensions were infiltrated into the abaxial side of *A. thaliana* leaves and bacterial populations were estimated at 24 h. Figure 5D shows that, when infiltrated, the population of the *20167* mutant is at the same level as that of the WT strain.

DISCUSSION

The *E. coli* chemotactic pathway has provided a paradigm for sensory systems in general. However, the increasing number of sequenced bacterial genomes shows that, although the central sensory mechanism seems to be common to all bacteria, there is added complexity in a wide range of species (Krell *et al*., 2011; Wadhams and Armitage, 2004). Indeed, the analysis of bacterial genomes reveals that motile bacteria differ enormously in the number of chemoreceptors. Chemoreceptor abundance in a bacterial genome ranges from 64 in *Magnetospirillum magnetotacticum* (Alexandre *et al*., 2004) to a single chemoreceptor in other bacteria (Lacal *et al*., 2010a). Forty-seven chemoreceptors have been identified in *Dd*3937 through bioinformatics analyses. In comparison with its close Enterobacteria relatives (*Shigella*, *Salmonella*, *Yersinia*, *Escherichia* spp.), which show a small number of chemoreceptors, for instance five in *E. coli* (Lacal *et al*., 2010a; Wadhams and Armitage, 2004), *Dd*3937 possesses a large number of chemoreceptors. This is congruent with the hypothesis that microorganisms living in complex and variable environments must be able to detect a greater variability of changing conditions. Indeed, genome sequencing has revealed that many microorganisms from aquatic and soil environments possess large numbers of chemoreceptors (Alexandre *et al*., 2004). Moreover, to assess the relationship between the abundance of chemoreceptors and lifestyle, regardless of taxonomic classification, Lacal *et al*. (2010a) analysed 264 bacterial genomes for the presence of chemoreceptors, and similar results were obtained, showing that the number of chemoreceptors per genome depends mostly on bacterial lifestyle and metabolic diversity, but weakly on genome size. Among the 264 bacterial genomes analysed, only three phytopathogenic bacteria, *Xylella fastidiosa*, *Xanthomonas oryzae* pv. *oryzae* and *Agrobacterium tumefaciens* C58, were included. These authors also analysed the topology of the chemoreceptors defining several classes. In the case of *Dd*3937, the most abundant topology is that of class Ia, as it mainly occurs in bacteria (Lacal *et al*., 2010a). For the rest of the classes, similar results to those shown for bacteria were obtained in *Dd*3937. Moreover, the analysis of the LBR length in the most abundant topology class (Ia) also showed the same clustering into two defined groups described by Lacal *et al*. (2010a).

The detailed analysis of the distribution of *Dd*3937 LBRs among known bacterial genomes allowed us to determine that more than 50% of the LBRs show high similarity only with sequences from other plant-pathogenic bacteria. These *in silico*-based data suggest that these chemoreceptors would be implicated in the perception of plant-derived compounds. The large number of putative chemoreceptors involved in this process demonstrates the relevance of the perception of plant compounds during the interaction. Indeed, the identification of host–microbe interaction factors in the genome of *Dd*3937 with supervised machine learning shows that methyl-accepting chemotaxis genes are highly enriched among the predicted host–microbe interaction factors in this bacterium (Ma *et al*., 2014).

The results from the 'in-plug' chemotaxis assays indicated that more than one mutant strain showed an altered chemotactic response towards JA or xylose with respect to the WT strain. A possible explanation of these findings would be that a given compound can be detected by more than one chemoreceptor. Indeed, there are several examples in the literature showing this (Kato *et al*., 2008). The best studied example is the recognition of amino acids by the paralogous receptors PctA, PctB and PctC of *Pseudomonas aeruginosa* (Rico-Jiménez *et al*., 2013). However, the opposite phenomenon has also been described: one chemoreceptor is able to sense more than one compound. For example, McpS, a *Pseudomonas putida* KT2440 chemoreceptor for tricarboxylic acid (TCA) cycle intermediates, is involved in the chemotactic response towards different ligands, such as malate, succinate and fumarate (Lacal *et al*., 2010b). Other examples of multi-ligand chemoreceptors are as follows: Tcp receptor in *S. typhimurium*, which mediates positive chemotaxis to citrate and negative chemotaxis to phenol (Yamamoto and Imae, 1993), PA2652 of *P. aeruginosa*, which mediates chemotaxis to malate, but not to any of the remaining TCA cycle intermediates (Alvarez-Ortega and Harwood, 2007), and McpX of *Vibrio cholera*, which functions as a chemoreceptor for multiple amino acids (Nishiyama *et al*., 2012). Another hypothesis which cannot be ruled out is that a mutation in a given chemoreceptor alters the chemotactic response of the remaining receptors, through gene expression reprogramming. There is much information missing about the gene regulation of chemoreceptors. For example, it is unknown whether all receptors are constitutively expressed, whether they are inducible under certain stimuli and whether there is fine regulation depending on the metabolic needs. It has been shown that the responses of *E. coli* to two opposing chemoattractant gradients are dependent on the chemoreceptor ratio (Kalinin *et al*., 2010). Therefore, if *Dd*3937 chemoreceptor ratios are being altered through individual mutations, one can expect several changing chemotactic behaviours, as found in this study.

The analysis of the chemotactic response to xylose and JA using a quantitative methodology based on capillary assays showed

that, of the five mutants found to be altered in chemotaxis towards JA in the in-plug assay, only two were affected in the capillary assays. In the case of xylose, of the four mutants altered, only two consistently showed a reduced chemotaxis in the capillary assays. This discrepancy could be a result of the nature of the medium employed, i.e. semi-solid agar in the 'in-plug' assay and liquid medium in the capillary assay. Both the availability of the chemoattractant to be perceived and the effects of the conditions during the assay on the motility ability could be responsible for the different results obtained. An important element to take into account is that chemotaxis is a flagellum-dependent phenomenon and the number of flagella is intimately linked to the characteristics of the medium (Jarrell and McBride, 2008).

Although other chemoreceptors could be involved in JA perception, we focused our studies on the role of ABF-0018541 and ABF-0020167 in the active entry process through wounds as the corresponding mutant strains consistently showed a decreased chemotaxis towards this hormone in the two assays performed. Our experiments showed that only the *20167* mutant displayed impaired entry ability to the plant apoplast, which suggested a pivotal role for this receptor during this process. In the case of xylose, although *20252* and *46680* mutants showed reduced chemoattraction towards this sugar in both 'in-plug' and quantitative assays, only *46680* was impaired in the entry process, which also suggested a key role for this receptor during the entry to the plant apoplast. However, when analysing the colonization ability, although *20167* and *46680* mutants displayed impaired entry ability, only the *20167* strain was altered in the establishment of a sufficiently large population to prosper in the hostile plant conditions. When *20167* was infiltrated, the bacterial population reached the same level as that of the WT strain. Therefore, the altered perception of compounds produced in wounds, such as JA, is the main contributor to the impaired colonization ability of this strain, which does not seem to be related to the colonization ability or fitness inside the plant tissue. In the case of *46680*, although under laboratory conditions the mutant is not affected in its colonization ability, a role in this process under natural conditions cannot be ruled out.

It is noteworthy that, in our *in silico* analysis, similar sequences to the LBR from ABF-0020167 were only found in plant-pathogenic bacterial genomes. It is also noteworthy that no matches with other *Dickeya* strains were found. Moreover global BLAST analyses did not show any match of this LBR with *Pectobacterium carotovorum*, a phylogenetically related bacterium which does not sense JA (Antunez-Lamas *et al*., 2009). Therefore, ABF-0020167 could be considered as a potential receptor for JA, although further analyses are required to demonstrate the interaction of its LBR with this plant hormone. It is noteworthy that ABF-0018541 has also been identified as a putative JA chemoreceptor in our capillary assays. The sequence similarity between the LBR of these two chemoreceptors is very low, but it cannot be ruled out that a similar tertiary structure exists, allowing the binding of the same type of molecules. Indeed, both LBR sequences show homology with the annotated Pfam domain 4HB_MCP_1 (four helix bundle sensory module for signal transduction) (Ulrich and Zhulin, 2005), which is a ubiquitous sensory module in prokaryotic signal transduction. A detailed structural analysis of these regions will shed light on their implications in JA binding.

With regard to the ABF-0046680 chemoreceptor, the *46680* mutant was impaired in the ability to sense xylose, but further analyses are required to demonstrate whether: (i) the LBR of ABF-0046680 binds directly to xylose; and (ii) xylose is being released at wound sites.

Moreover, it is expected that the contribution of these chemoreceptors to the efficiency of infection by *Dd*3937 will depend on the host. As demonstrated previously, the chemotaxis process depends on the presence of a concentration gradient of the compound being sensed (Kalinin *et al*., 2010), and the gradient created in a wound for a given compound may be different for different hosts.

In summary, the study of *Dd*3937 chemoreceptors by bioinformatics analyses and experimental approaches has allowed us to identify the ABF-0020167 and ABF-0046680 chemoreceptors as putative receptors for biomolecules of great importance in plant defence: JA and xylose, respectively. Moreover, entry assays through wounds performed in *A. thaliana* leaves suggest that the perception of plant compounds released from a wound by the chemoreceptors ABF-0020167 and ABF-0046680 is pivotal to the entry into the plant apoplast and to the achievement of a population density sufficient to prosper in the plant apoplast. Further research is needed to demonstrate the interaction between the LBRs and these molecules.

EXPERIMENTAL PROCEDURES

Identification of MCP domains

For sequence analyses, the *Dd*3937 genome sequence was downloaded from the ASAP database (http://asap.ahabs.wisc.edu/asap/home.php). The *Dd*3937 genome was searched for genes encoding chemoreceptors. The search for the MCP domain was performed using an internal pipeline. The MCP domain was downloaded from the Pfam website of the Sanger Centre (http://pfam.sanger.ac.uk/) using the Pfam identifier PF00015. This alignment was used to build hidden Markov models (HMMs) by means of HMMER v2.3.2 (Eddy, 1998). Finally, a customized BioPerl (Stajich *et al*., 2002) script combined with HMMER searched for HMM hits using an E-value of 0.005 to search for MCP-containing proteins.

Identification of TMs and determination of LBR sequence

TM domains were identified one by one using the DAS algorithm at the DAS server (Cserzo *et al*., 1997, 2002). The complete sequences of

Table 2 Bacterial strains and plasmids used in this study.

Strain or plasmid	Relevant characteristics	Source or reference
Strains		
E. coli		
DH5α	*supE44 ΔlacU169 (Φ80 lacZΔM15) hsdR17 recA1 endA1 gyrA96 thi-1 relA1*	Hanahan (1983)
CC118 λ*pir*	*Δ (ara, leu) araD ΔlacX74 galE galK phoA20 thi-1 rps rpoB argE*(Am) *recA1* lysogenized with λ*pir* phage; Sp^r	Herrero *et al*. (1990)
S17-1λ*pir*	λ*pir* lysogen of S17-1	de Lorenzo, CIB, Madrid, Spain
UU1250	*Δaer-1 Δ(tar-tap)5201 Δtsr-7028 Δtrg-100 ygjG*::Gm *zbd*::Tn*5*	Bibikov *et al*. (2004)
K12	Wild-type strain	CBGP, Madrid Spain
D. dadantii 3937		
3937	Wild-type strain	Kotoujansky *et al*. (1985)
14536	*14536*::Tn*5* Sp^r *derivative of 3937*	This work
17662	*17662*::Tn*5* Sp^r *derivative of 3937*	This work
17665	*17665*::Tn*7* Km^r *derivative of 3937*	This work
17668	*17668*::Tn*7* Km^r *derivative of 3937*	This work
17824	*17824*::miniTn*5* Sp^r *derivative of 3937*	This work
17863	*17863*::miniTn*5* Sp^r *derivative of 3937*	This work
18541	*18541*::miniTn*5* Sp^r *derivative of 3937*	This work
18754	*18754*::miniTn*5* Sp^r *derivative of 3937*	This work
20167	*20167*::pKNG101 Sm^r *derivative of 3937*	This work
20252	*20252*::Tn*5* Sp^r *derivative of 3937*	This work
46680	*46680*::pKNG101 Sm^r *derivative of 3937*	This work
Plasmids and BACs		
pKNG101	R6K origin *sacB* marker exchange vector; mob^+, Sm^r	Kaniga *et al*. (1991)
pGEM-T easy®	Ap^r	Invitrogen
pGEM-T easy-18754	Ap^r	This work
pGEM-T easy-4387	Ap^r	This work
TrueBlue-BAC2®	Cm^r	Genomics One
TrueBlue-BAC2-18541	Cm^r	This work
TrueBlue-BAC2-20167	Cm^r	This work
TrueBlue-BAC2-20252	Cm^r	This work
TrueBlue-BAC2-46680	Cm^r	This work

chemoreceptors were searched for TMs considering a cut-off value of –2.5. Then, the LBR was extracted from the protein sequence as: (i) the sequence in between two TM domains when present; (ii) the sequence in between the amino terminus and the TM when only one TM was present; (iii) the sequence in between the amino terminus and the MCP domain when no TM was present; or (iv) the sequence in between the carboxyl terminus and the MCP domain when no TM was present. When three TMs were found, only the periplasmic LBR was considered. LBRs were identified in all MCPs and aligned with CLUSTAL W2, which renders the percentage identity matrix. The sizes of LBRs for the chemoreceptors belonging to topology Ia were clustered into groups of 10-amino-acid intervals.

Identification of Pfam domains in the LBR sequence

HMMs based on protein families were downloaded from the Pfam (Finn *et al*., 2014) website (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/). We focused on the Pfam component, so-called Pfam-A, corresponding to a number of high-quality, manually curated families. LBRs of each of the 47 MCP-associated proteins in Dd3937 were searched for the above Pfam domains using a customized BioPerl (Stajich *et al*., 2002) script. Only hits with an E-value of less than or equal to 0.005 were retained.

Global BLAST analysis of LBR sequences

Amino acid sequences of LBRs were searched against a database of bacterial genome sequences. For this, the complete bacterial sequences of 2997 chromosomes and 2164 plasmids, available as of 14 February 2014, were downloaded from GenBank Refseq. Local BLAST (E-value $\leq 10^{-30}$) was performed using an *ad hoc* PERL script. The output file was processed manually.

Microbiological methods

The bacterial strains and plasmids used in this study are listed in Table 2. Unless indicated, strains of *E. coli* were grown at 37 °C in Luria–Bertani (LB) medium (Sambrook *et al*., 1989). Unless indicated, strains of *Dd*3937 were cultivated at 28 °C in nutrient broth (NB) (per litre: yeast extract, 1 g; beef extract, 2 g; NaCl, 5 g; bactopeptone, 5 g) or King's B (KB) medium (King *et al*., 1954). When required, antibiotics were used at the following final concentrations (μg/mL): ampicillin, 100; chloramphenicol, 30; gentamycin, 20; kanamycin, 20; spectinomycin, 50; streptomycin, 20. Bacterial growth was monitored at an optical density at 600 nm (OD_{600}) using a Jenway 6300 spectrophotometer Jenway (Stone/Staffordshire/United Kingdom). For *Dd*3937 strains used in this study, $OD_{600} = 1.0$ corresponds approximately to 5×10^8 colony-forming units (cfu).

DNA manipulation

The plasmids used in this study are described in Table 2. To analyse the functions of the genes of interest in *Dd*3937, mutants in these genes were tracked by genomic polymerase chain reaction (PCR) from a mutant grid

available in our laboratory using two gene-specific primers and a transposon-specific primer. Those mutants not found in the grid were made by two different approaches: In short, the *17662* and *17668* mutants were constructed by Tn*7 in vitro* mutagenesis using a Genome Priming System (GPS-1) kit (New England Biolabs, Ipswich/Massachusetts/USA), and marker exchange was performed as described previously by Roeder and Collmer (1985). Construction of the *20167* and *46680* mutants was carried out by amplifying with *Pfu* DNA polymerase (Biotools, Madrid/Spain) an internal sequence between 700 and 800 bp from the *Dd*3937 genome and cloning into pKNG101 (Kaniga *et al*., 1991). The resulting plasmid was transferred into *Dd*3937 via triparental mating (de Lorenzo and Timmis, 1994). As pKNG101 cannot replicate in *Dd*3937, single crossover integrants were selected by resistance to streptomycin. In all cases, the mutations were verified by DNA sequencing.

For gain-of-function assays, ABF-0018754 and AIH-0004387 genes were amplified using a *Pfu* DNA polymerase from *Dd*3937 and *E. coli* K12 genomic DNA, respectively. The primers (5'→3') used were as follows: 18754F, GCCCATGTCATTTGATAC; 18754R, AGCGATTAAAAGGTTTCCC; 04387F, ATTCACTCGCGCTAATCTCC; 04387R, CTTATCGGGCATTTTCAT GGC. PCR products were cloned into pGEM-Teasy® vector and the resulting plasmids were transferred into *E. coli* UU1250 via electroporation (Sambrook *et al*., 1989).

For complementation of the mutants, ABF-0020167 and ABF-0046680 genes were amplified using a *Pfu* DNA polymerase from *Dd*3937 genomic DNA employing the following primers (5'→3'): 18541F_PstI, *CTGCA G*TAGCAATTGTCTCCAATTG; 18541R_XmaI, *CCCGG*GTTGTAACGCCATCA TC; 20167F_PstI, *CTGCAG*TTTGCGTTCAAGAGCCTC; 20167R_XmaI, *CCCG GG*TATCAAGGTGCAGTGC; 20252F_PstI, *CTGCAG*TACATGGAATGACGT TCACA; 20252R_XmaI, *CCCGGG*CCAAACAGGTGTGATGGTTA; 46680F_ BamHI, *GGATCC*ATCATTAATTCTCATGTTG; 46680R_PstI, *CTGCAG*TGTA TCAGCGTTGACCGC; the restriction sites used for cloning into TrueBlueBAC2© (Genomics One) are shown in italic type. The cloned genes were sequenced to confirm that no PCR errors had occurred, and TrueBlueBAC2-20167 or TrueBlueBAC2-46680 was introduced into the corresponding mutant by electroporation (Sambrook *et al*., 1989).

Gain-of-function assays

Starter cultures of *E. coli* cells were grown overnight in liquid LB medium to the late exponential phase and then 20-fold diluted into liquid LB medium and incubated at 37 °C to an OD_{600} of 0.6. Cultures were centrifuged at 6500 ***g*** and washed twice with 10 mM $MgCl_2$. Sediments were used for inoculation. Tryptone broth (TB) (Parkinson, 1976) 0.35% agar plates, supplemented or not with 1 mM serine (final concentration), were inoculated with bacterial sediments using a toothpick. Plates were incubated for 48 h at 37 °C. Five plates were inoculated for each experiment.

Swimming assays

*Dd*3937 cells were scraped from NB plates incubated overnight and suspended in 10 mM $MgCl_2$ to an OD_{600} of 1.0. A 5-μL drop from each bacterial suspension was inoculated at the centre of each of five KB 0.25% agar plates. Plates were incubated for 24 h at 28 °C with high humidity. After that, bacterial haloes were recorded. Significant differences between the CIs of WT and mutant strains were determined and statistically compared according to Student's *t*-test ($P < 0.05$).

The chemical 'in-plug' bacterial chemotaxis assay

*Dd*3937 cells were scraped from NB plates incubated overnight and suspended in 10 mM $MgCl_2$ at the desired final concentration. Then the bacterial suspension was mixed with Minimal Medium A (MMA) agar medium (Antúnez-Lamas *et al*., 2009) to a final concentration of 0.25% agar and 2.5×10^8 cfu/mL. Once polymerized, a 5-μL drop containing a given chemical was placed on the agar surface (Tso and Adler, 1974). Xylose, serine and JA were used at 10 mM (final concentration). Plates were incubated for 20 min at room temperature. After that, bacterial haloes were recorded. Five replicates were analysed for each experiment. Water was used as a solvent for xylose and serine, and JA was dissolved in ethanol.

Capillary assays

The chemotaxis of *Dd*3937 towards several compounds was measured using a modified Palleroni assay (Palleroni, 1976). CI was defined as the number of cells that accumulated after 20 min within a 5-μL tip containing 2 μL of a given compound divided by the number of cells accumulated within a 5-μL tip containing 2 μL of the solvent in which the compound was dissolved. Values of CI significantly greater than 1.0 were identified as a positive response of the cells to the chemical compound. Values of CI significantly less than 1.0 were identified as a negative response of the cells to the chemical compound. Three replicates (capillaries) were analysed in each experiment. Xylose, serine and JA were tested at 1 mM. Significant differences between the CIs of WT and mutant strains were determined and statistically compared according to Student's *t*-test ($P < 0.05$).

Plant bioassays

Entry and entry-dependent colonization assays of *Dd*3937 strains in *A. thaliana* leaves were carried out as described previously by Antunez-Lamas *et al*. (2009) using 4-week-old Col-0 ecotype plants. A 10-μL drop containing 10^5 cfu was placed on a hole made with a needle on the adaxial side of leaves. Ten leaves were inoculated with each strain. Three leaves were mock infiltrated with 10 mM $MgCl_2$. After 2 or 24 h at 28 °C, leaves were cut and washed with sterile water, and viable cell counts in leaves were determined by serial dilution and plating.

For non-entry-dependent colonization assays, 10 leaves were infiltrated into the abaxial side with a suspension containing 2×10^6 cfu/mL by syringe infiltration. Three leaves were mock infiltrated with 10 mM $MgCl_2$. After 24 h at 28 °C, leaves were processed as described above.

In both kinds of experiment, significant differences between WT and mutant cells inside the leaves were determined and statistically compared according to Student's *t*-test ($P < 0.05$).

ACKNOWLEDGEMENTS

This research was supported by the Spanish Plan Nacional I+D+I grants AGL-2009-12757 and AGL2012-32516. I. Río-Álvarez was supported by the FPI program (MICINN-Spain). C. Muñoz-Gómez was supported by the FPI program (MINECO-Spain). P. M. Martínez-García was supported by the Campus de Excelencia Internacional Andalucía Tech and AGL2011-30343-C02-01. We thank Dr Tino Krell for critical reading of the manuscript and helpful comments, and for providing the *E. coli* UU1250 strain. We also thank Dr Miguel Angel Torres for helpful review of the manuscript.

REFERENCES

Alexander, R.P. and Zhulin, I.B. (2007) Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. *Proc. Natl. Acad. Sci. USA*, **104**, 2885–2890.

Alexandre, G., Greer-Phillips, S. and Zhulin, I.B. (2004) Ecological role of energy taxis in microorganisms. *FEMS Microbiol. Rev.* **28**, 113–126.

Alvarez-Ortega, C. and Harwood, C.S. (2007) Identification of a malate chemoreceptor in *Pseudomonas aeruginosa* by screening for chemotaxis defects in an energy taxis-deficient mutant. *Appl. Environ. Microbiol.* **73**, 7793–7795.

Antunez-Lamas, M., Cabrera, E., Lopez-Solanilla, E., Solano, R., Gonzalez-Melendi, P., Chico, J.M., Toth, I., Birch, P., Pritchard, L., Liu, H. and Rodriguez-Palenzuela, P. (2009) Bacterial chemoattraction towards jasmonate plays a role in the entry of *Dickeya dadantii* through wounded tissues. *Mol. Microbiol.* **74**, 662–671.

Antúnez-Lamas, M., Cabrera-Ordonez, E., Lopez-Solanilla, E., Raposo, R., Trelles-Salazar, O., Rodriguez-Moreno, A. and Rodriguez-Palenzuela, P. (2009) Role of motility and chemotaxis in the pathogenesis of *Dickeya dadantii* 3937 (ex *Erwinia chrysanthemi* 3937). *Microbiology*, **155**, 434–442.

Barabote, R.D., Johnson, O.L., Zetina, E., San Francisco, S.K., Fralick, J.A. and San Francisco, M.J.D. (2003) *Erwinia chrysanthemi tolC* is involved in resistance to antimicrobial plant chemicals and is essential for phytopathogenesis. *J. Bacteriol.* **185**, 5772–5778.

Barras, F., van Gijsegem, F. and Chatterjee, A.K. (1994) Extracellular enzymes and pathogenesis of soft-rot *Erwinia*. *Annu. Rev. Phytopathol.* **32**, 201–234.

Bauer, D.W., Bogdanove, A.J., Beer, S.V. and Collmer, A. (1994) *Erwinia chrysanthemi hrp* genes and their involvement in soft rot pathogenesis and elicitation of the hypersensitive response. *Mol. Plant–Microbe Interact.* **7**, 573–581.

Bibikov, S.I., Miller, A.C., Gosink, K.K. and Parkinson, J.S. (2004) Methylation-independent aerotaxis mediated by the *Escherichia coli* Aer protein. *J. Bacteriol.* **186**, 3730–3737.

Briegel, A., Li, X., Bilwes, A.M., Hughes, K.T., Jensen, G.J. and Crane, B.R. (2012) Bacterial chemoreceptor arrays are hexagonally packed trimers of receptor dimers networked by rings of kinase and coupling proteins. *Proc. Natl. Acad. Sci. USA*, **109**, 3766–3771.

Charkowski, A., Blanco, C., Condemine, G., Expert, D., Franza, T., Hayes, C., Hugouvieux-Cotte-Pattat, N., López-Solanilla, E., Low, D., Moleleki, L., Pirhonen, M., Pitman, A., Perna, N., Reverchon, S., Rodríguez-Palenzuela, P., San Francisco, M., Toth, I., Tsuyumu, S., van der Waals, J., van der Wolf, J., Van Gijsegem, F., Yang, C.H. and Yedidia, I. (2012) The role of secretion systems and small molecules in soft-rot Enterobacteriaceae pathogenicity. *Annu. Rev. Phytopathol.* **50**, 425–449.

Costechareyre, D., Chich, J.F., Strub, J.M., Rahbé, Y. and Condemine, G. (2013) Transcriptome of *Dickeya dadantii* infecting *Acyrthosiphon pisum* reveals a strong defense against antimicrobial peptides. *PLoS ONE*, **8**, e54118.

Cserzo, M., Wallin, E., Simon, I., von Heijne, G. and Elofsson, A. (1997) Prediction of TM alpha-helices in prokaryotic membrane proteins: the Dense Alignment Surface method. *Protein Eng.* **10**, 673–676.

Cserzo, M., Eisenhaber, F., Eisenhaber, B. and Simon, I. (2002) On filtering false positive transmembrane protein predictions. *Protein Eng.* **15**, 745–752.

Cuppels, D.A. (1988) Chemotaxis by *Pseudomonas syringae* pv tomato. *Appl. Environ. Microbiol.* **54**, 629–632.

De Lorenzo, V. and Timmis, K.N. (1994) Analysis and construction of stable phenotypes in gram-negative bacteria with Tn*5*- and Tn*10*-derived minitransposons. *Methods Enzymol.* **235**, 386–405.

Dellagi, A., Rigault, M., Segond, D., Roux, C., Kraepiel, Y., Cellier, F., Briat, J.F., Gaymard, F. and Expert, D. (2005) Siderophore-mediated upregulation of *A. thaliana* ferritin expression in response to *Erwinia chrysanthemi* infection. *Plant J.* **43**, 262–272.

Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Fagard, M., Dellagi, A., Roux, C., Perino, C., Rigault, M., Boucher, V., Shevchik, V.E. and Expert, D. (2007) *Arabidopsis thaliana* expresses multiple lines of defense to counterattack *Erwinia chrysanthemi*. *Mol. Plant–Microbe Interact.* **20**, 794–805.

Falke, J.J. and Hazelbauer, G.L. (2001) Transmembrane signaling in bacterial chemoreceptors. *Trends Biochem. Sci.* **26**, 257–265.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L., Tate, J. and Punta, M. (2014) Pfam: the protein families database. *Nucleic Acids Res.* **42**, 222–230.

Hanahan, D. (1983) Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* **166**, 557–580.

Hawes, M.C. and Smith, L.Y. (1989) Requirement for chemotaxis in pathogenicity of *Agrobacterium tumefaciens* on roots of soil-grown pea plants. *J. Bacteriol.* **171**, 5668–5671.

Hazelbauer, G.L. and Lai, W.C. (2010) Bacterial chemoreceptors, providing enhanced features to two-component signaling. *Curr. Opin. Microbiol.* **13**, 124–132.

Herrero, M., de Lorenzo, V. and Timmis, K.N. (1990) Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in gram-negative bacteria. *J. Bacteriol.* **172**, 6557–6567.

Jarrell, K.F. and McBride, M.J. (2008) The surprisingly diverse ways that prokaryotes move. *Nat. Rev. Microbiol.* **6**, 466–476.

Kalinin, Y., Neumann, S., Sourjik, V. and Wu, M. (2010) Responses of *Escherichia coli* bacteria to two opposing chemoattractant gradients depend on the chemoreceptor ratio. *J. Bacteriol.* **192**, 1796–1800.

Kamoun, S. and Kado, C.I. (1990) Phenotypic switching affecting chemotaxis, xanthan production, and virulence in *Xanthomonas campestris*. *Appl. Environ. Microbiol.* **56**, 3855–3860.

Kaniga, K., Delor, I. and Cornelis, G.R. (1991) A wide-host-range suicide vector for improving reverse genetics in gram-negative bacteria: inactivation of the *blaA* gene of *Yersinia enterocolitica*. *Gene*, **109**, 137–141.

Kato, J., Kim, H.E., Takiguchi, N., Kuroda, A. and Ohtake, H. (2008) *Pseudomonas aeruginosa* as a model microorganism for investigation of chemotactic behaviors in ecosystem. *J. Biosci. Bioeng.* **106**, 1–7.

Kim, S.H., Prive, G.G., Pandit, J., Koshland, D.E., Yeh, J.I. and Biemann, H.P. (1996) High-resolution structures of the ligand binding domain of the wild-type bacterial aspartate receptor. *J. Mol. Biol.* **262**, 186–201.

King, E.O., Ward, M.K. and Raney, D.E. (1954) Two simple media for the demonstration of pyocyanin and fluorescin. *J. Lab. Clin. Med.* **44**, 301–307.

Kotoujansky, A., Diolez, A., Boccara, M., Bertheau, Y., Andro, T. and Coleno, A. (1985) Molecular cloning of *Erwinia chrysanthemi* pectinase and cellulase structural genes. *EMBO J.* **4**, 781–785.

Krell, T., Lacal, J., Muñoz-Martínez, F., Reyes-Darias, J.A., Cadirci, B.H., García-Fontana, C. and Ramos, J. (2011) Diversity at its best: bacterial taxis. *Environ. Microbiol.* **13**, 1115–1122.

Lacal, J., García-Fontana, C., Muñoz-Martínez, F., Ramos, J.L. and Krell, T. (2010a) Sensing of environmental signals: classification of chemoreceptors according to the size of their ligand binding regions. *Environ. Microbiol.* **12**, 2873–2884.

Lacal, J., Alfonso, C., Liu, X., Parales, R.E., Morel, B., Conejero-Lara, F., Rivas, G., Duque, E., Ramos, J.L. and Krell, T. (2010b) Identification of a chemoreceptor for TCA cycle intermediates: ligand induced dimerization of chemoreceptor sensor domain determines magnitude of chemotaxis. *J. Biol. Chem.* **285**, 23 126–23 136.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

León, J., Rojo, E. and Sánchez-Serrano, J.J. (2001) Wound signalling in plants. *J. Exp. Bot.* **52**, 1–9.

Llama-Palacios, A., López-Solanilla, E., Poza-Carrión, C., García-Olmedo, F. and Rodríguez-Palenzuela, P. (2003) The *Erwinia chrysanthemi phoP-phoQ* operon plays an important role in growth at low pH, virulence and bacterial survival in plant tissue. *Mol. Microbiol.* **49**, 347–357.

Llama-Palacios, A., Lopez-Solanilla, E. and Rodriguez-Palenzuela, P. (2005) Role of the PhoP-PhoQ system in the virulence of *Erwinia chrysanthemi* strain 3937: involvement in sensitivity to plant antimicrobial peptides, survival at acid pH, and regulation of pectolytic enzymes. *J. Bacteriol.* **187**, 2157–2162.

López-Solanilla, E., García-Olmedo, F. and Rodríguez-Palenzuela, P. (1998) Inactivation of the *sapA* to *sapF* locus of *Erwinia chrysanthemi* reveals common features in plant and animal bacterial pathogenesis. *Plant Cell*, **10**, 917–924.

López-Solanilla, E., Llama-Palacios, A., Collmer, A., García-Olmedo, F. and Rodríguez-Palenzuela, P. (2001) Relative effects on virulence of mutations in the *sap*, *pel*, and *hrp* loci of *Erwinia chrysanthemi*. *Mol. Plant–Microbe Interact.* **14**, 386–393.

Ma, B., Charkowski, A.O., Glasner, J.D. and Perna, N.T. (2014) Identification of host–microbe interaction factors in the genomes of soft rot-associated pathogens *Dickeya dadantii* 3937 and *Pectobacterium carotovorum* WPP14 with supervised machine learning. *BMC Genomics*, **15**, 508.

Maggiorani Valecillos, A., Rodriguez Palenzuela, P. and Lopez-Solanilla, E. (2006) The role of several multidrug resistance systems in *Erwinia chrysanthemi* pathogenesis. *Mol. Plant–Microbe Interact.* **19**, 607–613.

Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., Ronald, P., Dow, M., Verdier, V., Beer, S.V., Machado, M.A., Toth, I., Salmond, G. and Foster, G.D.

(2012) Top 10 plant pathogenic bacteria in molecular plant pathology. *Mol. Plant Pathol.* **13**, 614–629.

Moens, S. and Vanderleyden, J. (1996) Functions of bacterial flagella. *Crit. Rev. Microbiol.* **22**, 67–100.

Munzinger, M., Budzikiewicz, H., Expert, D., Enard, C. and Meyer, J.M. (2000) Achromobactin, a new citrate siderophore of *Erwinia chrysanthemi*. *Z. Naturforsch.* **55**, 328–332.

Nichols, N.N., Lunde, T.A., Graden, K.C., Hallock, K.A., Kowalchyk, C.K., Southern, R.M., Soskin, E.J. and Ditty, J.L. (2012) Chemotaxis to furan compounds by furan-degrading *Pseudomonas* strains. *Appl. Environ. Microbiol.* **78**, 6365–6368.

Nishiyama, S., Suzuki, D., Itoh, Y., Suzuki, K., Tajima, H., Hyakutake, A., Homma, M., Butler-Wu, S.M., Camilli, A. and Kawagishi, I. (2012) Mlp24 (McpX) of *Vibrio cholera* implicated in pathogenicity functions as a chemoreceptor for multiple amino acids. *Infect. Immun.* **80**, 3170–3178.

Oku, S., Komatsu, A., Tajima, T., Nakashimada, Y. and Kato, J. (2012) Identification of chemotaxis sensory proteins for amino acids in *Pseudomonas fluorescens* Pf0-1 and their involvement in chemotaxis to tomato root exudate and root colonization. *Microbes Environ.* **27**, 462–469.

Palleroni, N.J. (1976) Chamber for bacterial chemotaxis experiments. *Appl. Environ. Microbiol.* **32**, 729–730.

Parkinson, J. (1976) *cheA*, *cheB*, and *cheC* genes of *Escherichia coli* and their role in chemotaxis. *J. Bacteriol.* **126**, 758–770.

Persmark, M., Expert, D. and Neilands, J.B. (1989) Isolation, characterization, and synthesis of chrysobactin, a compound with siderophore activity from *Erwinia chrysanthemi*. *J. Biol. Chem.* **264**, 3187–3193.

Porter, S.L., Wadhams, G.H. and Armitage, J.P. (2011) Signal processing in complex chemotaxis pathways. *Nat. Rev. Microbiol.* **9**, 153–165.

Rahman, H., King, R.M., Shewell, L.K., Semchenko, E.A., Hartley-Tassell, L.E., Wilson, J.C., Day, C.J. and Korolik, V. (2014) Characterization of a multi-ligand binding chemoreceptor CcmL (Tlp3) of *Campylobacter jejuni*. *PLoS Pathog.* **10**, e1003822.

Rennie, E.A. and Scheller, H.V. (2014) Xylan biosynthesis. *Curr. Opin. Biotechnol.* **26C**, 100–107.

Reverchon, S. and Nasser, W. (2013) *Dickeya* ecology, environment sensing and regulation of virulence programme. *Environ. Microbiol. Rep.* **5**, 622–636.

Rico-Jiménez, M., Muñoz-Martínez, F., García-Fontana, C., Fernandez, M., Morel, B., Ortega, A., Ramos, J.L. and Krell, T. (2013) Paralogous chemoreceptors mediate chemotaxis towards protein amino acids and the non-protein amino acid gamma-aminobutyrate (GABA). *Mol. Microbiol.* **88**, 1230–1243.

Rio-Alvarez, I., Rodríguez-Herva, J.J., Cuartas-Lanza, R., Toth, I., Pritchard, L., Rodriquez-Palenzuela, P. and López-Solanilla, E. (2012) Genome-wide analysis of the response of Dickeya dadantii 3937 to plant antimicrobial peptides. *Mol. Plant Microbe Interact.* **25**, 523–533.

Roeder, D.L. and Collmer, A. (1985) Marker-exchange mutagenesis of a pectate lyase isozyme gene in *Erwinia chrysanthemi*. *J. Bacteriol.* **164**, 51–56.

Sambrook, J., Frits, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Sourjik, V. and Wingreen, N.S. (2012) Responding to chemical gradients: bacterial chemotaxis. *Curr. Opin. Cell Biol.* **24**, 262–268.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D. and Birney, E. (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res.* **12**, 1611–1618.

Sweeney, E.G., Henderson, J.N., Goers, J., Wreden, C., Hicks, K.G., Foster, J.K., Parthasarathy, R., Remington, S.J. and Guillemin, K. (2013) Structure and proposed mechanism for the pH sensing *Helicobacter pylori* chemoreceptor TlpB. *Structure*, **20**, 1177–1188.

Taurino, M., Abelenda, J.A., Río-Alvarez, I., Navarro, C., Vicedo, B., Farmaki, T., Jiménez, P., García-Agustín, P., López-Solanilla, E., Prat, S., Rojo, E., Sanchez-Serrano, J. and Sanmartin, M. (2014) Jasmonate-dependent modifications of the pectin matrix during potato development function as a defense mechanism targeted by *Dickeya dadantii* virulence factors. *Plant J.* **77**, 418–429.

Tso, W.W. and Adler, J. (1974) Negative chemotaxis in *Escherichia coli*. *J. Bacteriol.* **118**, 560–576.

Ulrich, L.E. and Zhulin, I.B. (2005) Four-helix bundle: a ubiquitous sensory module in prokaryotic signal transduction. *Bioinformatics*, **21** (Suppl 3), iii45–iiii8.

Wadhams, G.H. and Armitage, J.P. (2004) Making sense of it all: bacterial chemotaxis. *Nat. Rev. Mol. Cell Biol.* **5**, 1024–1037.

Yamamoto, K. and Imae, Y. (1993) Cloning and characterization of the *Salmonella typhimurium*-specific chemoreceptor Tcp for taxis to citrate and from phenol. *Proc. Natl. Acad. Sci. USA*, **90**, 217–221.

Yang, C.H., Gavilanes-Ruiz, M., Okinaka, Y., Vedel, R., Berthuy, I., Boccara, M., Wei-Ta Chen, J., Perna, N.T. and Keen, N.T. (2002) *hrp* genes of *Erwinia chrysanthemi* 3937 are important virulence factors. *Mol. Plant–Microbe Interact.* **5**, 472–480.

Yao, J. and Allen, C. (2006) Chemotaxis is required for virulence and competitive fitness of the bacterial wilt pathogen *Ralstonia solanacearum*. *J. Bacteriol.* **188**, 3697–3708.

Yao, J. and Allen, C. (2007) The plant pathogen *Ralstonia solanacearum* needs aerotaxis for normal biofilm formation and interactions with its tomato host. *J. Bacteriol.* **189**, 6415–6424.

Yap, M.N., Yang, C.H., Barak, J.D., Jahn, C.E. and Charkowski, A.O. (2005) The *Erwinia chrysanthemi* type III secretion system is required for multicellular behavior. *J. Bacteriol.* **187**, 639–648.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Fig. S1 Complementation of non-chemotactic strain *Escherichia coli* UU1250 with putative serine chemoreceptors. Tryptone broth (TB) 0.35% agar plates (A) and TB 0.35% agar plates supplemented with 1 mM serine (B) were inoculated with wild-type (WT), UU1250 or transformed *E. coli* UU1250 strains expressing the *E. coli* serine chemoreceptor (AIH-0004387) or the putative serine chemoreceptor of *Dd*3937 (ABF-0018754). Results were recorded at 48 h. Similar results were obtained in three independent experiments.

Fig. S2 'In-plug' chemotactic response of *Dd*3937 strains towards serine. A drop containing water or serine (10 mM) was deposited on the (0.25%) soft-agar bacteria. All the *Dd*3937 strains tested [wild-type (WT) and mutants] showed no chemotactic response towards water; a representative image is shown. Haloes were recorded after 20 min. Similar results were obtained in three independent experiments.

Fig. S3 Swimming ability of *Dd*3937 strains in rich medium. King's B (KB) 0.25% agar plates were inoculated with *Dd*3937 wild-type (WT) and mutant strains at the exponential phase. Diameters of haloes were recorded at 24 h. The means and standard errors are shown. Error bars represent standard error of the mean (SEM). Similar results were obtained in three independent experiments.

Table S1 Percentage identity matrix of the Dd3937 ligand-binding region (LBR).

Table S2 Pfam matches for Dd3937 ligand-binding regions (LBRs). (a) Nomenclature from the ASAP database (https://asap.ahabs.wisc.edu/asap/logon/php). In grey, methyl-accepting chemotaxis proteins (MCPs) under study in this work.

Table S3 Global BLAST of Dd3937 ligand-binding regions (LBRs). Amino acid sequences of LBRs were searched against a database of bacterial genome sequences using BLAST (E-value $\leq 10^{-30}$). * indicates that the periplasmic LBR was selected when two LBRs were present. (a) Nomenclature from the ASAP database (https://asap.ahabs.wisc.edu/asap/logon/php).