AN ABSTRACT OF THE DISSERTATION OF

Jeffrey A. Kimbrel for the degree of Doctor of Philosophy in Molecular and Cellular Biology presented on March 13, 2012.

Title: Genome-enabled Discovery and Characterization of Type III Effector-encoding Genes of Plant Symbiotic Bacteria.

Abstract approved:

Jeff H. Chang

Symbiosis is the close and protracted interaction between organisms. The molecular interactions that occur during symbiosis are complex with multiple barriers that must be overcome. Many Gram-negative, host-associated bacteria use a type III secretion system to mediate associations with their eukaryotic hosts. This secretion system is a specialized apparatus for the injection of type III effector proteins directly into host cells, which in the case of plant pathogens, are collectively necessary to modulate host defense. The type III secretion system is not a mechanism exclusive to pathogens, however, as many strains of commensal *Pseudomonas fluorescens* and mutualistic rhizobia demonstrably require a type III secretion system to interact with their host plants. The work presented in this thesis describes genome-enabled approaches for characterizing type III effector genes across the range of plant symbiosis. Using high-throughput sequencing technology, draft genome

sequences were generated for the plant pathogen, *Xanthomonas hortorum* pv. *carotae* M081, the plant commensal, *Pseudomonas fluorescens* WH6, and six strains from the plant mutualists *Sinorhizobium fredii* and *Bradyrhizobium japonicum*. Analyses of the draft genome sequences and publicly available finished sequences contributed insights into mechanisms of host-association and to increasing the inventory of type III effector sequences as well as developing methods directly applicable for agriculture. Finally, characterization of the genetic diversity of type III effectors from rhizobia shows that collections of type III effectors of mutualists are static, with little diversity in content and sequence variation. This represents the first comprehensive cataloging of type III effector from species of mutualistic bacteria and the first to provide evidence for purifying selection of this important class of genes.

# Genome-enabled Discovery and Characterization of Type III Effector-encoding Genes of Plant Symbiotic Bacteria

by

Jeffrey A. Kimbrel

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented March 13, 2012
Commencement June 2012

Doctor of Philosophy dissertation of Jeffrey A. Kimbrel presented on March 13, 2012.

APPROVED:

Major Professor, representing Molecular and Cellular Biology

Director of the Molecular and Cellular Biology Program

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Jeffrey A. Kimbrel, Author

## ACKNOWLEDGEMENTS

I would like to thank Jeff Chang, for his excellent mentorship and for providing me the inspiration to make the most out of my PhD training. Thank you to my graduate committee, and the faculty and staff that have helped me on my way. Thanks to all my friends and fellow graduate students, particularly Sam Fox and Bill Thomas, who made the lows seem a little higher. I would like to thank my family, especially my mother, Cindy Muniz, for the endless encouragement. Finally, an enormous thank you to Liz Harper, for giving me strength with her selfless support and motivation.

## CONTRIBUTION OF AUTHORS

Chapter 1: JAK wrote the manuscript.

Chapter 2: JAK sequenced, assembled and analyzed the *Xhc* genome, type III effector characterization and did all bioinformatic analyses except for annotation of the genome, which was done by SAG. TNT did all PCR assays. JAK, KBJ and JHC wrote the manuscript.

Chapter 3: JAK prepared the DNA for sequencing, assembled and analyzed the genome sequence of WH6, as well as drafted the manuscript. SAG annotated the genome. ABH sequenced the DNA flanking the Tn*5* insertions, did the preliminary analyses of gene functions, and helped draft the manuscript. ALC assisted with analyzing the different assemblies. DIM, GMB, DJA, and JHC conceived of the study and drafted the manuscript.

Chapter 4: JAK and WJT prepared the DNA for sequencing and assembled the genomes. JAK did all subsequent bioinformatic analyses. WJT, with assistance from JAK ALC, and CAT cloned and characterized candidates for translocation. JAK, WJT and JHC wrote the manuscript.

Chapter 5: JAK and JHC wrote the manuscript.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

Figure Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF TABLES

Table Page

LIST OF APPENDICES

Appendix Page

LIST OF APPENDICES (Continued)

Appendix Page

LIST OF APPENDICES (Continued)

LIST OF APPENDIX FIGURE

Figure Page

LIST OF APPENDIX TABLE

# Introduction: Plant-microbe interactions

Jeffrey A. Kimbrel

## INTRODUCTION

The environment in which a plant lives is teeming with microbial life. Many of the encounters between a plant and bacteria are incidental, but many interactions lead to meaningful relationships whose outcomes vary with the players involved. These outcomes include a full spectrum from mutualism to commensalism to parasitism. Plant defense plays a central role in dictating the boundaries of these interactions, and, regardless of the type of interaction, host-establishment can be determined by a microbe's ability to circumvent these boundaries and initiate a symbiotic relationship.

Most of what is known about bacterial avoidance of plant defense has been elucidated by work with plant pathogens (Dangl and Jones, 2001; Jones and Dangl, 2006). The central idea is that plants are able to recognize bacteria and trigger a battery of innate immune responses that collectively suppress infection. This understanding of plant immunity forms a basic framework for plant interactions with any microbe, not only for potentially pathogenic bacteria. Bacteria with a commensal or mutualistic lifestyle must also avoid triggering or have mechanisms to suppress the plant's immune responses before host-association can be established. Bacteria achieve this goal by using many different strategies, and a prominent strategy described in this thesis is the use of the type III secretion system (T3SS) to deliver type III effector proteins (T3Es) into the host plant cell (Hueck, 1998). Throughout this introduction, the use of the T3SS in pathogens is explored, and parallels are drawn between how pathogens, commensal, and mutualist bacteria use the T3SS.

## PATHOGENS AS A MODEL FOR PLANT-MICROBE INTERACTIONS

Plant innate immunity begins following the perception of non-plant features such as the flg22 component of the flagellum (Felix et al., 1999). These recognized bacterial features are termed PAMPs (or MAMPs) for Pathogen-associated (or Microbe-associated) Molecular Patterns (Jones and Dangl, 2006; Schwessinger and Zipfel, 2008). The defense response mediated through the recognition of PAMPs or MAMPs is called PAMP- or Pattern-Triggered Immunity (PTI).

The recognition of PAMPs is mediated by pattern-recognition receptors (PRR). PRR proteins are exterior surface proteins of a plant cell that internalize a signal to induce a variety of physiological changes (Segonzac and Zipfel, 2011). These changes include immediate defense responses such as the release of calcium and reactive-oxygen species (ROS) signaling molecules, and subsequent responses such as changes in gene expression and cell wall strengthening via the deposition of callose (Luna et al., 2011). Cytosolic $Ca^{2+}$ concentrations rise within minutes of PAMP perception, and this $Ca^{2+}$ is involved with signaling for nitric oxide (NO) production, as well as with inducing transcriptional changes through the induction of $Ca^{2+}$-dependent protein kinases (Ma, 2011). In addition to ROSs primarily acting as signaling molecules, $H_2O_2$ can also cross-link glycoproteins to the cell wall to strengthen the physical barrier (Torres et al., 2006). Induction of mitogen-activated protein kinases (MAPK) through the action

of PRRs also lead to transcriptome changes, though the direct targets of MAPK phosphorylation are currently not well-defined (Tena et al., 2011).

PTI must be avoided, delayed or overcome in order to initiate a symbiotic relationship (Soto et al., 2009). Many bacteria do this by secreting or injecting effector proteins into a host cell via specialized secretion systems to alter the host's perception or response to the bacterium (Cambronne and Roy, 2006). One of the most heavily characterized secretion systems of plant- and animal-interacting bacteria is the type III secretion system (T3SS), which injects type III effector proteins (T3E) (Hueck, 1998). Inside the host cell, the T3Es function to block or delay many aspects of PTI (Grant et al., 2006). For example, several T3Es from *Pseudomonas syringae* pv. *tomato* DC3000 (*Pto*DC3000) have been shown to interfere with the host-signaling induced by flg22 perception (Felix et al., 1999; Li et al., 2005). Another *Pto*DC3000 T3E, hopU1, has been shown to ADP-ribosylate the arabidopsis AtGRP7 RNA-binding protein, which is thought to alter host defense by adjusting RNA metabolism (Fu et al., 2007).

The number of T3Es in a pathogen's arsenal ranges from as few as twelve in *Erwinia*, to over 70 in *Ralstonia* (Cunnac et al., 2004; Nissinen et al., 2007). Among T3E-possessing species of plant-associated bacteria, host-specificity is not necessarily correlated with a bacterium's collection of T3Es. In fact, among *P. syringae*, the strains with the fewest number of T3Es often have the broadest host range (O'Brien et al., 2010). The narrowing of host range and subsequent swelling of the T3E collection of pathogens are a consequence of co-evolution,

whereby avoidance of host defenses have shaped the members and functions of a pathogen's T3E collection (McCann and Guttman, 2008).

T3Es can also be negative host-range determinants, through their recognition by the host, thereby betraying the presence of the bacterium. This layer of defense is referred to as effector-triggered immunity (ETI), and is basically an amplified and more rapid PTI response (Jones and Dangl, 2006). Because the recognized T3Es render the pathogen avirulent, these T3E were historically called avirulence (Avr) proteins (Flor, 1971; van der Biezen and Jones, 1998). Avr recognition is either directly or indirectly mediated by resistance proteins (R proteins, products of "*R* genes"), which typically have Nucleotide-Binding and Leucine Rich Repeat domains (NB-LRR; (Dangl and Jones, 2001; Jones and Dangl, 2006)). Early studies of *Avr* and *R* genes in the flax/flax-rust interaction led to the gene-for-gene hypothesis which states that a single host *R* gene "recognizes" a cognate *avr* gene of a pathogen (Flor, 1956; Ellis and Dodds, 2007). Detection of an Avr through an R protein triggers ETI, which culminates in a localized cell death termed the Hypersensitive Response (HR; Greenberg and Yao, 2004).

Despite the many characterized Avr-R "interactions", direct binding between the two proteins has only been shown in a few instances (Scofield et al., 1996; Tang et al., 1996; Jia et al., 2000; Krasileva et al., 2010). Evidence suggested an alternate mechanism for the action of some R proteins, establishing the "guard hypothesis" - that most R proteins guard cellular machinery likely targeted by an Avr, and recognize the Avr through indirect means (Dangl and

Jones, 2001). One well-studied guard protein is RPS2, which perceives Avr-dependent modifications to its guardee RIN4, e.g., the cleavage of RIN4 by AvrRpt2 (Axtell and Staskawicz, 2003; Mackey et al., 2003; Chisholm et al., 2005).

## HOST-ASSOCIATION AND USE OF T3SS BY COMMENSAL BACTERIA

T3SS-encoding loci have been identified in the genomes of environmental microbes without a known host-associated lifestyle, as well as soil and plant-commensals. The latter group of bacteria are often ubiquitous in soil, and often have unknown interactions with plants in the rhizosphere. Despite their name, these interactions between plants and bacteria traditionally called “commensals” can result in consequences bordering on mutualism or parasitism. This dual effect can be seen with *Pseudomonas fluorescens* WH6, which can be both beneficial and harmful to plants. WH6 acts as a plant-growth promoting rhizobacterium (PGPR) towards dicots, through the action of a WH6-produced compound that inhibits the germination of monocots (Banowetz et al., 2008). Other PGPRs may influence their host through direct means, such as the production of molecules like phytohormones, or by indirect means by keeping pathogens at bay (Haas and Défago, 2005).

For commensal bacteria, the T3SS has been intensively studied in *P. fluorescens* (Rezzonico et al., 2004; Pallen et al., 2005). Based on PCR using oligonucleotide primers to *hrcN*, a core T3SS gene, it was concluded that T3SS-encoding loci were widespread (Preston et al., 2001; Rezzonico et al., 2004).

However, the function conferred by the T3SS to *P. fluorescens* remains unclear. In *P. fluorescens* SBW25, for example, the "necessary" *hrcN* gene appears absent and mutations in other T3SS-related genes had no effect on the ability of SBW25 to grow in association with its host. Nevertheless, when genetically modified to constitutively express the T3SS, SBW25 was able to induce an AvrB-dependent HR suggesting the potential for functionality despite the absence of *hrcN* (Preston et al., 2001). Similarly, a T3SS mutant of a different strain of *P. fluorescens*, KD, showed no decrease in rhizosphere fitness. A T3SS mutant of KD, however, showed a marked reduction in its thwarting of damping-off disease by the oomycete *Pythium ultimum*, suggesting the possibility that KD-mediated biocontrol is accomplished by a direct interaction between bacteria and the oomycete (Rezzonico et al., 2005).

## COMPLEX MOLECULAR DIALOG BETWEEN RHIZOBIA AND LEGUMES

Some species of mutualistic rhizobia encode for functional T3SSs. Rhizobia are critical to the nitrogen cycle, replenishing the soil with fixed nitrogen through their interaction with legumes (Masson-Boivin et al., 2009). Rhizobia are α-proteobacteria belonging to six different genera, with the earliest common ancestor giving rise to the "fast-growing" and "slow-growing" rhizobia some 500 million years ago, ~100 million years before mycorrhizae, and ~400 million years before the first leguminous plant (Dresler-Nurmi et al., 2009). Rhizobia have two distinct lifestyles, free-living in the rhizosphere, and differentiated inside a nodule where they reduce atmospheric $N_2$ into ammonia. Like most bacterial mutualists,

rhizobia exchange nutrients that are difficult for the host to obtain, for water and carbon products provided by their host (Dresler-Nurmi et al., 2009).

The molecular events leading to the formation of $N_2$-fixing nodules are vast and complex, with many layers of specificity. One of the first layers involves the perception and response to flavonoids and nod chemical signals, synthesized by plant and mutualist, respectively (Spaink, 2000). Flavonoids are host-specific phenolic compounds exuded from roots of legumes that can act as an attractant for receptive rhizobia (Cooper, 2004). Flavonoids can vary between plant species, and certain plants can make several different flavonoid compounds (Pueppke et al., 1998). If perceived by rhizobia, the flavonoid leads to a cascade of signaling events inside the rhizobium. This cascade starts with binding of the flavonoid to the rhizobial NodD protein (Peck et al., 2006). The flavonoid binding activates NodD and allows it to bind to a *cis*-element called the *nod*-box, driving the expression of the downstream "*nod*" genes. Indeed, NodD bound by flavonoids from non-host legumes greatly reduce the ability of NodD to induce expression of the *nod*-box genes (Peck et al., 2006). This first level of specificity requires proper flavonoid/NodD binding for activation, and ensures NodD activation only at the root of a receptive host.

The products of the *nod* genes are nod-factors (NFs), lipochitooligosaccharide signal molecules that trigger the early events of nodule formation in receptive legumes. The NodABC proteins synthesize a basic NF backbone, and other Nod proteins modify and decorate the NF in a manner that also confers specificity through recognition by its corresponding host plant

(Spaink, 2000). These decorations are strain specific, and the appropriate NF is necessary to prime a host plant for symbiosis. Importantly, PTI can be induced upon root perception of a NF from an incompatible rhizobial strain, adding another layer of specificity to the plant-rhizobia symbiosis (Spaink, 2000).

NFs, perceived by plant lysM family extracellular receptors, initiate chemical and morphological changes to the root, including calcium ion spiking, cytoskeletal changes resulting in root hair curling, and eventual internalization of the rhizobia (Segonzac and Zipfel, 2011). NFs also stimulate root cortical cells to begin mitotic division, forming the nodule primordium, which will eventually house the rhizobia (Gage, 2002). The rhizobia move from the curled root hair tip to the root cortex via the infection thread, an inward invagination of the root hair.

Up to this point, the rhizobium is still extracellular to the plant, and must be internalized. Inner root cortex cells endocytose individual bacteria, forming the symbiosome, the beginnings of a microenvironment where $N_2$-fixation can occur (Brewin, 2004). Here, rhizobia differentiate and continue to divide without cytokinesis, resulting in large polyploid bacteroids (Kereszt et al., 2011). In this plant organ called the nodule, plant-derived leghemoglobin protects the bacteroid nitrogenase proteins from environmental oxygen. The rhizobia FixLJ two-component proteins sense the low oxygen environment and induce the genes involved in fixing $N_2$ (Bobik et al., 2006; Kereszt et al., 2011).

## RHIZOBIA AND PLANT DEFENSE

Until recently, it was unclear what role plant defense plays in interactions with rhizobia. Recent evidence suggests that plants respond to rhizobia with a defense response early during interactions, and rhizobia must therefore have mechanisms to suppress host defense (Deakin and Broughton, 2009; Soto et al., 2009; Zamioudis and Pieterse, 2012). For example, analysis of microarrays of *Lotus japonicus* roots nodulated by *Mesorhizobium loti* revealed extensive and early expression of defense genes but subsequent suppression within hours, indicating that even in compatible interactions, rhizobia initially elicit PTI (Kouchi et al., 2004). Additionally, the pathogenesis-related 2 (*PR2*) gene of *M. truncatula* is down-regulated upon perception of NFs from a compatible rhizobia, but is not down-regulated when interacting with a NF-deficient rhizobia strain (Mitra and Long, 2004).

Rhizobia have potentially evolved multiple ways to limit detection by their hosts. For example, rhizobia and many other alpha-proteobacteria do not encode for the potent flg22 PAMP, and thus do not seem to trigger PTI in legumes or arabidopsis (Felix et al., 1999; Gómez-Gómez et al., 1999). Another rhizobial PAMP, the bacterial elongation factor Tu (Ef-Tu), elicits a PTI response in the non-host arabidopsis, but not in legumes (Kunze et al., 2004; Boller, 2005). However, it is unclear whether the absence of Ef-Tu recognition is a result of changes to Ef-Tu, or the lysM receptors as a means to facilitate rhizobial colonization (Segonzac and Zipfel, 2011).

Rhizobia also have mechanisms to cope with, protect against, or even turn host defense responses against the host. The rhizobial polysaccharide layer for example is a potent physical protective barrier. Polysaccharides themselves can dampen host defense. Priming with a host-compatible polysaccharide, for example, can reduce the subsequent HR induced by an incompatible bacterium (Graham et al., 1977; Dow et al., 2000). Furthermore, in pathogens, their polysaccharides also appear to function to chelate calcium ions as a countermeasure against defense signaling by the host. Whether rhizobial-synthesized polysaccharides have analogous functions is unknown (Aslam et al., 2008).

It has been speculated that rhizobia may have mechanisms to turn PTI responses, such as the production of ROS by the host plant, for its gain (Jones et al., 2007). ROS can have negative effects on rhizobia by aborting infection threads (Vasse et al., 1993). Not surprisingly, rhizobia encodes for enzymes to deal with ROS. Mutants of superoxide dismutase or catalase-encoding genes have nodulation defects ranging from ROS sensitivity to an inability to invade a host (Sigaud et al., 1999; Jamet et al., 2003; 2005; Davies and Walker, 2007). However, despite the negative effects of ROS to rhizobia, evidence also suggests that PTI-induced $H_2O_2$ can crosslink plant-derived glycoproteins to the cell walls of the infection thread. This generates a mechanical force that facilitates the elongation inwards of the infection thread (Santos et al., 2001; Brewin, 2004). Similarly, defense-generated Nitric Oxide (NO) is also thought to play a role in nodule function, and has been found in nitrogen-fixing nodules of *Medicago*

*truncatula* (Baudouin et al., 2006). Array studies in *M. truncatula* have shown differential expression of NO-responsive genes during pathogen and rhizobial interactions (Ferrarini et al., 2008).

Finally, rhizobia can modify host defense through suppression of defense responses. Several secretion system-encoding loci have been identified in genomes of many strains of rhizobia. The T3SS is one of the more studied secretion system for mutualistic rhizobia. The nature of the T3SS in plant-rhizobia interactions is slowly unraveling, but it is becoming clear that the use of pathogenic strategies by mutualists is perhaps a widespread mechanism.

## CONCLUSIONS

In this thesis, I describe my work in characterizing the T3E collections of a phytopathogen, commensal, and multiple rhizobia mutualists. In the following chapter, I detail my work in sequencing the genome of the carrot pathogen *Xanthomonas hortorum* pv. *carotae* str. M081. I mined the genome for T3E-encoding genes, and demonstrated the ability of the AvrBs2 and XopQ proteins to induce an HR in *Nicotiana benthamiana* carrying the *Bs2 R* gene or *N. tabacum*, respectively. Therefore, corresponding *R* genes, if present in the carrot germplasm, could be introgressed into commercially-grown carrot cultivars for control against *Xhc*. Furthermore, using genome information, we designed and validated molecular markers to help facilitate molecular detection of this pathogen in carrot seed lots.

In chapter 3, I describe my work in sequencing the genome of *P. fluorescens* WH6. I identified a T3SS-endocing locus suggesting a possible role in host association. However, using a homology-based approach, I found evidence for very few T3E-encoding genes. This work also helped identify coding regions in the GAF metabolic pathway, a compound with potential use for control against grassy weeds.

Chapter 4 details the mining and characterization of candidate T3E proteins from two species of rhizobia for T3SS-dependent translocation. This work represents the largest and most extensive analysis of T3Es from mutualistic bacteria. Our analyses led to the very surprising conclusion that T3E genes in rhizobia were highly conserved. We noted little variation in the content of T3E collections, little variation in their sequences, and little evidence for diversifying selection. This contrasts with T3E collections of pathogens, which exhibit strong evidence for diversifying selection as evidenced by dramatic variation in collection content and variation in sequences across T3E families.

Chapter 5 is a summary of my thesis.

# Genome sequencing and comparative analysis of the carrot bacterial blight pathogen, *Xanthomonas hortorum* pv. *carotae* M081, for insights into pathogenicity and applications in molecular diagnostics

Jeffrey A. Kimbrel, Scott A. Givan, Todd N. Temple, Kenneth B. Johnson, and Jeff H. Chang

Molecular Plant Pathology
2011 vol. 12 (6) pp. 580-594
PMID: 21722296

**SUMMARY**

*Xanthomonas hortorum* pv. *carotae* (*Xhc*) is an economically important pathogen of carrots. Its ability to epiphytically colonize foliar surfaces and infect seeds can result in bacterial blight of carrots when grown in warm and humid regions. We used high-throughput sequencing to determine the genome sequence of isolate M081 of *Xhc*. The short reads were *de novo* assembled and resulting contigs were ordered using a syntenic reference genome sequence from *X. campestris* pv. *campestris* ATCC 33913. The improved, high-quality draft genome sequence of *Xhc* M081 is the first for its species. Despite its distance from other sequenced xanthomonads, *Xhc* M081 still shared a large inventory of orthologous genes, including many clusters of virulence genes common to other foliar pathogenic species of *Xanthomonas*. We also mined the genome sequence and identified at least twenty-one candidate type III effector genes. Two were members of the *avrBs2* and *xopQ* families that demonstrably elicit effector-triggered immunity. We show that expression *in planta* of these two type III effectors from *Xhc* M081 was sufficient for eliciting resistance gene-mediated hypersensitive responses in heterologous plants, indicating a possibility for resistance-gene mediated control of *Xhc*. Finally, we identified regions unique to the *Xhc* M081 genome sequence, and demonstrated their potential in the design of molecular diagnostics for this pathogen.

## INTRODUCTION

Xanthomonads are a diverse group of Gram-negative, γ-proteobacteria. Its members are successful pathogens capable of infecting many agriculturally important crop plants. *Xanthomonas hortorum* pv. *carotae* (*Xhc*, synonyms: *X. campestris* pv. *carotae, X. carotae*; CABI, 2010) causes bacterial leaf blight of carrot, an important disease in most regions of the world. Like most other members of the *Xanthomonas* genus, *Xhc* has the ability to epiphytically colonize foliar surfaces of its host. *Xhc* is also pathogenic and when weather conditions are sufficiently warm and humid, *Xhc* can incite foliar disease which can result in defoliation of host plants with consequential loss of yield.

Several *Xanthomonas* isolates belonging to just a few of the many species of this genera have finished genome sequences, including *X. campestris* pv. *campestris* (*Xcc*), *X. campestris* pv. *vesicatoria* (*Xcv*, syn. *X. vesicatoria*, *X. axonopodis* pv. *vesicatoria*), *X. axonopodis* pv. *citri* (*Xac*; syn. *X. citri* pv. *citri*), and *X. oryzae* pv. *oryzae* (*Xoo;* (Da Silva et al., 2002; Thieme et al., 2005; Salzberg et al., 2008)). Comparisons indicate that most isolates share a high percentage of orthologous genes and long-range synteny (Da Silva et al., 2002; Thieme et al., 2005; Blom et al., 2009). *Xoo* isolates have similar numbers of genes but are exceptional in their genome organization. Indeed, *Xoo* genomes contain the largest number of insertion element families compared to all other sequenced xanthomonads and have undergone significant numbers of large-scale rearrangements (Salzberg et al., 2008).

*X. hortorum* is one of the many less-characterized species within the *Xanthomonas* genus. *X. hortorum* was initially defined based on DNA hybridization studies (Vauterin et al., 1995). Subsequent phylogenetic and multilocus sequence analysis using *gyrB* or four housekeeping genes, respectively, confirmed its isolates represent a distinct species and furthermore, grouped *X. hortorum* with *X. cynarae* and *X. gardneri* to form a diverse clade (Parkinson et al., 2007; Young et al., 2008; Parkinson et al., 2009). Determining the genome sequence for its members will thus be an important contribution to understanding the diversity and evolution of the *Xanthomonas* genus and help resolve this heterogeneous clade.

The process by which *Xhc* infects its host plant is presumed to be similar to that of other foliar pathogens of *Xanthomonas*. Xanthomonads typically gain access to their hosts through natural openings and wounds to colonize and proliferate in the intercellular spaces. Pathogenesis by most xanthomonads is dependent on a type III secretion system (T3SS), a molecular injection apparatus that delivers bacterial-encoded proteins directly into host cells (Büttner and Bonas, 2009). Once inside the host cell, these so-called type III effector (T3E) proteins function to perturb host processes such as PAMP-triggered immunity (PTI; (Jones and Dangl, 2006)).

T3Es also have the potential to elicit plant defenses. In effector-triggered immunity (ETI), direct or indirect perception of a single T3E by a corresponding plant resistance (R) protein results in a robust resistance response (Jones and Dangl, 2006). A classic hallmark of ETI is the hypersensitive response (HR),

which is visualized as a rapid and localized cell death of the infected area (Greenberg and Yao, 2004). Therefore, T3Es are collectively necessary for many Gram-negative pathogens to infect host plants but just a single T3E can limit the host range of a pathogen through its perception by a corresponding host R protein.

The production of commercial carrot crops depends greatly on planting seeds with zero or low contamination of *Xhc*. The semi-arid climates where carrot seeds are produced permit epiphytic growth of *Xhc* (Toit et al., 2005). The bacteria associated with seeds provide the inoculum for bacterial blight when carrots are grown in warm and humid regions. Once the pathogen is established in a crop, suppression of bacterial blight with bactericides is difficult. Sensitive, specific, and facile methods are therefore needed for detecting *Xhc* in seed lots (Meng et al., 2004). Development of cost-effective management methods against *Xhc* is also greatly desired (Toit et al., 2005). A sequenced genome can be an important resource for these purposes and has been used in developing molecular markers to distinguish between two important *Xanthomonas* rice pathogens (Lang et al., 2010).

We present an improved, high-quality draft genome sequence of the M081 isolate of *X. hortorum carotae*. *Xhc* M081 is distinct from other sequenced isolates of *Xanthomonas* and despite the degree of phylogenetic distance from other xanthomonads, *Xhc* M081 shares many orthologous genes and shows high synteny to *Xcc*, *Xac*, as well as *Xcv*. We characterized the genome of *Xhc* to gain insight into its mechanisms of pathogenesis. We describe potential virulence

genes and provide evidence that *Xhc* M081 encodes two members of the ETI-eliciting AvrBs2 and XopQ type III effector families (Kearney and Staskawicz, 1990; Wei et al., 2007). These two T3Es are therefore potential targets for development of control measures against *Xhc* in carrot. Finally, we identified several regions unique to *Xhc*, designed and validated several pairs of primers that specifically amplified products from *Xhc* but not other tested bacteria. These regions have potential use in the design of molecular diagnostics for *Xhc*.

## RESULTS

### Isolation and preliminary typing of *Xanthomonas hortorum* pv. *carotae* M081

Bacteria were isolated from diseased carrot plants (*Daucus carota* L.) grown in a seed production field located near Madras, Oregon USA. We used several phenotypic and molecular markers to preliminarily type M081. This isolate was selected based on its ability to grow on XCS medium, which is semi-selective for *Xanthomonas hortorum* pv. *carotae* (Williford and Schaad, 1984). The colony morphology of M081 grown on YDC medium was also suggestive of it belonging to the *Xanthomonas* genus (data not shown; (Schaad and Stall, 1988)). The 16-23S intergenic spacer region and the elongation factor alpha-encoding gene of M081 had the highest similarities to that of *Xcc* type strain ATCC 33913. Additionally, PCR of M081 with the 3S and 9B primer sets resulted in products that were consistent in size to other isolates of *Xhc* (Meng et al., 2004; Temple and KB, 2009). Finally and most importantly, we were able to show that M081 could achieve large epiphytic populations (>$10^7$ cfu/gm leaf tissue) and cause

symptoms consistent with bacterial blight of carrots in greenhouse and growth chamber experiments (data not shown). We therefore classified isolate M081 as a member of *X. hortorum*, which was later substantiated by phylogenetic analysis (see below).

**Sequencing and assembling an improved, high-quality draft genome sequence**

We used an Illumina IIG to sequence the genome of *Xhc* and generated nearly 30 million paired end (PE) reads, of which approximately 19.1 million and 10.4 million were 32mer and 70mer pairs, respectively. The theoretical coverage of all filtered PE reads was estimated to be 525x, assuming *Xhc* M081 had a genome size of 5.1 megabases. We elected to use a *de novo* approach to assemble the PE reads because of our desire to identify unique regions. We also lacked any data that could direct us to a suitable reference genome for a guided approach to assemble the short reads; previous phylogenetic analyses placed *Xhc* in a clade separate from other *Xanthomonas* species with sequenced isolates (Young et al., 2008; Parkinson et al., 2009).

We sought to develop an improved, high-quality draft genome sequence for *Xhc*. New standards for genome sequences were established in response to next generation sequencing to provide an assessment of quality (Chain et al., 2009). The improved, high-quality standard requires the use of additional work to eliminate discernable misassemblies and resolve gaps and is sufficient for assessing genomes for gene content and comparisons with other genomes.

To this end, we used the software program Velvet version 0.7.55 and a variety of parameter settings to *de novo* assemble the PE reads and generated a

total of approximately 30 different assemblies (Zerbino and Birney, 2008). We identified a single high-quality *de novo* assembly based on consensus support. In other words, we had greater confidence in the quality of this particular assembly because the majority of its contigs were supported by contigs of other assemblies derived using different Velvet parameter settings (Kimbrel et al., 2010). The one high-quality assembly we selected was derived from approximately 20 million PE reads with an actual coverage of approximately 110x and had 153 contigs larger than one kilobase (kb) for a sum total of 5.06 megabases (Mb). The average contig size was approximately 32 kb, the largest contig was 232 kb, and the N50 was 26 contigs; one-half of the genome was represented by the 26 largest contigs.

**Comparative and phylogenomic analyses**

We compared the contigs from *Xhc* M081 to finished genome sequences from representative isolates of *Xanthomonas* to search for a genome sequence with sufficient structural similarities to use as a reference for ordering contigs (Da Silva et al., 2002; Thieme et al., 2005; Salzberg et al., 2008). Our preliminary analysis indicated that the contigs of *Xhc* M081 had sufficient within-synteny to the genome of *Xcc* ATCC 33913 and it was thus used to order all contigs of *Xhc* M081 larger than one kb. The resulting assembly was deemed an improved, high-quality draft genome sequence (Chain et al., 2009). The genome is depicted as a single circular chromosome with physical gaps depicted by red tick marks (Fig. 2.1). We found no evidence for plasmids in *Xhc* M081, which thus far have only been found in *Xac* and *Xcv* (Da Silva et al., 2002; Thieme et al., 2005).

The genome of *Xhc* M081 shares several characteristics with genomes of other representative foliar pathogenic isolates of *Xanthomonas* (Table 1). *Xhc* M081 has a high GC content of 63.7% and the size of its genome was within the range of other sequenced genomes of isolates of *Xanthomonas* (Table 1). We used an automated approach to identify and annotate 4493 coding sequences (CDSs) resulting in a genome coding percentage of 87.4% (Giovannoni et al., 2008; Kimbrel et al., 2010). Given the similarities of the latter two characteristics to other *Xanthomonas* genomes, we are confident that the majority of the *Xhc* M081 genome is present in the assembly and our automated annotation approach was acceptable.

With the exception of *Xoo*, genomes of the foliar pathogenic xanthomonads share long-range synteny (Da Silva et al., 2002; Thieme et al., 2005; Blom et al., 2009). To determine whether *Xhc* M081 had similar syntenic relationships, we parsed the genome sequence of *Xcc*, *Xac*, *Xcv*, and *Xoo* into all possible 25mer DNA sequences and aligned the unique 25mers to the genome sequence of *Xhc* M081 (Fig. 2.2). Our results confirmed our initial findings by showing the greatest synteny to the genome of *Xcc*. The genome of *Xhc* M081 was also syntenic to the genomes of *Xac* as well as *Xcv* with the exception of a few large insertion-deletions (indels) and inversion events that appeared to be localized to the predicted terminator sequence (data not shown). Similar to others, the genome of *Xhc* showed little structural similarity to the genome of *Xoo*.

The *Xhc* M081 genome encodes a high percentage of proteins orthologous to proteins of other species of *Xanthomonas*. Reciprocal best hit

analysis using BLASTP showed that 74.2%, 73.8%, 72.9%, and 61.1% of the proteins encoded by *Xhc* M081 were also found in *Xcc*, *Xac*, *Xcv*, and *Xoo*, respectively. The greater than 70% orthology was similar to levels observed in previous comparisons between different *Xanthomonas* species (Da Silva et al., 2002; Thieme et al., 2005; Blom et al., 2009). The lower amount of orthology to *Xoo* was not surprising considering that *Xoo* appears to be the most distinct of the sequenced foliar pathogenic isolates.

We used two approaches to examine the relationship of *Xhc* M081 to other xanthomonads. In the first, we used a phylogenomic approach to determine the species relationship of *Xhc* M081 to representative isolates of *Xanthomonas* with finished genome sequences. We identified a core of 1776 translated sequences common to the 10 examined isolates and produced a species tree based on the comparison of a super alignment from their sequences (Fig. 2.3A). Each of the previously sequenced species grouped as expected with *Xhc* M081 forming a branch by itself. We used multilocus sequence analysis (MLSA) to examine the relatedness of *Xhc* M081 to other isolates of *X. hortorum* (Fig. 2.3B; complete tree is provided as figure 2.6; (Young et al., 2008)). *Xhc* M081 grouped with other pathovars of *X. hortorum* within the heterogeneous clade that also includes *X. cynarae* and *X. gardneri* isolates (Young et al., 2008).

**Candidate virulence genes of *Xhc* M081**

We identified several clusters of virulence genes important for pathogenesis by xanthomonads. Xanthan is an exopolysaccharide produced by xanthomonads with important roles in biofilm formation and pathogenesis (Katzen

et al., 1998). Synthesis is dependent on a cluster of 12 *gum* genes, *gumB-gumM*. We identified a similarly arranged cluster present on a single contig in *Xhc* M081 (XHC_2807 to XHC_2795), flanked by homologs of *gumN-P* on one side and a tRNA-encoding gene on the other. The presence of the tRNA-encoding gene has been implicated as evidence for acquisition of this gene cluster by horizontal gene transfer (Lu et al., 2008). We did not find any evidence of insertion sequences in this cluster.

The *rpf* (regulation of pathogenicity factors) cluster of genes encodes positive regulators of extracellular enzymes and proteins that synthesize and perceive an intracellular diffusible factor (DSF) cis-11-methyl-2-dodecenoic acid, important for virulence (Wang *et al.*, 2004). Four genes are demonstrably important; RpfB and RpfF are involved in biosynthesis of the DSF whereas RpfG and RpfC are hypothesized to perceive DSF. *Xcc* mutants of *rpfF* and *rpfC*, for example, are compromised in manipulating plant stomatal closure, an important step during infection (Gudesblat et al., 2009). *Xhc* M081 encodes a cluster of approximately 23 kb with eight genes homologous to *rpf* genes, including *rpfB, F, G,* and *C* that were all present on one contig.

Other examples of homologous virulence genes included the four-gene operon of *hmsHFRS* that encode hemin storage systems involved in biofilm formation in *Yersinia pestis* (Abu Khweek et al., 2010). *Xhc* M081 has homologs of *wzm* and *wzt* involved in synthesis of lipopolysaccharides (Rocchetta and Lam, 1997). However, like *X. fuscans* subsp. *aurantifolii* types B and C, the *wzt* gene of *Xhc* M081 appears to encode a C-terminal truncated protein, which is predicted to

be affected in substrate binding (Cuthbertson et al., 2005; Moreira et al., 2010). *Xhc M081* encodes a homolog of *yapH*, a plant adhesion protein, *hlyD* and *hlyB* genes for hemolysin secretion, and a homolog of the *Pseudomonas aeruginosa asnB* gene, involved in asparagine and O-antigen biosynthesis (Holland et al., 2005; Augustin et al., 2007; Das et al., 2009).

The type IV secretion system (T4SS) is an apparatus used by many bacteria to interact with their hosts (Alvarez-Martinez and Christie, 2009). Clusters of T4SS-encoding genes have been identified in genome sequences of xanthomonads but the role of the T4SS in virulence of xanthomonads is still unclear (Da Silva et al., 2002). *Xhc* M081 appears to encode for a T4SS (XHC_2815-XHC_2830) but it was difficult for us to determine whether the T4SS is functional because the cluster of genes was distributed across three separate and adjoining contigs with physical gaps that corresponded to what appears to be three different copies of *virB4*. Whether the three copies are *bona fide* or an artifact of misassembly is unresolved. The three contigs that encode the T4SS were approximately 10 kb away from the *gum* cluster. In the genome of *Xcc* ATCC 33913, which was used as a reference to order the contigs of *Xhc* M081, these two gene clusters are nearly 26 kb apart (Da Silva et al., 2002).

The type III secretion system (T3SS) is another apparatus used by many Gram-negative pathogens to interact with their hosts. In plant pathogens, the T3SS is required for pathogenesis and is encoded by a single cluster of genes called the *hrp* genes (Niepold et al., 1985; Lindgren et al., 1986). All *hrp* genes of *Xhc* M081 (CDSs XHC_1407 to XHC_1426) except for *hrcC,* were found in their

entirety and clustered on a single contig of 25 kb in length. Approximately 50 nt of the C-terminal coding portion of *hrcC* was absent. We used PCR with primers that spanned the region and sequencing of the product to very that *hrcC* was indeed complete. The organization of the *hrp* cluster in *Xhc* was identical to corresponding *hrp* genes of other *Xanthomonas* foliar pathogens. Additionally, we did not identify any polymorphisms in the CDSs that would overtly affect function (data not shown). The T3SS of *Xhc* is consequently predicted to be complete and functional.

We mined the *Xhc* M081 genome sequence for candidate T3E genes. In *Xanthomonas* and *Ralstonia* spp., T3E genes are often preceded by a *cis* regulatory motif recognized by the transcriptional regulator HrpX (Mukaihara et al., 2004; Koebnik et al., 2006). We identified 118 putative Plant-Inducible-Promoter boxes (PIP-box) in the genome of *Xhc* M081. Of these 118, only 38 had a CDS within 300 bp of the putative PIP-box. We further eliminated 26 CDSs because their translated sequences were homologous to proteins with functions atypical of T3Es. We also used BLASTP searches to find ten more CDSs with translated sequences homologous to known T3Es (Table 2).

Three of the candidate T3E genes required additional characterizations because of evidence for potential assembly artifacts. Two CDSs both with homology to *xopX* were found in tandem in one contig. The translated sequences of the *xopX* homologs were 61% identical (77% similar) to each other and the genes have a similar arrangement in *Xcc* ATCC 33913, suggesting that this was not an artifact of short-read assembly (White et al., 2009). Nonetheless, we used

PCR of *Xhc* M081 genomic DNA and sequencing of the product to confirm the presence of tandem copies of *xopX*. Similarly, we found two CDSs with homology to *xopR* in tandem on a contig. One of the CDSs appeared to be full-length relative to *xopR* of *Xcc* ATCC 33913. The other CDS had weaker homology to *xopR* and was shorter. Both *xopR* homologs, however, had putative PIP-boxes less than 100 bp upstream of their predicted start codons suggesting the two to be separate candidate T3E genes. We again confirmed this arrangement using PCR and sequencing of the amplified product (data not shown). Finally, *xopAD* was found on more than one contig. We speculate that the repeats, ~126 bp in length, were difficult to assemble. The repeats also made it difficult to design specific primers for PCR and gap closure. Therefore, we were unable to determine if $xopAD_{Xhc}$ encodes a full-length product, or is a pseudogene similar to *xopAD* of *Xcv* (White et al., 2009). Not including *xopAD*, *Xhc* encodes at least 21 candidate T3Es with three unique to *Xhc* M081.

Two of the candidate T3E genes of *Xhc* M081 are highly homologous to demonstrable ETI-eliciting T3Es. AvrBs2 (88% identity) was first characterized in *Xcv* and is very widespread based on a survey of various races of *Xcv* and other pathovars of *X. campestris* (Kearney and Staskawicz, 1990). Given the prevalence of *avrBs2* in *Xanthomonas* spp., it was not at all surprising that at least 10 kb of DNA sequence flanking either side of this T3E gene in *Xhc* M081 was also conserved in *Xcc*, *Xcv*, *Xac* and *Xoo*. Furthermore, the regions surrounding and including *avrBs2* had a GC% that was not significantly different from the

genome average of 63.7%. This observation suggests that *avrBs2* was likely present in an ancestor common to the foliar pathogenic species of *Xanthomonas*.

XopQ (92% identical) is also prevalent in xanthomonads. XopQ is a member of the HopQ1-1 family of T3Es first discovered in *Pseudomonas syringae* pv. *tomato* DC3000 (Guttman et al., 2002; Petnicki-Ocwieja et al., 2002; Chang et al., 2005; Wei et al., 2007). In contrast to $avrBs2_{Xhc}$, comparisons of the five kb of DNA sequences flanking *xopQ* of *Xhc* M081 showed a variable pattern that was difficult to interpret. We also noted that *xopQ* resides in a 20 kb region with an average GC content of 55%, a large deviation from the *Xhc* M081 genome average of 63.7%. Finally, the CDS just upstream of $xopQ_{Xhc}$ encodes an IS4 family transposase, the presence of which appears to be unique to *Xhc* M081. In total, these data suggest that the different foliar pathogenic *Xanthomonas* spp. we examined may have acquired *xopQ* independently after they diverged from their common ancestor.

We determined whether $AvrBs2_{Xhc}$ and $XopQ_{Xhc}$ could elicit a hypersensitive response (HR) in plants. The *R* gene corresponding to *avrBs2* has been cloned (Tai et al., 1999). *Bs2* encodes nucleotide binding and leucine-rich repeat motifs typical of many resistance proteins and *Bs2* is sufficient to elicit an HR in transgenic *N. benthamiana* plants when co-expressed with *avrBs2* (Tai et al., 1999). HopQ1-1 elicits an HR in wild-type *N. benthamiana* as well as *tabacum*. The *R* gene corresponding to *hopQ1-1* has yet to be identified but nevertheless could provide another opportunity for potential control against *Xhc*.

Neither of the *R* genes has been identified in carrots so we elected to use tobacco and *Agrobacterium tumefaciens*-mediated transient expression to test for elicitation of the HR. *Agrobacterium* carrying a CaMV 35S-expressing *avrBs2* from *Xhc* M081 caused a rapid HR 24 hours post infiltration (hpi) in transgenic *N. benthamiana* constitutively expressing Bs2 (Fig. 2.4; (Tai et al., 1999)). In contrast, no phenotypes were visible in wild-type *N. benthamiana* lacking *Bs2* following infiltration with *Agrobacterium* carrying the same DNA construct. Transient expression of CaMV 35S-expressing *xopQ* and *hopQ1-1* cloned from *P. syringae* pv. *tomato* DC3000 also resulted in strong, rapid HRs in 75% and 71%, respectively, of infiltrated leaves of wild-type *N. tabacum*. No phenotypes were observed following challenge of tobacco plants with *Agrobacterium* lacking T3E genes (data not shown). These results suggest that $avrBs2_{Xhc}$ and $xopQ_{Xhc}$ are sufficiently similar to their original founding family members for their translated products to be perceived by a corresponding R protein and elicit ETI. We cannot, however, exclude the possibility that HopQ1-1 and XopQ are perceived by two different R proteins of tobacco, though both appeared to elicit ETI in an age-dependent manner, with more robust HRs in older leaves.

**Development of molecular markers for *Xhc***

As an important step towards developing molecular markers for diagnosing *Xhc* contamination in lots of carrot seeds, we searched the *Xhc* M081 genome for unique regions based on comparisons to genomes of other *Xanthomonas* species. Over 500 kb of sequences distributed over 171 different regions larger than one kb were identified (Fig. 2.1, track 5). We focused our efforts on 16

regions based on the criteria of size, location in the genome relative to each other, and uniqueness to *Xhc* based on BLASTN results to the NCBI nt database. Using these 16 regions as templates, we designed 18 different primer pairs (XhcPP = *X. hortorum* pv. *carotae* Primer Pair). We used PCR of genomic DNA from *Xhc* M081 as a template to test these primers. Seven pairs led to amplifications of expected-sized products (Fig. 2.5). PCR with XhcPP05 resulted in a second fragment but it appeared to amplify with far less efficiency and was larger in size than the expected product (not shown). PCR with XhcPP14 also resulted in a less abundant second product that was smaller in size.

We tested the seven primers pairs in PCR with genomic DNA from *Xhc* isolates found in four world-wide regions that produce carrot seeds to determine whether the primer pairs are broadly applicable for *Xhc* (Table 3). PCR with primer pairs XhcPP08, 13 and 14 yielded different sized products or failed to amplify a fragment from all tested *Xhc* seed isolates. In contrast, PCR with primer pairs XhcPP02, 03, 04, and 05 amplified an expected sized fragment from all tested *Xhc* seed isolates (see also fig. 2.1; asterisks track 5).

We also tested the seven primer pairs, or a subset of them, against available type strains, related strains, and 23 isolates of bacteria from 14 genera that are commonly associated with plant surfaces. None of the seven tested pairs amplified a fragment from genomic DNA extracted from *X. campestris* pvs. *campestris* ATCC 33913, *vesicatoria* MTV1, *coriander* ID-A (*Xccor* ID-A) or *X. hortorum* pv. *hederae* (*Xhh*) ATCC 9653 (Table 3). We did note a very faint band of approximately 600 bp with XhcPP03 when tested with DNA from *Xhh*. The

inability of our primer pairs to amplify products from *Xccor* ID-A was encouraging because the 3S and 9B primer pairs could not distinguish between *Xccor* ID-A and isolates of *Xhc* (Meng et al., 2004; Poplawsky et al., 2004). PCR with XhcPP02, 03, 04, and 05 failed to yield a single product of expected size from any of the 23 bacterial isolates associated with plant surfaces (Table 2.6). Reactions with XhcPP04 resulted in multiple, faint, and non-specific amplified products from many of the tested isolates. Therefore, amongst the bacteria that we tested, the primer pairs, XhcPP02, 03, 04, and 05 appeared to be specific to *Xhc*.

We did *post hoc* analysis to determine if regions that XhcPP02-05 annealed to corresponded to any CDSs. One of the XhcPP02 primers annealed to XHC_2981 (TetR regulator) and the other annealed to an intergenic region. Genes homologous to XHC_2981 are present in other soil bacteria, but none were detected in genomes of xanthomonads. XhcPP03 annealed to XHC_4113 and XHC_4114 (both are annotated as “hypothetical”). Homologs are present in other xanthomonads but unlike *Xhc* M081, are not clustered and not within an amplifiable distance. XhcPP04 annealed to XHC_4117 (membrane fusion protein). Here too, homologs are present in other xanthomonads but we observed very little nucleotide homology. Finally, XhcPP05 annealed to XHC_4175 (“hypothetical”), which is unique to *Xhc* M081, and XHC_4176 (patatin-like protein). The genomic regions corresponding to these four primer pairs are thus strong candidates for use in developing molecular diagnostic tools for detecting contamination of *Xhc* on carrot seeds, irrespective of the global regions in which the seeds are produced.

## DISCUSSION

The semi-arid climate of the Pacific Northwest region of the United States and Canada is ideal for carrot seed production. In this region, epiphytic populations of *Xanthomonas hortorum* pv. *carotae* on umbels can infect developing seeds without eliciting foliar disease. As a consequence, *Xhc* can be unknowingly abundant on seeds harvested from production fields, which then serves as the inoculum for bacterial blight on carrots grown in other regions more conducive for disease development (Toit et al., 2005). The asymptomatic 'epidemic' of epiphytic colonization of the seed crops frequently necessitates hot water treatment of seed lots to reduce the seed-borne populations of *Xhc*. Hot water treatment, however, is expensive and potentially injurious to seeds.

Diagnosing seed lots for contamination by *Xhc* is therefore critical but often time-consuming owing to the common use of dilution plating to enumerate *Xhc* before and after treatment. As a step towards developing confident and facile detection methods for *Xhc*, we used an Illumina IIG to determine the genome sequence of *Xhc* M081, an isolate found on infected carrots grown in central Oregon. Short-reads were *de novo* assembled and contigs ordered based on a syntenic reference genome sequence to develop an improved, high-quality draft genome sequence. The genome sequence of isolate M081 of *Xhc* is thus the first for this clade and is important towards filling in the gaps along the phylogenetic tree of *Xanthomonas* for understanding evolutionary relationship and genetic diversity within this genus of important plant pathogens.

The observed long-range synteny to representative isolates of foliar xanthomonads except *Xoo* could simply be a consequence of our efforts to use the *Xcc* ATCC 33913 genome to order the contigs of *Xhc* M081. Several points suggest otherwise. The genome of *Xhc* M081 was *de novo* assembled and synteny to examined genomes of *Xanthomonas* species was found both within and across its contigs, with the exception of *Xoo*. Furthermore, the majority of genomic regions greater than one kb and unique to *Xhc* as well as the regions inverted relative to *Xac* and *Xcv* were wholly contained within contigs. Finally, analysis of GC skew showed a general bias of guanine in the leading strand indicating that there was no overt incorrect ordering of contigs (Fig. 2.1; (Rocha, 2004; Arakawa and Tomita, 2007)). Together, these observations justified the use of the *Xcc* ATCC 33913 genome to order the contigs of *Xhc* M081 and indicated that the observed synteny is a true reflection of similarities in genome structures rather than an artifact of our *in silico* efforts to improve the genome.

The high-quality draft genome sequence of *Xhc* was derived from short reads and improved using only *in silico* approaches. This requires an acknowledgement of potential limitations. Unlike finished genomes, the draft assembly of *Xhc* still had a considerable number of ambiguous bases, which we did not attempt to resolve. These did not appear to have a significant effect on our analysis because the number of annotated CDSs was similar to other xanthomonads. Repeated sequences such as tandem repeats or duplicate regions in the genome are particularly challenging for short-read assembly (Pop and Salzberg, 2008). For example, *Xcc* ATCC 33913 is reported to have two

rRNA-encoding operons; we only identified one in *Xhc* M081. We suspect our sequenced isolate also has two rRNA-encoding operons but they collapsed into one contig.

Analysis of the *Xhc* M081 genome provided insights into mechanisms of pathogenesis. We found a complete *gum* gene cluster, which is hypothesized to have been acquired by horizontal gene transfer and an important acquisition for the evolution of pathogenesis by xanthomonads (Lu et al., 2008). Similarly, *Xhc* M081 encoded for a cluster of *rpf* genes, which is not surprising considering that these genes were likely present early in the evolution of Xanthomonadaceae (Lu et al., 2008). We also found clusters of genes that encode a type IV and type III secretion system. The T4SS-encoding gene cluster of *Xhc* M081 was fragmented and we hesitate in speculating on its functionality. In contrast, inspection of the T3SS-encoding region suggested it to be functional and the necessity for T3SS in pathogenesis by foliar pathogens of *Xanthomonas* is well demonstrated. The T3SS delivers type III effector proteins with demonstrable roles in dampening and eliciting host defense (White et al., 2009). We identified 21 candidate type III effector genes and the products from two, AvrBs2$_{Xhc}$ and XopQ$_{Xhc}$ elicited HRs when transiently overexpressed in *Nicotiana* spp.

These so-called ‘avirulence’ proteins are potential targets for the development of carrot cultivars resistant to *Xhc* through introgression of the corresponding *R* genes, assuming *Bs2* and the *R* gene corresponding to *xopQ* are present in the carrot germplasm. It is, however, unclear whether resistance gene-mediated control can provide durable control. *AvrBs2* is nearly ubiquitous

and is required for full virulence by foliar *Xanthomonas* pathogens. However, Bs2 resistance in pepper incurs such a strong negative selective pressure that *Xcv* isolates with mutations in *avrBs2* with little to no cost in virulence are prominent (Swords et al., 1996; Gassmann et al., 2000; Leach et al., 2001).

We found no evidence for genes of *Xhc* M081 that encode for members of the transcriptional-activator like (TAL) family of type III effectors (Boch et al., 2009; Moscou and Bogdanove, 2009). TAL effectors are characterized by a variable number of amino acid repeating motifs, nuclear localization signals and acidic transcriptional activation domain. The repeated sequences could be difficult to accurately assemble, but hallmarks of TAL effector genes should still be detectable. We searched but failed to identify any hallmarks of TAL-encoding genes and thus suspect the absence of TAL-encoding genes from *Xhc* M081 is a true reflection of its repertoire of type III effectors rather than of the limitations of short-read assembly.

The genome sequence of *Xhc* M081 will be useful for developing molecular detection methods for diagnosing *Xhc* contamination on carrot seeds. We developed primer pairs XhcPP02-05 that specifically amplified a fragment of DNA from eight globally dispersed isolates of *Xhc* but not from other species of *Xanthomonas*, another pathovar of *X. hortorum*, or other bacteria commonly associated with plants. Practical implementation of these primers, however, will require additional testing against a larger collection of bacteria associated with carrot seed. Furthermore, to advance molecular diagnostics for *Xhc*, we intend to use the genomic regions corresponding to these primer pairs to develop primers

for use in loop-mediated isothermal PCR (LAMP; (Mori and Notomi, 2009; Temple and KB, 2009)). This method shows tremendous potential because it is quick to perform and obviates the need for dilution plating. LAMP also shows superior performance to quantitative PCR because of its robustness in the presence of PCR inhibitors (Temple and Johnson, unpublished data), and can be performed in various conditions/facilities with limited equipment and resources (Mori and Notomi, 2009).

The improved, high-quality draft genome sequence of *Xhc* M081 also has potential use towards molecular typing of *Xhc*. Primer pairs XhcPP08, 13, and 14 yielded similarly-sized PCR products from isolates from the Northern hemisphere (*Xhc*PNW1, *Xhc*PNW2 and *Xhc*Fr1). In contrast, these primer pairs resulted in polymorphic banding patterns when diagnosing isolates of *Xhc* from the Southern hemisphere. Primer pairs XhcPP13 and XhcPP14 failed to yield a product from *Xhc*_Ch1, while primer pairs XhcPP08 and XhcPP13 failed to yield a product from *Xhc*_Ch3. Primer pair XhcPP13 failed to amplify a fragment of DNA and primer pair XhcPP08 yielded a larger sized product from isolate *Xhc*_Ar1.

The standard we attempted to adhere to is considered sufficient for genome mining and comparative approaches (Chain et al., 2009). Genome sequencing and comparative genomic analysis have been done for numerous bacterial plant pathogens to identify virulence factors, better understand phylogenetic relationships among closely-related bacterial species, and identify sequences of DNA novel to a species for potential application to molecular detection technologies. Our sequencing and generation of an improved, high-

quality draft genome sequence for an isolate of *X. hortorum* has strong potential for developing diagnostic tools for management of *Xhc* and provided a greater understanding of this economically important bacterial pathogen.

## EXPERIMENTAL PROCEDURES

### Bacterial strains and plasmids

All bacterial strains and plasmids used in this study are listed in Table 4. *Xanthomonas* isolates and *Pseudomonas syringae* pv. *tomato* DC3000 were grown in King's B media at 28°C. *E. coli* and *Agrobacterium tumefaciens* GV2260 were grown in Luria-Bertani (LB) media at 37°C and 28°C, respectively. The following concentrations of antibiotic were used: 50 μg/ml rifampicin (100 μg/ml for *A. tumefaciens*), 25 μg/ml gentamycin, 30 μg/ml kanamycin (100 μg/ml for *A. tumefaciens*), and 50 μg/ml cycloheximide.

### Molecular techniques

To prepare genomic DNA for high-throughput sequencing, we isolated genomic DNA from *Xhc* M081 using osmotic shock and alkaline lysis followed by a phenol/chloroform extraction. DNA was prepared for Illumina sequencing according to the instructions of the manufacturer (Illumina, San Diego, CA). We used paired-end sequencing of 36mers (3 channels) and 76mers (1 channel).

To test primer pairs for molecular diagnostics of *Xhc*, PCRs were carried out in 25 μl reaction mixtures containing 1 X ThermoPol Buffer (20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2.0 mM $MgSO_4$, 0.1 % Triton X-100, pH 8.8 @ 25°C), 250 μM of dNTPs, 3.0 mM $MgCl_2$, 1.0 μM of each primer (Table 5), 2.0 units *Taq* and *Pfu* DNA polymerase (25:1 mixture), and 2.5 μl of

template genomic DNA. Cycling parameters for PCR were: 94°C for 2 min, followed by 35 cycles of 94°C for 15 sec, 62.7°C for 20 sec, and 72°C for 45 sec, with a final extension of 72°C for 1 min.

To clone candidate type III effector genes, we used primer pairs in a two-step PCR for the CDSs of *avrBS2* and *xopQ* ((Chang et al., 2005); Table 5). DNA fragments were cloned into pDONR207 using BP Clonase following the instructions of the manufacturer (Invitrogen, Carlsbad, CA). The CDS for *hopQ1-1* was previously cloned in pDONR207 (Chang and Dangl, unpublished). All three CDSs were cloned into pGWB14 using LR Clonase ((Nakagawa et al., 2007); Invitrogen, Carlsbad, CA).

To prepare PCR fragments for Sanger sequencing, products were treated with ExoI and SAP for 20 minutes at 37°C followed by 40 minutes at 80°C. Sanger and Illumina sequencing were done at the Center for Genome Research and Biocomputing Core Labs (CGRB; Oregon State University, Corvallis, OR).

**Genome assembly**

The last four and six bases were trimmed off from the 36mer and 76mer reads, respectively, and all short reads that had ambiguous bases, as well as its paired read, were removed. We used Velvet 0.7.55 to *de novo* assemble the reads, and the highest-quality assembly was identified using methods previously described (Zerbino and Birney, 2008; Kimbrel et al., 2010). We used Mauve Aligner 2.3 (default settings) and the genome sequence of *X. campestris* pv. *campestris* str. ATCC 33913 as a reference to order the *Xhc* M081 contigs (Da Silva et al., 2002; Rissman et al., 2009). We used an automated method, as

previously described, to annotate the improved high-quality draft genome sequence of *Xhc* M081 (Delcher et al., 1999; Giovannoni et al., 2008; Kimbrel et al., 2010).

**Bioinformatic analyses**

Circular diagrams were plotted using DNAplotter ((Carver et al., 2009); http://www.sanger.ac.uk/Software/ Artemis/circular/).

Synteny plots were generated by first identifying all the unique 25mers present within the genome sequences for Xcc str. ATCC 33913, Xac str. 306, Xcv str. 85-10, and Xoo PXO99A (Da Silva et al., 2002; Thieme et al., 2005; Salzberg et al., 2008). We used CASHX to map the unique 25mers to the genome sequence of *Xhc* M081 and R to display perfect matching 25mers relative to their coordinates in the respective genomes (Fahlgren et al., 2009; R Development Core Team).

Phylogenomic relationships of *Xhc* M081 to other isolates of *Xanthomonas* were determined using HAL (Table 4; (Robbertse et al., 2006); http://aftol.org/pages/Halweb3.htm). The tree was visualized using the Archaeopteryx and Forester Java application (http://www.phylosoft.org/archaeopteryx/ (Zmasek and Eddy, 2001)).

For MLSA, we extracted partial sequences of *dnaK*, *fyuA*, *gyrB*, and *rpoD* from the genome sequence of *Xhc* M081. Sequences were concatenated and used in comparisons with corresponding sequences as previously determined (Young et al., 2008). Neighbor-joining trees were generated with 1000 bootstrap replicates.

To identify regions unique to *Xhc* M081, we used BLASTN to compare the *Xhc* M081 genome to *Xcc*, *Xac*, *Xcv* and *Xoo* (e-value > 1 x $10^{-7}$). Contiguous regions larger than 1 kb were used as queries in BLASTN searches of the NCBI nr/nt database. The NCBI Primer-BLAST was used to design PCR primers specific to these regions of the *Xhc* M081 genome by excluding those that could potentially amplify known xanthomonad sequences (taxid:338; http://www.ncbi.nlm.nih.gov/tools/primer-blast/).

**Identifying candidate type III effector genes**

We used a Perl regular expression to search for the PIP-box sequence, TTCGB-$N_{15}$-TTCGB, where B is any base except adenine (Fenselau and Bonas, 1995; Tsuge et al., 2005). CDSs were classified as candidate type III effectors if they met both of the following criteria. The CDS must be no more than 300 bp downstream of an identified PIP-box. Secondly, the translated sequence of the CDS must have no or low homology to sequences with annotated functions not normally associated with type III effector proteins. Amino acid sequences of confirmed type III effectors from xanthomonads and other phytopathogens were also used as queries in BLASTP searches (e-value < 1 x $10^{-7}$) against the 4,493 *Xhc* M081 translated CDSs (www.xanthomonas.org; (White et al., 2009; Kimbrel et al., 2010)).

***Agrobacterium*-mediated transient expression**

Binary vectors carrying candidate type III effector genes were mobilized into *A. tumefaciens* GV2260 via three-way conjugation. Bacteria were grown overnight in King's B media, washed in 10 mM $MgCl_2$, and resuspended to an $OD_{600}$ of 1.0. A blunt syringe was used to infiltrate bacteria into leaves of six-week

old wild-type *Nicotiana tabacum*, *Nicotiana benthamiana*, or transgenic *N. benthamiana* constitutively expressing the *Bs2* resistance gene (Tai et al., 1999). Leaves were scored 24 hpi.

Plants were maintained in a growth chamber cycling 9 hours light/25°C daytime, and 15 hours dark/20°C.

## ACKNOWLEDGEMENTS

We thank Chris Sullivan and Mark Dasenko in the Center for Genome Research and Biocomputing for computational support and DNA sequencing, Philip Hillebrand for his assistance, Jason Cumbie and Dr. Joey Spatafora for their advice, Dr. Brian Staskawicz for providing transgenic *N. benthamiana* seeds and Dr. Virginia Stockwell and Dr. Joyce Loper for providing their culture collection. *Xhc* M081 was originally isolated by Fred Crowe at Oregon State University, Central Oregon Research and Extension Center, in Madras, OR. *Xhc*-infested carrot seed lots, *Xcv* strain MTV1 and *Xcc*or ID-A were provided by Lindsey du Toit and Mike Derie of Washington State University, Mt Vernon, WA. This research was supported in part by startup funds from OSU and the National Research Initiative Competitive Grant no. 2008-35600-18783 from the USDA's National Institute of Food and Agriculture, Microbial Functional Genomics Program to JHC and the California Fresh Carrot Advisory Board to KBJ.

This Whole Genome Shotgun project has been deposited at DEBJ/EMBL/GenBank under the accession AEEU00000000. The version described in this paper is the first version, AEEU01000000

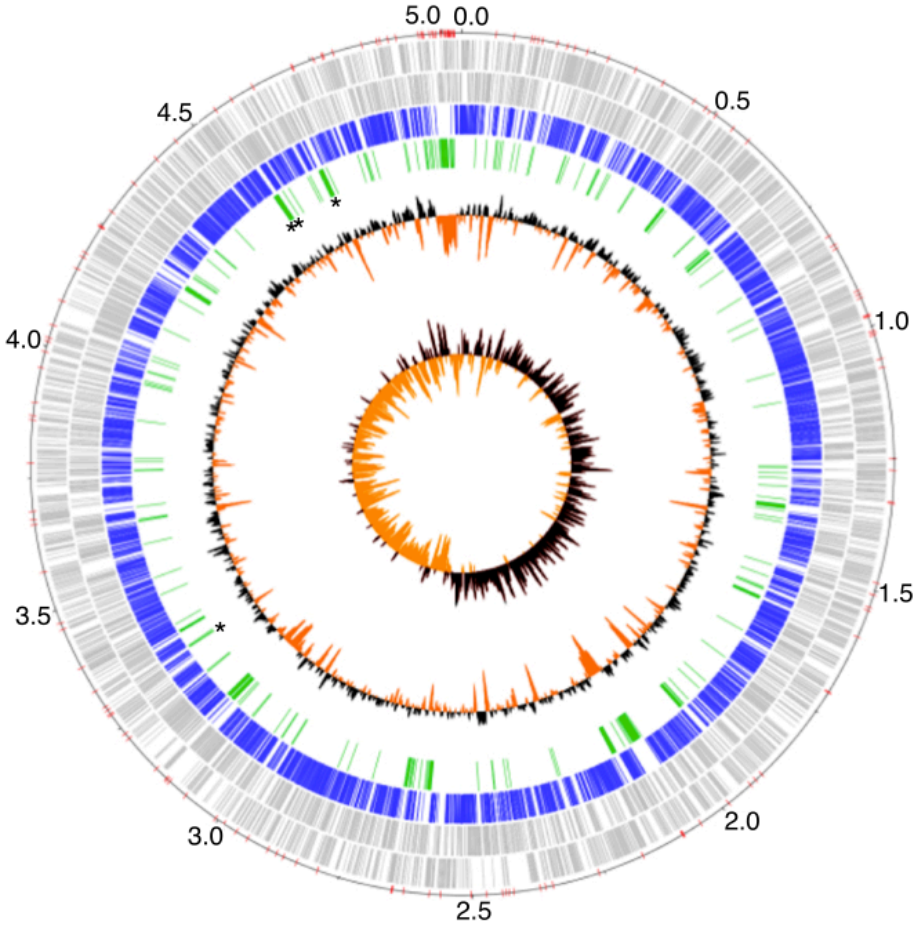



**Figure 2.1. Circular representation of the improved, high-quality draft genome sequence of *Xhc* M081.**

**Circle 1** (outer) designates the coordinates of the genome in half million base pair increments (red tick marks denote contig breaks); **circles 2** and **3** show predicted CDSs as gray lines on the positive and negative strands, respectively; **circle 4** shows CDSs with homology to at least one other gene in other xanthomonads (e-value < 1 x $10^{-7}$); **circle 5** shows regions of *Xhc* M081 larger than one kb with no detectable homology to genomes of other xanthomonads; asterisks highlight the locations for the *Xhc*-specific primer sets XhcPP02 (between genome coordinates 3.0-3.5) XhcPP03-05 (in numerical order starting near genome coordinate 4.5; see also tables 2, 5 and fig. 4); **circle 6** shows deviations from the average GC% of 63.7% (black = greater and orange = lower); and **circle 7** shows GC skew with bias for and against guanine as black and orange, respectively.

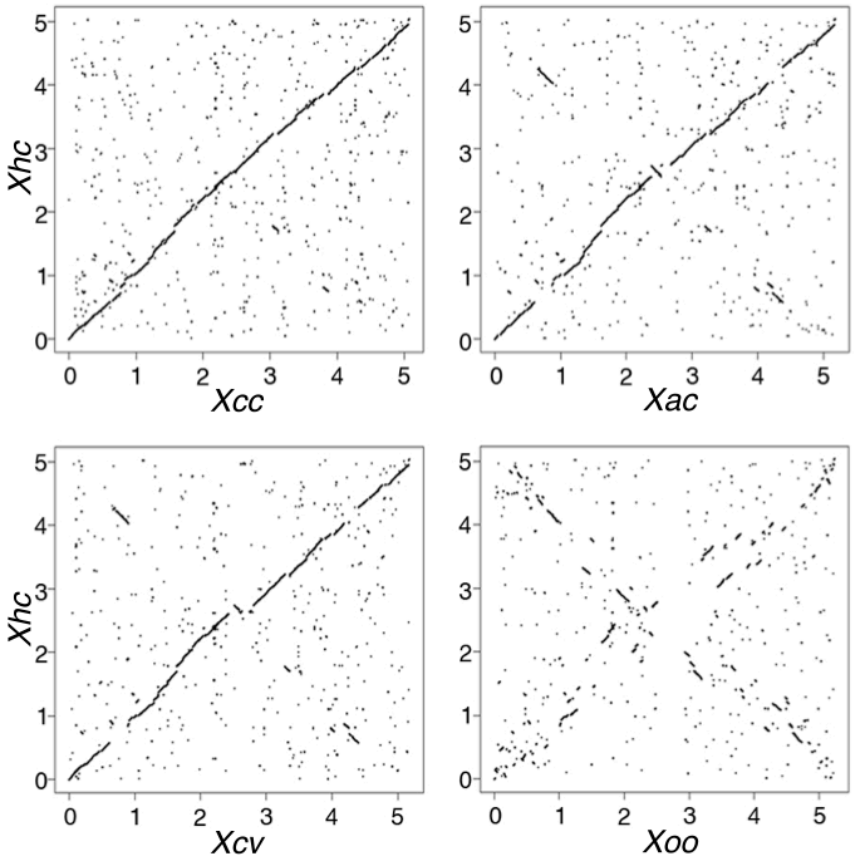



**Figure 2.2. Synteny plots comparing the genome structure of *Xhc* M081 to genomes of other *Xanthomonas* species.**

Unique 25mers from *Xcc*, *Xac*, *Xcv*, and *Xoo* were compared to the improved, high-quality draft genome sequence of *Xhc* M081. The start positions of all matching pairs were plotted in an XY graph with the coordinates of the genome of *Xhc* M081 along the y-axis and the coordinates of the genomes of *Xcc*, *Xac*, *Xcv*, and *Xoo* along the x-axis (see table 4). The termini are located at the approximate mid-way point for each comparison. Genome scales are shown in one Mb increments.

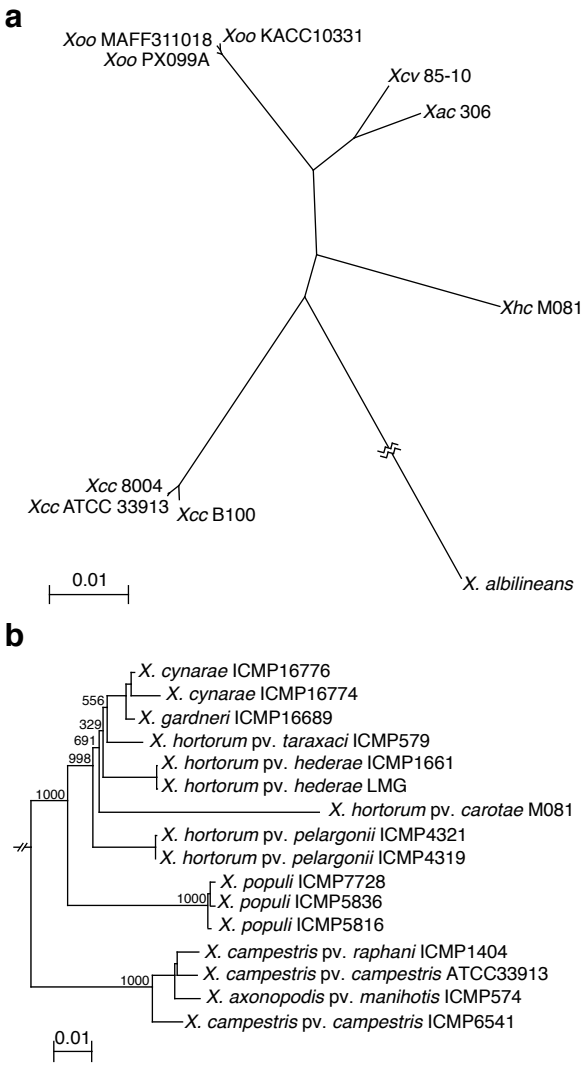



**Figure 2.3. Isolate M081 groups with *X. hortorum*.**

**a)** Unrooted phylogenomic tree of ten *Xanthomonas* isolates based on a super alignment of 1776 translated sequences. Abbreviations are as described in the text and accession numbers are presented in table 4. Bootstrap support for nodes (r = 1000) were all 100. The branch length for *X. albilineans* was 0.232. **b)** Neighbor joining tree of concatenated nucleotide sequences for partial *dnaK*, *fyuA*, *gyrB*, and *rpoD* genes from *Xanthomonas* strains. A portion of the tree is presented, focusing on the *X. hortorum-cynarae-garnderi* group. Numbers indicate bootstrap support (r = 1000). The scale bars indicate the number of amino acid substitutions per site.

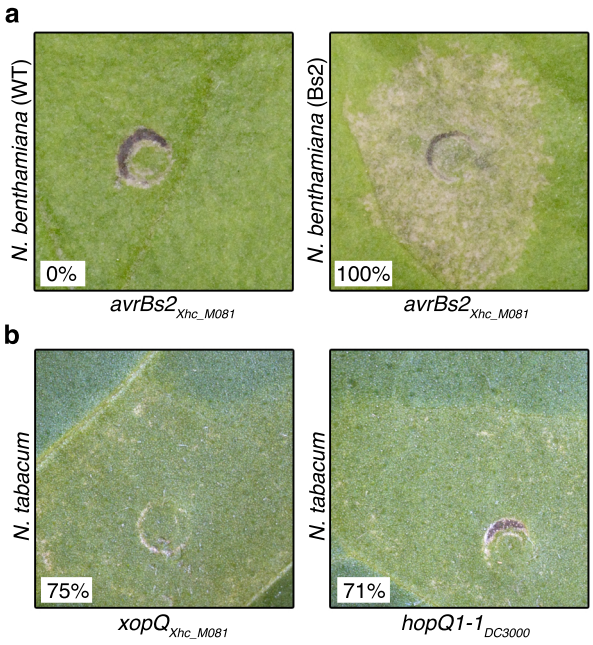



**Figure 2.4. AvrBs2$_{Xhc}$ and XopQ$_{Xhc}$ elicit a hypersensitive response in tobacco plants.**

Plants were challenged with $OD_{600}$ = 1.0 of *Agrobacterium tumefaciens* carrying *avrBs2*, *xopQ*, or *hopQ1-1* under the regulation of the CaMV35S promoter. **(a)** Wild-type and transgenic *N. benthamiana*, expressing *Bs2*, were challenged with *A. tumefaciens* carrying *avrBs2* (Tai et al., 1999). **(b)** Wild-type *N. tabacum* was challenged with *A. tumefaciens* carrying *xopQ* or *hopQ1-1*. Twenty-four leaf panels were infiltrated per experiment and experiments were repeated three times with identical results (percentage with a response are shown). Plant responses were scored 24 hours post infection.

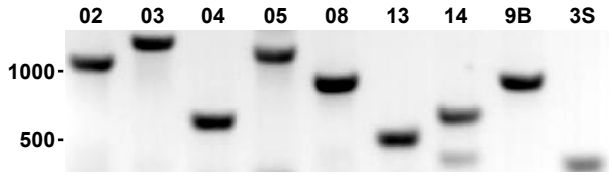



**Figure 2.5. A panel of molecular markers specific to *Xhc*.**

An inverse image of a 1X TAE agarose gel showing amplified products from genomic DNA extracted from *Xhc* M081. The primer pairs used are shown along the top. Their expected fragment lengths were 1041, 1266, 620, 1119, 875, 517, 651, 900, and 300 bp, respectively. Primer pairs 9B and 3S are from (Meng et al., 2004). The 500 and 1000 base markers from the 100 bp DNA ladder (NEB, Quick-Load) are shown.

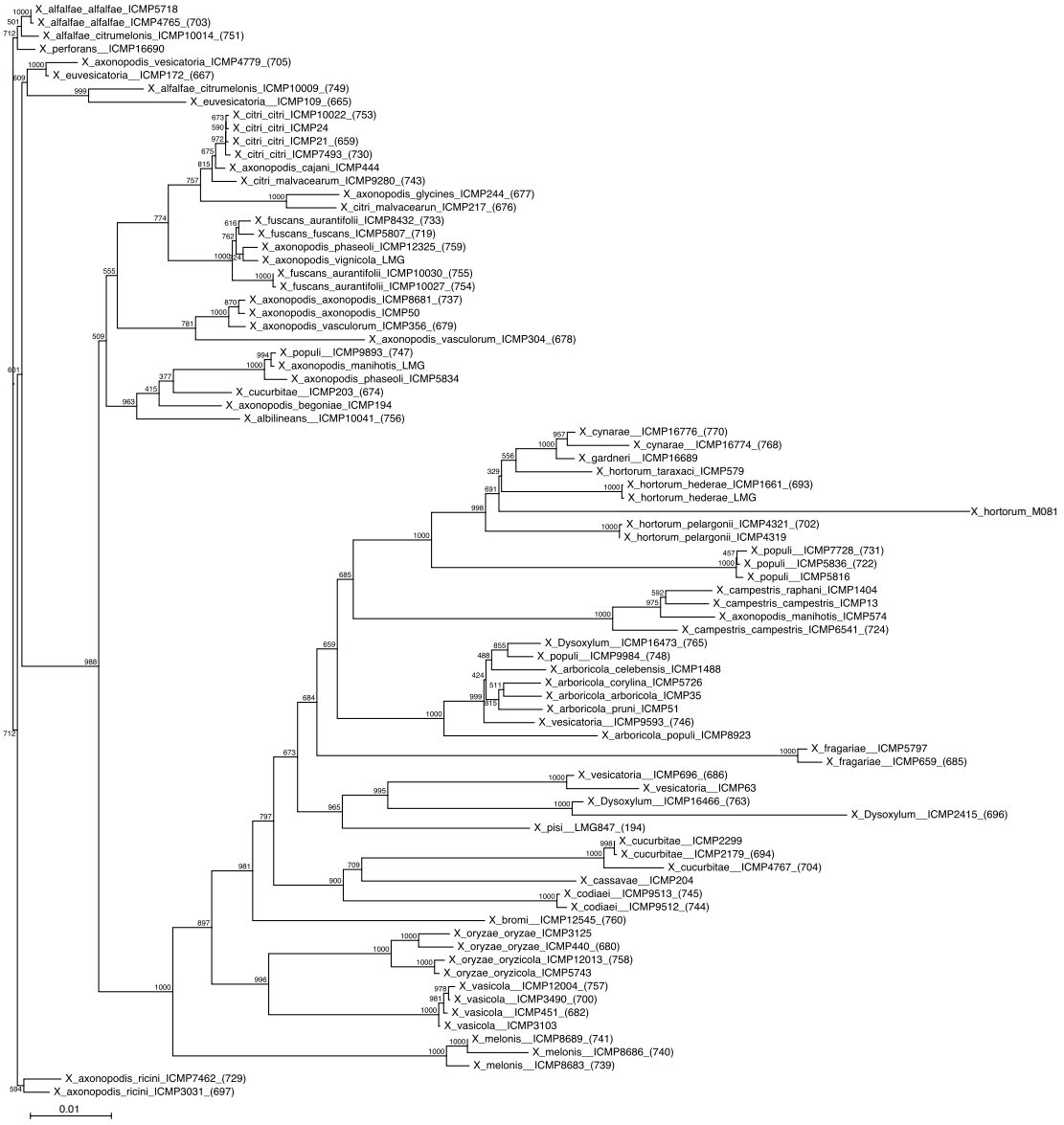



**Figure 2.6. Complete neighbor joining tree of concatenated nucleotide sequences for partial *dnaK, fyuA, gyrB,* and *rpoD* genes from *Xanthomonas* strains (Young et al., 2008).**

Numbers indicate bootstrap support (r = 1000). The scale bars indicate the number of amino acid substitutions per site.

**Table 2.1. Comparison of *Xanthomonas* genome characteristics**

| Isolate* | *Xhc*† | *Xcc* | *Xcv* | *Xac* | *Xoo* |
|---|---|---|---|---|---|
| Genome Size (Mb) | 5.062 | 5.076 | 5.178 | 5.176 | 5.240 |
| GC% | 63.7 | 65.1 | 64.7 | 64.8 | 63.6 |
| # CDSs | 4493 | 4179 | 4487 | 4312 | 4988 |
| Avg. length of CDSs (bp) | 985 | 1032 | 1005 | 1032 | 856 |
| Coding % | 87.4 | 84.9 | 87.4 | 86.2 | 81.8 |

*See table 2.4; †Based on improved, high-quality draft genome sequence

**Table 2.2. Candidate type III effectors of *Xhc* M081**

| Gene* | e-value† | Name/function | Distance from PIP-box (bp)‡ |
|---|---|---|---|
| 0064 | 0.0 | *avrBs2* | 64 |
| 0287 | $6 \times 10^{-62}$ | *xopR* | 63 |
| 0288¥ | $1 \times 10^{-21}$ | *xopR* | 76 |
| 0818 | 0.0 | *xopAG* | not found |
| 1217 | 0.0 | *xopQ* | 74 |
| 1402 | 0.0 | *xopF1* | 863 |
| 1403 | $7 \times 10^{-107}$ | *xopZ* | not found |
| 1405 | $2 \times 10^{-119}$ | *hrpW* ∫ | not found |
| 1411 | $1 \times 10^{-68}$ | *hpaA* | not found |
| 1431 | 0.0 | *xopX* | not found |
| 1432 | 0.0 | *xopX* | not found |
| 4256/4257§ | 0.0 | *xopAD* | not found |
| 4368 | $8 \times 10^{-22}$ | *xopAE*/*hpaF* | 41 |
| 4426 | $1 \times 10^{-21}$ | *avrXccA1* | not found |
| 4439 | $2 \times 10^{-72}$ | *xopT* | not found |
| 0140 | n/a | Hypothetical | 34 |
| 0803 | n/a | Hypothetical | 472 |
| 1218 | n/a | Hypothetical | 70 |
| 2239¥ | n/a | Hypothetical | 151 |
| 2558 | n/a | Hypothetical | 68 |
| 3774 | n/a | Hypothetical | 166 |
| 4437¥ | n/a | Hypothetical | 204 |

*CDS identifier number, †BLASTX; ‡Distance of predicted start codon from the 3' end of the PIP-box; ¥Unique to *Xhc* M081; ∫Helper protein secreted by T3SS; n/a = not applicable; §Potential pseudogene (see corresponding text).

**Table 2.3: Evaluation of oligonucleotide primers for specific detection of *Xhc***

| | Primer Pair (XhcPP)* | | | | | | | | Primer Pair* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Strain | 02 | 03 | 04 | 05 | 08 | 13 | 14 | | 9B[‡] | 3S[‡] | LAMP[§] |
| *Xhc* M081 | + | + | + | + | + | + | + | | + | + | + |
| *Xhc*PNW1 | + | + | + | + | + | + | + | | + | + | + |
| *Xhc*PNW2 | + | + | + | + | + | + | + | | + | + | + |
| *Xhc*Fr1 | + | + | + | + | + | + | + | | - | + | + |
| *Xhc*Ar1 | + | + | + | + | +[†] | - | + | | + | + | + |
| *Xhc*Ch1 | + | + | + | + | + | - | - | | + | + | + |
| *Xhc*Ch2 | + | + | + | + | + | + | + | | + | + | + |
| *Xhc*Ch3 | + | + | + | + | - | - | + | | + | + | + |
| *Xcc* ATCC 33913 | - | - | - | - | - | - | - | | - | - | - |
| *Xcv* MTV1 | - | - | - | - | - | - | - | | - | - | - |
| *Xccor* ID-A | - | - | - | - | - | - | - | | + | + | + |
| *Xhh ATCC 9653* | - | - | - | - | - | - | - | | - | - | + |

*+ = Primer pair yielded a product of expected size (see Fig. 4); - = Primer pair failed to yield a product; [†]Product was not of expected size (1.2 kb). [‡](Meng et al., 2004); [§](Temple and KB, 2009).

**Table 2.4. Strains and plasmids used in this study**

| Strain or plasmid | Relevant Information | Reference or source |
|---|---|---|
| Strain | | |
| *Escherichia coli* DH5α | $F^-$Φ80dlacZΔM15 *recA1 endA1 gryA96 thi-1 hsdR17* ($rK^-$ $mK^+$)*supE44 relA1 deoR*Δ(*lacZY-argF*)U169 | Gibco-BRL |
| *Xanthomonas hortorum* pv. *carotae* M081 | Wild-type | This study |
| *Xanthomonas hortorum* pv. *carotae* PNW1* | Isolated from carrot seed lot produced in USA | This study |
| *Xanthomonas hortorum* pv. *carotae* PNW2* | Isolated from carrot seed lot produced in USA | This study |
| *Xanthomonas hortorum* pv. *carotae* Fr1* | Isolated from carrot seed lot produced in France | This study |
| *Xanthomonas hortorum* pv. *carotae* Ar1* | Isolated from carrot seed lot produced in Argentina | This study |
| *Xanthomonas hortorum* pv. *carotae* Ch1* | Isolated from carrot seed lot produced in Chile | This study |
| *Xanthomonas hortorum* pv. *carotae* Ch2* | Isolated from carrot seed lot produced in Chile | This study |
| *Xanthomonas hortorum* pv. *carotae* Ch3* | Isolated from carrot seed lot produced in Chile | This study |
| *Xanthomonas hortorum* pv. *hederae* ATCC 9653 | Type strain | Vauterin et al., 1995 |
| *X. campestris* pv. *coriander* ID-A | Isolated from coriander seed lot produced in Oregon, USA | |
| *Xanthomonas campestris* pv. campestris ATCC 33913 | Type strain | |
| *Xanthomonas campestris* pv. *vesicatoria* MTV1 | Wild-type | This study |
| *Agrobacterium tumefaciens* GV2260 | Wild-type, $Rif^R$ | |
| *Pseudomonas syringae* pv. *tomato* DC3000 | Wild-type, $Rif^R$ | |
| Plasmids | | |
| pDONR207 | Gateway entry vector, $Gm^R$ | Invitrogen |
| pDONR207:*avrBs2* | CDS of *avrBs2* in entry vector | This study |
| pDONR207:*xopQ* | CDS of *xopQ* in entry vector | This study |
| pDONR207:*hopQ1-1* | CDS of *hopQ1-1* in entry vector | Chang and Dangl, unpublished |
| pGWB14 | Gateway destination vector, plant expression binary; CaMV35S, C-term 3XHA, $Kan^R$ | |
| pGWB14:*avrBs2* | Plant expression binary with CDS of *avrBs2* | This study |
| pGWB14:*xopQ* | Plant expression binary with CDS of *xopQ* | This study |
| pGWB14:*hopQ1-1* | Plant expression binary with CDS of *hopQ1-1* | This study |
| Used for genome comparisons | | |

| | | |
|---|---|---|
| *Xanthomonas alblineans* | NC_013722 | |
| *Xanthomonas campestris* pv. *campestris* str. ATCC33913 | NC_003902 | |
| *Xanthomonas campestris* pv. *campestris* str. 8004 | NC_007086 | |
| *Xanthomonas campestris* pv. *campestris* str. B100 | NC_010688 | |
| *Xanthomonas axonopodis* pv. *citri* str. 306 | NC_003919 | |
| *Xanthomonas campestris* pv. *vesicatoria* str. 85-10 | NC_007508 | |
| *Xanthomonas oryzae* pv. *oryzae* KACC10331 | NC_006834 | |
| *Xanthomonas oryzae* pv. *oryzae* MAFF 311018 | NC_007705 | |
| *Xanthomonas oryzae* pv. *oryzae* PXO99A | NC_010717 | |

*Isolates recovered from seed lots were typed as *Xhc* based on growth on XCS and YDC media and for positive amplification with *Xhc*-specific PCR and LAMP primers (data not shown).

**Table 2.5. Sequences of oligonucleotides used in this study**

| Primer Pair | Sequence of top strand primer (5' to 3') | Sequence of bottom strand primer (5' to 3') | Product Length (bp) |
|---|---|---|---|
| XhcPP01 | TTGCGGCCGGCAAAATGCAC | CTACGATCAGGCGGCGCAGG | 1323 |
| XhcPP02 | ACGCAGGCAGACACGACACG | GCGCTTTCGCTCAATGGCGG | 1041 |
| XhcPP03 | TGGGTGCCATAGCGTTGCGG | TGCGCCTCTGGTTGCACTCG | 1226 |
| XhcPP04 | CTCCACGCGCAGGTCCAGTG | GAGAAGCCTGGCTGACGCCG | 620 |
| XhcPP05 | ACAGGCCGAGTCGCAACAGC | TGCTGCCGCGAAACCCGATT | 1119 |
| XhcPP06 | AATGGATGTGGCCGCACGGG | GGTTGCGCTGGATGCGGTCT | 963 |
| XhcPP07 | AGATCGATGCGCTCGGCAGC | TTCCGACGCCGTCACCTTGC | 1405 |
| XhcPP08 | GCGCATCCATTTGCCAGCCG | CCCGCTCTTGCTCACCTGCC | 875 |
| XhcPP09 | GTTGCTTGGCGTGCGTGGTG | CGCGTGGTGGGAGCGTTCTT | 1038 |
| XhcPP10 | AGCTGTTGCCGGAACTCGCC | GCGCAGACCACGAAGTCGCT | 1207 |
| XhcPP11 | GCTGGGCTCGTCGGCGTATC | GGGAATGCCGCGTTGGTGGA | 1224 |
| XhcPP12 | TAGCTGTTGCTGCACGGCCC | TCGTTGCGCCCTCGTTGTCC | 1437 |
| XhcPP13 | AGCGGCAGCCGAGAACAACC | GCGCGCGCTACGAGATGAGT | 517 |
| XhcPP14 | ATCGGCCTGTGCAACGGTGG | ACGCGCTGCGCTGAAGAGTT | 651 |
| XhcPP15 | CATTGCGCGCATACACCGCC | CGTTGGCGCAGGTGGGGATT | 552 |
| XhcPP16 | TGTCGAACAGCCGCCCGAAC | CAGCAGTGCGGACACGCAGA | 979 |
| XhcPP17 | TCGGGCACTTGAAGGCGCAG | GTCGGTCGCGCCGTAGATGG | 523 |
| XhcPP18 | AACTCGCGCGTTCTTGCGGA | TGGCGCAACGGGGATTGGTC | 666 |
| *9B | CATTCCAAGAAGCAGCCA | TCGCTCTTAACACCGTCA | 900 |
| *3S | TGCCTGGCTACGGAATTA | ATCCACATCCGCAACCAT | 300 |
| XopQ | CAAAAAAGCAGGCTCCATGG ATTCCATCAGGCATCGCCCC | GAAAGCTGGGTGTTTTTCAGA AGCAAGCGCCAC | 1379 |
| AvrBs2 | CAAAAAAGCAGGCTCCATGC GTATTGGTCCTTTGCAACC | GAAAGCTGGGTGCTGCTCCGG CTCGATCTGTTTGGC | 2157 |
| XopX | AGCTTGGTCGCATGTTCACC | TCTGCGAAACAGAGCATTGG | 828 |
| XopR | CATTGACGGCAGCTGCTTGC | ATAACGATGCGATACAGCG | 666 |
| HrcC, hpa1, hpa2 | ATACCGATCAGACCGATCTG | GGCAATCCGCGATGTATCC | 600 |
| B1 and B2 | GGGGACAAGTTTGTACAAAAA AGCAGGCT | AGATTGGGGACCACTTTGTAC AAGAAAGCTGGGT | Gateway cloning |

*(Meng et al., 2004).

**Table 2.6. Testing XhcPP against other plant-associated bacteria**

| Bacteria tested[†] | XHCPP02[‡] | XHCPP03[‡] | XHCPP04[‡] | XHCPP05[‡] |
|---|---|---|---|---|
| *Xhc* M081 | 1041 | 1266 | 620 | 1119 |
| *Acinetobacter* sp. | N | N | N | N |
| *Agrobacterium radiobacter* | N | N | * | N |
| *Agrobacterium tumefaciens* | * | N | [¥] | * |
| *Bacillus cereus* | N | N | [¥] | N |
| *Bacillus licheniformis* | * | N | [¥] | * |
| *Bacillus mycoides* | * | N | N | N |
| *Bacillus* sp. | N | N | N | * |
| *Dickeya* spp. | * | N | N | * |
| *Enterobacter cloacae* | N | N | * | N |
| *Erwinia persicina* | * | N | [¥] | [¥] |
| *Frondihabitans* spp. | * | N | [¥] | N |
| *Massilia* spp. | * | N | [¥] | * |
| *Methylobacterium* spp. | * | N | * | N |
| *Pantoea agglomerans* | N | * | * | * |
| *Pectobacterium betavasculorum* | * | N | N | * |
| *Pectobacterium carotovorum* subsp. *atrosepticum* | N | N | N | N |
| *Pectobacterium carotovorum* subsp. *carotovorum* | * | * | * | N |
| *Pseudomonas putida* | N | N | N | N |
| *Pseudomonas reactans* | * | * | N | [¥] |
| *Pseudomonas syringae* | N | N | [¥] | N |
| *Pseudomonas tolaasii* | N | N | [¥] | * |
| *Ralstonia metallidurans* | * | * | [¥] | N |
| *Sphingomonas* spp. | N | N | N | * |

[†]Tested bacteria are from a culture collection maintained by V. O. Stockwell and J. E. Loper (USDA ARS Horticultural Laboratory in Corvallis, OR). Isolates were typed based on sequencing of the 16S-23S rRNA intergenic spacer region and other biochemical methods (V. O. Stockwell, personal communication). For PCR, cell suspensions of the bacteria were prepared at 5 x $10^5$ cfu/ml and boiled for 10 min. Five µl of boiled suspensions were used as templates for PCR. All reactions were done in duplicate with similar results. PCR conditions were as described in experimental procedures.

[‡]Sizes of expected fragments based on those amplified from *Xhc* M081 are shown; N = no amplified product; *single, faint, and non-specific band of ≤ 600 bp; [¥]two~four non-specific bands.

# An improved, high-quality draft genome sequence of the Germination-Arrest Factor-producing *Pseudomonas fluorescens* WH6

Jeffrey A. Kimbrel, Scott A. Givan, Anne B. Halgren, Allison L. Creason, Dallice I. Mills, Gary M. Banowetz, Donald J. Armstrong, and Jeff H. Chang

BMC Genomics
2010, vol. 511, p 522
PMID: 20920191

## ABSTRACT

**Background:** *Pseudomonas fluorescens* is a genetically and physiologically diverse species of bacteria present in many habitats and in association with plants. This species of bacteria produces a large array of secondary metabolites with potential as natural products. *P. fluorescens* isolate WH6 produces Germination-Arrest Factor (GAF), a predicted small peptide or amino acid analog with herbicidal activity that specifically inhibits germination of seeds of graminaceous species.

**Results:** We used a hybrid next-generation sequencing approach to develop a high-quality draft genome sequence for *P. fluorescens* WH6. We employed automated, manual, and experimental methods to further improve the draft genome sequence. From this assembly of 6.27 megabases, we predicted 5876 genes, of which 3115 were core to *P. fluorescens* and 1567 were unique to WH6. Comparative genomic studies of WH6 revealed high similarity in synteny and orthology of genes with *P. fluorescens* SBW25. A phylogenomic study also placed WH6 in the same lineage as SBW25. In a previous non-saturating mutagenesis screen we identified two genes necessary for GAF activity in WH6. Mapping of their flanking sequences revealed genes that encode a candidate anti-sigma factor and an aminotransferase. Finally, we discovered several candidate virulence and host-association mechanisms, one of which appears to be a complete type III secretion system.

**Conclusions:** The improved high-quality draft genome sequence of WH6 contributes towards resolving the *P. fluorescens* species, providing additional impetus for establishing two separate lineages in *P. fluorescens*. Despite the high levels of orthology and synteny to SBW25, WH6 still had a substantial number of unique genes and represents another source for the discovery of genes with implications in affecting plant growth and health. Two genes are demonstrably necessary for GAF and further characterization of their proteins is important for developing natural products as control measure against grassy weeds. Finally, WH6 is the first isolate of *P. fluorescens* reported to encode a complete T3SS. This gives us the opportunity to explore the role of what has traditionally been thought of as a virulence mechanism for non-pathogenic interactions with plants.

## BACKGROUND

*Pseudomonas fluorescens* is a diverse species of bacteria that is found throughout natural habitats and associated with plants. Contributing to their diverse lifestyles is their ability to produce an equally diverse array of secondary metabolites that affect interactions with hosts and other inhabitants of their ecosystems. Some isolates benefit plants by producing growth promoting hormones or antimicrobial compounds to control against pathogens (Haas and Défago, 2005). Others are deleterious and have the capacity to synthesize and secrete novel compounds that negatively affect growth of plants (Li et al., 2003; Flores-Vargas and O'Hara, 2006; Armstrong et al., 2009).

The physiological diversity of *P. fluorescens* is mirrored by its tremendous genetic diversity. However, the genetic diversity may reflect the possibility that *P. fluorescens* is not a single species, but rather a complex of at least two lineages. Molecular phylogenetic studies of 16 isolates suggested *P. fluorescens* should be represented by the *P. chlororaphis* and *P. fluorescens* lineages (Yamamoto et al., 2000). Alternatively or additionally, *P. fluorescens* may have an open pan genome (Medini et al., 2005; Tettelin et al., 2008). Finished genome sequences are available for the SBW25, Pf-5, and Pf0-1 isolates of *P. fluorescens* (Paulsen et al., 2005; Silby et al., 2009). Their genomes exceed 6.4 megabases and their relatively large sizes are not unexpected for free-living bacteria (Merhej et al., 2009). Comparative analyses of the three isolates of *P. fluorescens* revealed substantial variation in diversity of genome content and heterogeneity in genome organization (Silby et al., 2009). Each genome has 1,000 to nearly 1,500 unique genes when compared to each other.

Plant-associated isolates of *P. fluorescens* potentially have mechanisms for interacting with plants. Many Gram-negative bacteria use a type III secretion system (T3SS) to interact with their hosts (Galán and Wolf-Watz, 2006). The T3SS is the most complex of the bacterial secretion systems and is typically encoded by a large cluster of genes arranged as a single superoperon. Its function is to inject type III effector proteins directly into host cells (Galán and Wolf-Watz, 2006; Grant et al., 2006). These type III effectors are important host-range determinants of plant pathogenic bacteria because they perturb and potentially elicit plant defenses (Jones and Dangl, 2006).

It is unclear as to how prevalent T3SS-encoding regions are in *P. fluorescens*. Nearly 60% of a surveyed collection of *P. fluorescens* strains had a homolog of *rscN*, which encodes the ATPase of the T3SS (Rezzonico et al., 2004). However, it is not known whether all genes necessary to complete the T3SS are present in these isolates. Of the three completed genomes, genes encoding the T3SS are present only in SBW25. Several important or typically conserved genes are missing or truncated in the T3SS-encoding locus of SBW25 (Preston et al., 2001). Despite the cryptic appearance of the T3SS, when constitutively expressing the transcriptional regulator of the T3SS, SBW25 could deliver a heterologous type III effector into plant cells, suggesting the T3SS may still be functional (Preston et al., 2001).

The role of the T3SS for the lifestyle of *P. fluorescens* is still unclear. In SBW25, despite the cryptic T3SS, single mutants of some but not all the remaining T3SS-encoding genes were reduced in fitness in the rhizosphere of sugar beets (Jackson et al., 2005). This is not unheard of, as mutants of seemingly cryptic T3SS in pathogens are compromised in virulence (Ideses et al., 2005). However, in the case of SBW25, the T3SS mutants were also compromised in growth *in vitro* (Jackson et al., 2005). A T3SS mutant of the biocontrol isolate *P. fluorescens* KD was compromised in its ability to protect cucumbers against damping-off disease caused by *Pythium ultimum* (Rezzonico et al., 2005). This may be a result of KD requiring a functional T3SS to elicit host defenses, thereby indirectly protecting against *P. ultimum* or potentially as a direct mechanism against the pathogen.

We are interested in exploiting *P. fluorescens* for control of grassy weeds. We have previously reported the selection, isolation, and characterization of five strains of *P. fluorescens* that inhibit germination of seeds of grassy weeds (Banowetz et al., 2008). Further characterizations led to the identification of Germination-Arrest Factor (GAF) produced by these isolates. GAF is a small, extremely hydrophilic secreted herbicide that reacts with ninhydrin and possesses an acid group, suggestive of a small peptide or amino acid analog (Armstrong et al., 2009; Banowetz et al., 2009). In particular, the high specificity of GAF towards grasses and inhibitory activity at only certain developmental stages during seed germination provides promise for its potential as a natural herbicide for the control of grassy weeds in grass seed production and turf management settings.

We selected *P. fluorescens* WH6 as the archetypal GAF-producing isolate. WH6 was extracted from the rhizosphere of *Poa* sp. and *Triticum aestivum* at the Hyslop Research Farm in Benton County, Oregon, USA (Banowetz et al., 2008). We sequenced and developed an improved high-quality draft sequence for WH6 using a hybrid Illumina and 454-based sequencing approach. This standard is considered sufficient for our purposes of assessing gene inventory and comparing genome organization (Chain et al., 2009).

Comparative genomic analysis showed a high number of orthologous genes and strong similarity in genome organization between WH6 and SBW25. Phylogenomic analysis supported this observation and placed WH6 in the same lineage as SBW25, or the proposed *P. fluorescens* lineage. The high similarity in orthology and genome organization is in contrast to previous observations of *P.*

*fluorescens* and in comparisons of WH6 to Pf-5 or Pf0-1 (Silby et al., 2009). From a non-saturating Tn*5*-mutagenesis screen of WH6, we previously identified two mutants compromised in GAF activity (WH6-2::Tn*5* and WH6-3::Tn*5*; (Armstrong et al., 2009)). Mapping of DNA sequences flanking the two mutants revealed genes encoding proteins with potential functions in regulation and biosynthesis of GAF. Finally, inspection of the WH6 genome revealed several candidate host-association mechanisms, including what appears to be a complete type III and two complete type VI secretion systems.

## RESULTS AND DISCUSSION:

### Sequencing and developing an improved, high-quality draft genome sequence

We used an Illumina and a 454 FLX GS LR70 to sequence the genome of WH6 (Table 3.1). The theoretical coverage using all filtered reads was estimated to be 316x assuming a genome size of approximately 6.5 megabases. We employed a number of steps to meet the standards of an improved, high-quality draft genome sequence of WH6 for comparative purposes. We used Velvet 0.7.55, combinations of short-reads, and a variety of parameter settings to *de novo* assemble the short reads to generate approximately 75 different assemblies (Zerbino and Birney, 2008). We developed and used *ad hoc* Perl scripts with an associated visualization tool to compare each of the different assemblies to each other. This step enabled us to eliminate entire assemblies with large contigs not supported by any other assembly.

We identified a single high-quality *de novo* assembly based on nearly 24 million reads from all three sequencing methods (Table 3.1). The Velvet parameters were hash length of 31, expected coverage of 104, and a coverage cutoff of 20. Actual coverage of this assembly based on the total number of used reads was 65 ~ 120x, which was less than one-third the theoretical coverage. This assembly had 189 contigs greater than one kb and a total of 256 contigs greater than 100 bp. The largest contig was 264 kb and the N50 number and size were 26 contigs and 78 kb, respectively.

We used experimental and *in silico* approaches to improve the draft assembly by reducing the number of physical gaps. Of the 189 contigs greater than one kb, 139 contigs (74%) had significant homology to a reference sequence shared by the end of another contig. These 139 contigs potentially flanked 111 physical gaps (See Additional file 1: Table S1; http://www.biomedcentral.com/1471-2164/11/522/additional). We were able to amplify across 86 (77.5%) of the gaps using PCR. Physical gaps were subsequently resolved by reassembling the nearly 24 million short-reads with the 86 Sanger reads. Of the remaining scaffolds, we associated more based on *in silico* evidence. Some contigs shared long-range synteny to a reference genome (see below) and their ends had fifteen or more basepairs of sequence with 100% overlap to each other. This phenomenon is a result of Velvet failing to extend the contig because of low coverage. Secondly, some contigs could be paired together because their ends had partial coding regions with homology to a common reference gene. In total, nineteen more contigs were associated, resulting in a

final assembly of 115 scaffolds greater than 100 bp. The largest scaffold was 814 kb and the N50 number and size were 8 and 203 kb, respectively.

The improved, high-quality draft genome sequence had 67 sequence gaps totaling 258,650 Ns. There were 45 large sequence gaps with more than 300 Ns of which eight had more than 10,000 Ns each. We presumed these were artifacts of the Velvet assembly because the fragment size of our paired-end library was no larger than 300 bp. We corrected the sizes for 31 gaps to their corresponding length found in homologous reference sequences. In the other 14 cases, we simply reduced the number of Ns in the region to 300 bp, to reflect the maximum size of our paired-end library. Both approaches to correct the size of sequence gaps were validated using PCR of randomly selected regions (data not shown). In total, we reduced the number of Ns to 6,049 or ~2% of the original number of Ns.

The release of the finished genome sequence of SBW25 fortuitously coincided with our efforts of improving the draft genome sequence of WH6 (Silby et al., 2009). We noted that nearly 90% of the homologous sequences we found in the NCBI nt dataset using our BLASTN-based approaches were to *P. fluorescens* SBW25. We therefore surmised that the genome of WH6 would be similar to the finished genome of *P. fluorescens* SBW25 and used it as a reference for Mauve Aligner to reorder the 115 WH6 scaffolds (Rissman et al., 2009; Silby et al., 2009).

The genome of WH6 is presumed to be a single circular chromosome (Figure 3.1). A total of 53 scaffolds greater than one kb could be ordered using Mauve Aligner. The remaining 62 contigs could not be reordered and were

excluded from our circular representation of the genome. These 62 contigs were all smaller than one kb and their sum total was only 13 kb. Attempts to use Pf0-1 or Pf-5 as a reference for Mauve Aligner were largely unsuccessful, supporting our observation that WH6 and SBW25 had higher synteny than previously detected in *P. fluorescens* and suggesting our WH6 *de novo* assembly was of high quality. We found no evidence of plasmids in the genome of WH6.

This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AEAZ00000000. The version described in this paper is the first version, AEAZ01000000. The WH6 genome sequence and its associated tools can also be accessed from our website at: http://changbugs.cgrb.oregonstate.edu/microbes/org_detail.html?org=WH6-G3.

One challenge with *de novo* assembly is dealing with repeated sequences (Pop and Salzberg, 2008). Small repeated sequences are present in genomes of *P. fluorescens* but were not expected to have a large effect on our ability to assemble the WH6 genome because of the size of our paired-end fragments (Silby et al., 2009). Larger repeats, however, could not be resolved. We only observed one rRNA operon in the genome of WH6. We suspect that WH6 has five rRNA operons similar to SBW25 and Pf-5, but they collapsed into one contig. There was approximately 5X more coverage for the contig containing the one rRNA operon of WH6 compared to the other contigs. Similarly, nonribosomal peptide synthases (NRPSs) are encoded by large genes with repeated modules (Gross and Loper, 2009; Wilkinson and Micklefield, 2009). The modular domains either collapsed on one another in the assembly, or were assembled into short

contigs that we could not extend. A large fraction of these partial NRPS-encoding genes were found in the small contigs that we could not reorder using Mauve Aligner. Here too, we noticed higher coverage than the other scaffolds.

**Comparative and phylogenomic analyses of *P. fluorescens***

At a large scale, the genome of WH6 was similar to the genomes of the other *P. fluorescens* isolates (Table 3.2). The size of the genome is slightly smaller, which may be a consequence of the draft nature of our genome assembly. Nonetheless, the 5876 predicted coding sequences (CDSs) and 89.2% coding capacity were very similar.

Previous analyses of *P. fluorescens* found SBW25, Pf-5, and Pf0-1 to be divergent, with only ~61% of the genes shared and little long-range synteny (Silby et al., 2009). We used HAL to carry out similar analyses to determine the effect of the WH6 genome on the phylogenetic relationship of the *P. fluorescens* species and potential changes to the size of its pan genome (Robbertse et al., 2006). HAL uses a Markov Clustering algorithm based on e-values from reciprocal all-by-all BLASTP analysis to create clusters of orthologs. Core sequences from each species are concatenated and the super alignment is used in phylogenomic analysis. Using a core of 1966 translated sequences common to *P. fluorescens*, representative strains of *P. syringae*, and *P. aeruginosa* PAO1, HAL clustered the different species of *Pseudomonas* as expected (Stover et al., 2000; Buell et al., 2003; Feil et al., 2005; Paulsen et al., 2005; Silby et al., 2009). Further, HAL clearly defined two separate lineages for *P. fluorescens*, placing WH6 with SBW25 (Figure 3.2).

Within the *P. fluorescens* species as presently defined, 3115 genes formed the core and represented 53%, 52.6%, 50.7%, and 54.3% of the genomes of WH6, SBW25, Pf-5, and Pf0-1, respectively (Figure 3.3). This was nearly a 10% reduction relative to previous analysis of three genomes (Silby et al., 2009). A large fraction of the core genes was assigned to categories with general cellular processes such as energy production and conversion, amino acid transport and metabolism, translation, and transcription (Figure 3.4). Approximately 90% of the 3115 core genes had identical COG designations suggesting our automated annotation pipeline was accurate. There were some exceptions but their rarity and subtle differences did not warrant manual curation. For example, one cluster of orthologs had genes annotated as “arabinose efflux permeases” (COG2814) for genes from the published isolates of *P. fluorescens* but “permease of the major facilitator superfamily” (COG0477) for the ortholog of WH6.

A total of 4309 of the translated products of WH6 had an orthologous sequence in another isolate of *P. fluorescens*. Almost 69% of the WH6 genes had an orthologous sequence in SBW25, as compared to Pf-5 and Pf0-1 with 62% and 59%, respectively (Figures 3.1 and 3.3). We found similar levels of overlap using reciprocal BLASTP (data not shown). The 69% orthology between WH6 and SBW25 is much higher than previously observed between isolates of *P. fluorescens* (Silby et al., 2009). These levels were still lower than those between different pathovars of *P. syringae*, which had greater than 80% orthology (Feil et al., 2005; Joardar et al., 2005; Studholme et al., 2009). Therefore, the generalization that *P. fluorescens* have highly variable genomes still holds true.

The genomes of WH6 and SBW25 also showed extensive long-range synteny (Figure 3.5). This amount of synteny was unexpected given previous comparisons (Silby et al., 2009). When compared to Pf-5 or Pf0-1, we found little long-range synteny, which tended to be near the origin of replication. Synteny rapidly degraded away from the origin with an increase in inversions between the genomes (Silby et al., 2009). Taken together these lines of evidence all suggest WH6 and SBW25 to be similar and support, though perhaps prematurely, a redefinition of the *P. fluorescens* species (Yamamoto et al., 2000; Silby et al., 2009).

It could be argued that the high level of synteny we found with SBW25 was an artifact of using SBW25 to reorder the WH6 scaffolds. Though we cannot exclude this possibility, we highlight several points that suggest otherwise. We used a *de novo* approach to assemble the genome of WH6. The long-range synteny to the SBW25 genome was observed within each and across the *de novo* assembled scaffolds of WH6 (Figure 3.5). Furthermore, synteny with SBW25 was also supported by our ability to use SBW25 to successfully and substantially reduce the number of WH6 scaffolds and improve the WH6 genome sequence (Figure 3.1). Finally, analysis of GC skew gave higher confidence in the reordering of WH6 scaffolds (Figure 3.1, track 8). Genomes often have a bias of guanine in the leading strand (Rocha, 2004; Arakawa and Tomita, 2007). Inversions of GC skew in regions distant from the replication origins and termini are indicative of a recent recombination event (Parkhill et al., 2003). Barring these events, inversions of GC skew could also potentially indicate large-scale

misassemblies or incorrect reordering of contigs. For the most part, the genome of WH6 showed the expected bias of guanine in the leading strand; there are perhaps two small inversions in GC skew flanked by physical gaps between scaffolds near the terminator. Our use of SBW25 as a reference for reordering scaffolds is therefore acceptable and the observed synteny between WH6 and SBW25 appeared to reflect true similarities in genome organization.

More than 30% of the WH6 coding regions were unique (Figures 3.1 and 3.3). Examinations of their annotated functions suggested greater diversity in metabolic and host-association functions such as carbohydrate transport and metabolism, inorganic ion transport and metabolism, secondary metabolite biosynthesis, transport and catabolism, intracellular trafficking, secretion and vesicular transport, as well as defense mechanisms (Figure 3.4).

Examples of CDSs unique to WH6 and enriched in these functional categories include 35 candidate permeases of the major facilitator superfamily, a large and diverse superfamily of secondary active transporters that control movement of substrates across membranes (COG0477; (Law et al.)). WH6 also had 12 unique CDSs that encode for putative TonB-dependent receptors, involved in uptake of iron and potentially other substrates (COG1629; (Blanvillain et al., 2007); see also section entitled, “Regulators of gene expression”). Restriction modification (RM) systems are widespread defense mechanisms that protect prokaryotes from attack by foreign DNA (Tock and Dryden, 2005). RM systems are diverse and can vary dramatically in numbers. WH6 has at least 30 CDSs with annotated functions or domains common to proteins of RM systems.

PFWH6_5037-5039, for example, encode for a type I RM system that appears to be unique to WH6. Finally, other CDSs unique to WH6 and of direct interest to us are described in the following sections. The greater than 1500 genes unique to WH6 were dispersed throughout its genome with only a slight bias in location closer to the terminators (Figure 3.1). This bias was previously generalized for *P. fluorescens* (Silby et al., 2009).

**Mapping GAF mutants**

We previously identified two WH6 mutants from a non-saturating Tn*5*-mutagenesis screen for those affected in arresting the germination of *Poa* seeds (Armstrong et al., 2009). We cloned, sequenced and mapped their flanking sequences to identify the disrupted genes. Mutant WH6-3::Tn*5* had an insertion in PFWH6_3687. This CDS is annotated as a "predicted transmembrane transcriptional regulator (anti-$\sigma$ factor)". Its closest homolog, with 94% similarity is PrtR encoded by *P. fluorescens* LS107d2 (Burger et al., 2000). The Tn*5* element had inserted at nucleotide position 417 within codon Asp139. Because loss of *prtR* led to a loss of GAF activity, PrtR is likely an activator rather than a repressor, as was the case in *P. fluorescens* LS107d2 (Burger et al., 2000).

Just upstream of *prtR* in WH6 is *prtI*, which encodes a candidate ECF $\sigma^{70}$ factor. This arrangement is reminiscent of many sigma-anti-sigma factor pairs and suggests that the genes are potentially co-regulated and both may have roles in regulating GAF gene expression (Hughes and Mathee, 1998). It is peculiar that we failed to identify an insertion in *prtI* but one obvious explanation is that our

screen was not saturating. Regardless, it will be important to examine the necessity of PrtI for GAF activity to resolve its role.

Mutant WH6-2::Tn*5* had an insertion in PFWH6_5256, a gene encoding a candidate aminotransferase class III. The identification of an aminotransferase as necessary for GAF supports our previous findings suggesting that GAF contains an amino group and may be a small peptide or amino acid analog (Armstrong et al., 2009). Aminotransferases are pyridoxal phosphate (PLP)-dependent enzymes that catalyze the transfer of an amino group from a donor group (commonly an amino acid) to an acceptor molecule (Yoshimura et al., 1996). The Tn*5* element had inserted at nucleotide position 1124 within codon Lys375. Based on comparisons to the acetyl ornithine aminotransferase family the insertion is distal to the conserved residues that compose pyridoxal 5'-phosphate binding sites, the conserved residues that compose inhibitor-cofactor binding pockets, and the catalytic residue (Marchler-Bauer et al., 2009). Further characterization of WH6-2::Tn*5* is necessary to examine its enzymatic properties and role in biosynthesis of GAF.

**Regulators of gene expression**

Bacteria with large genomes tend to have complex regulatory networks to integrate and respond to a multitude of environmental signals. The extracytoplasmic function (ECF) $\sigma^{70}$ factors are a class of important transcriptional regulators of cell-surface signaling systems. Using a Hidden Markov Model (HMM) for ECF-encoding genes, we found 19, 26, 28, and 22 ECFs in WH6, SBW25, Pf-5 and Pf0-1, respectively (Staroń et al., 2009). Of the 19 identified in

WH6, ten are part of the core set common to all four sequenced *P. fluorescens* isolates and included *prtI* and *prtR*, which we identified as necessary for GAF activity. Because we had previously shown that Pf-5 and Pf0-1 do not have GAF activity, these results suggest that the putative PrtI/R-regulon may be different between the different isolates of *P. fluorescens* (Armstrong et al., 2009). Four of the 19 ECF-encoding genes were exclusive to the plant-associated strains WH6, SBW25, or Pf-5. Two of these were only shared with SBW25, of which one was *rspL* (see below). The other two lacked sufficient annotations to speculate on their functions. The remaining five ECFs were unique to WH6 and all are potentially co-expressed with genes encoding outer membrane receptors involved in iron perception or uptake (*chuA, fhuA,* and *fhuE*).

**Virulence factors**

Pseudomonads produce a wide-range of secondary metabolites with potential benefit or detriment to plants and microbes (Lindeberg et al., 2008; Gross and Loper, 2009). Many are synthesized by non-ribosomal peptide synthases (NRPS) or polyketide synthases (Lindeberg et al., 2008; Gross and Loper, 2009; Wilkinson and Micklefield, 2009). We found evidence for several NRPS-encoding genes. Because of their modular architecture, most NRPS-encoding genes of WH6 were fragmented and found on small contigs that failed to assemble or reorder. Therefore, it was not possible to determine the structure of the repeats or infer functions based on homology. We were, however, able to identify several other candidate toxins and virulence factors (Table 3.3).

We identified several secretion systems in WH6 unique to host-associated bacteria and/or necessary for full virulence of pathogenic bacteria. WH6 appears to encode a complete and functional type III secretion system (PFWH6_0718-0737; Figure 3.6a). We named its genes according to the nomenclature first proposed for SBW25 (Bogdanove et al., 1996; Preston et al., 2001). There is strong homology and synteny between the T3SS-encoding regions of WH6 and *P. syringae*, raising the possibility of a recent acquisition of the T3SS-encoding locus by WH6, similar to KD (Rezzonico et al., 2004). Phylogenetic analyses of *rscN*, however, placed WH6 with the group 8 of biocontrol isolates of *P. fluorescens* (data not shown; (Rezzonico et al., 2004)). Additionally, 15 kb of sequences on either side of the T3SS-encoding region of WH6 were highly syntenic to regions flanking the T3SS-encoding region of SBW25 with the exception of the type III effector gene, *ropE*. Together, these data argue against a recent acquisition of the T3SS-encoding region by WH6.

There were some differences between T3SS-encoding regions of WH6 and *P. syringae*. The *rspR, rspZ,* and *rspV* genes of WH6 were not present and we failed to detect any homology between the *rspF/hrpF, rspA/hrpA*, and *rspG/hrpG* genes. Data, however suggest these differences likely have little to no effect on T3SS function. HrpR and HrpS are highly similar and are functionally redundant. In some *Erwinia* strains, HrpS by itself is demonstrably sufficient for T3SS function (Wei et al., 2000; Hutcheson et al., 2001). Deletion mutants of *hrpZ* are still functional and HrpV functions as a negative regulator of the T3SS (Gail Preston, 1998; Alfano et al., 2003; Ortiz-Martín et al., 2010). HrpF and HrpA are

homologous to each other and are structural components of the T3SS. They are the most polymorphic proteins encoded within the T3SS-cluster and the absence of significant homology between *rspF* and *rspA* to their counterparts of *P. syringae* was therefore not surprising (Deng et al., 1998; Ramos et al., 2007). Our automated annotation approach failed to identify *rspG* but upon visual inspection, we noted a small CDS that encodes a potential product of 63 amino acids. BLAST searches failed to detect homology to *hrpG*, but given its position in the T3SS-encoding region and similarity in size to the translated product of *hrpG*, we have annotated it as *rspG*. In total, these data support the notion that WH6 encodes a complete and functional T3SS, although, its role in the lifestyle of WH6 remains unknown.

**Candidate type III effectors of WH6**

We used a homology-based approach to search for type III effector genes in the genome of WH6. Our database of type III effectors included those from T3SS-using phytopathogens and some mammal pathogens. We only identified one translated sequence with homology to PipB from *Salmonella*, and another with homology to HopI1 from *P. syringae* (e-value < $1x10^{-7}$, > 33% identity; (Knodler et al., 2004)). However, neither appeared to be strong candidates for a type III effector. We identified a homolog of *pipB* in the genome of Pf-5, which does not encode the T3SS. HopI1 encodes a J domain and its homolog in WH6 was annotated as the molecular chaperone, *dnaJ* (Jelenska et al., 2007). These results suggest that if WH6 does encode type III effectors, they are very divergent in sequence. SBW25, in contrast, had at least five genes with homology to known

type III effectors, of which two were expressed to sufficiently high levels and delivered by a heterologous T3SS-encoding bacterium (Vinatzer et al., 2005).

Computational approaches have been successfully used to identify candidate type III effectors from *P. syringae*, based in part on identifying a cis-regulatory element upstream of their genes and also some genes of the T3SS (Fouts et al., 2002; Chang et al., 2005; Ferreira et al., 2006). This so-called *hrp*-box is recognized by HrpL, an extracytoplasmic function (ECF) $\sigma^{70}$ factor encoded within the T3SS-encoding region of *P. syringae* (Innes et al., 1993). We therefore used a Hidden Markov Model (HMM) trained using 38 known HrpL-regulated genes of *P. syringae* pv *tomato* DC3000 to mine the genome of WH6 for *hrp*-boxes (Chang et al., 2005; Schechter et al., 2006).

We found 115 *hrp*-boxes in the genome of WH6 (bit score ≥ 3.0) but only 24 were within 500 bp of a CDS. Two were located upstream of *rspF* and *rscR* in the T3SS-encoding region, with bit scores of 7.9 and 3.2, respectively. We also identified a *hrp*-box upstream of *rspJ* but it had a lower bit score of 1.2. Fifteen of the CDSs downstream of candidate *hrp*-boxes had annotated functions not typically associated with type III effectors and we did not list them as possible candidates (data not shown). The remaining eight CDSs downstream of *hrp*-boxes were annotated as hypothetical proteins and the five with the highest bit scores for their corresponding *hrp*-boxes were not present in the genomes of Pf-5 and Pf0-1; all but PFWH6_1942 were unique to WH6 (Table 3.4). Further investigation of their first amino-terminal residues indicated that three have

characteristics suggestive of T3SS-dependent secretion (Guttman et al., 2002; Petnicki-Ocwieja et al., 2002; Chang et al., 2005).

Our two computational approaches yielded very few candidate type III effectors. One possible explanation is that because RspL and HrpL have only 50% identity (70% similarity), they recognize slightly different cis-regulatory sequences and our HMM was not adequately trained for the cis-regulatory sequence recognized by RspL. This is an unlikely scenario. Three sequences with strong similarity to the *hrp*-box of *P. syringae* were found in the T3SS-encoding regions for WH6 and SBW25 (Preston et al., 2001). Additionally, it has been observed that all HrpL-dependent phytopathogenic bacteria share an identical motif in the *hrp*-box despite having as little as 52% similarity (Nissan et al., 2005). Furthermore, in $\sigma^{70}$ factors, DNA binding specificity is conferred by the helix-turn-helix domain 4.2 (C B Harley, 1987; Potvin et al., 2008). Domain 4.2 of the WH6 RspL is highly similar (82.5%) to the corresponding domain of HrpL. An alternative explanation is that WH6 encodes very few type III effectors with little homology to those that have been identified. This is not unheard of. *P. aeruginosa* for example, has only three type III effectors (Feltman et al., 2001; Wolfgang et al., 2003).

**Type VI secretion systems**

The type VI secretion system (T6SS) is another secretion apparatus that is common to host-associated bacteria. Computational approaches suggest the T6SS may also be in *P. fluorescens* (Bingle et al., 2008; Shrivastava and Mande, 2008). We found evidence for two complete and functional T6SSs in WH6. We

have named these two systems T6SS-1 (Figure 3.6b; PFWH6_5796-5812) and T6SS-2 (Figure 3.6c; PFWH6_3251-3270). It is not uncommon for organisms to possess multiple T6SSs that are of different lineages and acquired independently (Bingle et al., 2008). Additionally, in other strains that have been characterized, different T6SSs appear to be independently regulated, suggesting each T6SS may have functions specific to different aspects of the lifestyle of the bacterium (Mougous et al., 2006). Whether this is also the case with WH6 awaits further characterization.

T6SS-1 belongs to the group A lineage and shares homology and synteny to HSI-I of *P. aeruginosa* PAO1 (Mougous et al., 2006). We therefore named the corresponding genes in WH6 according to the nomenclature established for HSI-I (Figure 3.6b). Synteny extended beyond the T6SS-encoding region and included the *tagQRST* genes bordering *ppkA*. We did not, however, find evidence for *tagJ1* in WH6 (Bingle et al., 2008; Hsu et al., 2009). T6SS-2 is a group B secretion system (Bingle et al., 2008). Less is understood about the group B secretion systems but T6SS-2 showed strong homology and synteny to a corresponding T6SS encoded in the genome of the phytopathogen *P. syringae* pv *tomato* DC3000 (Figure 3.6c; (Buell et al., 2003)).

There are few proteins that are demonstrable type VI effectors. Three homologs of VgrG and Hcp have been shown to be secreted by the T6SS but both likely have functions for the T6SS itself (Pukatzki et al., 2006; 2007; Wu et al., 2008). We found four *vgrG* genes, of which only one was associated with T6SS-1. The other three genes were found elsewhere in the genome. Whether

products from these latter three are secreted proteins of the T6SS or merely homologous in sequence is unknown. Both T6SSs of WH6 had a homolog of *hcp*. Recently, three additional proteins from *P. aeruginosa* PAO1 were shown to be secreted by the T6SS, but their orthologs were not found in WH6 (Hood et al., 2010).

## CONCLUSIONS

*P. fluorescens* is a genetically and physiologically diverse species found in many habitats. We sequenced the genome of the isolate WH6 because it produces Germination-Arrest Factor (GAF), an herbicide that specifically arrests seed germination of graminaceous species. Comparisons of the WH6 genome to genomes of SBW25, Pf-5, and Pf0-1 helped better define this species, with WH6 and SBW25 forming one lineage. Comparative studies revealing substantial similarity in gene inventory and synteny supported its placement and the argument of at least two major lineages of *P. fluorescens* (Yamamoto et al., 2000).

With the genome sequence, we were able to deduce potential functions for two genes necessary for GAF activity. One encoded a candidate anti-sigma factor. Our previous results suggest that PrtR is an activator and suggests it has a role in regulating expression of genes necessary for GAF. The second gene encoded a candidate aminotransferase, which tentatively supports our previous speculation that GAF is a small peptide or amino acid analog. Further studies are required to confirm their functions. A less labor-intensive and saturating screen

will also be necessary for a fuller understanding of the pathway controlling GAF expression and biosynthesis. The genome sequence will certainly facilitate such future endeavors.

We also identified a number of mechanisms that potentially affect plant health and some typically associated with host-associated bacteria. One of the more extensively characterized mechanisms is the type III secretion system. WH6 appears to encode the necessary repertoire of genes for a complete and functional T3SS. We also identified two T6SSs in WH6. Further studies are necessary to identify the role these secretion systems and their effectors play in the lifestyle of WH6.

## METHODS

### Sequencing DNA flanking Tn5-insertions

To determine the sites of Tn*5* insertion, genomic DNA from the two GAF mutants, WH6-2::Tn*5* and WH6-3::Tn*5* was digested with *Bam*HI or *Pst*I, respectively. We used Southern blotting with a biotinylated probe of the $Tet^R$ gene from pUTmini-Tn*5*gfp to identify the fusion fragments between the $Tet^R$ gene and flanking WH6 DNA (Matthysse et al., 1996). DNA fragments of corresponding size were cloned into pBluescript SK+ (Stratagene, La Jolla, CA), transformed into *E. coli* DH5∝, selected based on tetracycline resistance, isolated, and sequenced outwards using primers to the $Tet^R$ gene.

### *P. fluorescens* WH6 Genome Sequencing

We used the ZR Fungal/Bacterial DNA Kit to isolate genomic DNA from *P. fluorescens* WH6 grown overnight in LB at 28 °C (Zymo Research, Orange, CA).

Purity and concentration were determined using a Nanodrop ND1000 (Thermo Scientific, Waltham, MA). For Illumina-based sequencing, we prepared the DNA according to the instructions of the manufacturer and sequenced the DNA fragments on the Illumina GA I and II using 36-cycle (4 channels) and paired-end 76-cycle (1 channel) sequencing, respectively (Illumina, San Diego, CA). The Sanger and Illumina sequencing was done at the Center for Genome Research and Biocomputing Core Labs (CGRB; Oregon State University, Corvallis, OR). We also sequenced genomic DNA using the 454 FLX GS LR70 (454, Branford, CT). Preparation and sequencing by 454 was done at the Consortium for Comparative Genomics (University of Colorado Health Sciences Center, Denver, CO).

**Short-read assembly**

For Illumina-derived reads, the last four and six bases were trimmed from the 36mer and 76mer reads, respectively. We filtered out all Illumina-derived short reads that had ambiguous bases. For the paired-end reads, both reads were filtered out if one read of a pair had ambiguous bases. We used Velvet 0.7.55 to *de novo* assemble the reads (Zerbino and Birney, 2008). We assembled short reads from the different sequencing platforms independently, as well as in combination. We wrote *ad hoc* shell scripts to test different Velvet parameters of hash length, coverage cutoff, and expected coverage. In total, we generated approximately 75 different genome assemblies of WH6. Shell scripts are available for download (http://changlab.cgrb.oregonstate.edu).

**Improvements to the high-quality draft assembly**

We developed *ad hoc* Perl scripts to use BLASTN to compare between each of the WH6 assemblies and used congruency in contigs from the various assemblies to cull those with potential misassemblies (see next section for description of scripts; data were visualized using blast_draw.pl). We used Tablet 1.10.01.28 to inspect the remaining genome assemblies for depth of coverage and potential misassemblies (Milne et al., 2010). Finally, we used Mauve Aligner 2.3 and the genome sequence of *P. fluorescens* SBW25 as a reference to reorder WH6 contigs greater than 100 bp from our assembly with highest confidence (Rissman et al., 2009; Silby et al., 2009). Default settings were used for Mauve Aligner 2.3.

**Physical and sequence gap closure**

To identify contigs that potentially flanked a physical gap, we wrote and used Contig_end_blast_A.pl, to extract 300bp of sequence from the ends of each contig greater than one kb in size and use the contig ends as queries in a BLASTN search against the NCBI nt database. We also wrote and used Contig_end_blast_B.pl to find contig ends that shared significant homology (e-value ≤ 0.02) to the same reference sequence but aligned to different regions no more than one kb apart. The contigs corresponding to these ends were thus predicted to be physically linked in the genome of WH6. PCR using contig-specific primers and subsequent Sanger sequencing were used to close the physical gaps (See Additional file 2: Table S2; http://www.biomedcentral.com/1471-2164/11/522/additional). To correct the sizes for sequence gaps larger than 300

bp, we used a similar approach. PCR was used to validate our corrections for sequence gaps.

Contig_end_blast_A.pl, Contig_end_blast_B.pl, and blast_draw.pl are available for download from our website at: http://changlab.cgrb.oregonstate.edu.

**Genome Annotation**

We used a custom pipeline to annotate the improved high-quality draft assembly of WH6 as previously described (Giovannoni et al., 2008). The only exceptions were that we used Glimmer 3.02 rather than Glimmer 2 to predict coding regions and gene models were trained using the "long-orfs" option ((Delcher et al., 1999); http://www.cbcb.umd.edu/software/glimmer/).

**Bioinformatic analyses**

For analysis of synteny, we first parsed the genomes of SBW25, Pf-5 and Pf0-1 into all possible 25mers and identified their unique 25mer sequences. Next, we used CASHX to align all unique 25mers from each of three genomes to both strands of a formatted database from the WH6 genome sequence (Fahlgren et al., 2009). Only perfect matches were allowed. We identified the corresponding genome coordinates for each 25mer and the matching 25mer in the WH6 genome and used R to plot the start coordinates of each matching pair in an XY graph (R Development Core Team).

Phylogenomic relationships were determined using HAL ((Robbertse et al., 2006); http://aftol.org/pages/Halweb3.htm). HAL uses an all-by-all reciprocal BLASTP to create a similarity matrix from e-values. These are then used to group proteins into related clusters using a Markov Clustering algorithm. Clusters containing one protein sequence from each genome that identified each other as

best hits were extracted, concatenated within each proteome, and used to infer phylogenetic relationships. Phylogenetic trees were visualized using the Archaeopteryx & Forester Java application ((Zmasek and Eddy, 2001); http://www.phylosoft.org/archaeopteryx/).

Hidden Markov Models (HMMs) for *hrp*-boxes were trained from a set of 38 confirmed *hrp*-boxes in the *P. syringae* pv *tomato* DC3000 genome (Buell et al., 2003; Chang et al., 2005; Ferreira et al., 2006; Schechter et al., 2006). The HMM for the extracytoplasmic function $\sigma^{70}$ factors was downloaded from http://www.g2l.bio.uni-goettingen.de/software/f_software.html. Searches were done using HMMER 2.3.2 (http://hmmer.janelia.org/).

Circular diagrams were plotted using DNAplotter ((Carver et al., 2009); http://www.sanger.ac.uk/Software/ Artemis/circular/).

## ACKNOWLEDGMENTS

We thank Mark Dasenko and Chris Sullivan of the Center for Genome Research and Biocomputing (CGRB) for Illumina sequencing and computational support, as well as Don Chen and Philip Hillebrand for their assistance. We thank Dr. Joey Spatafora for providing us HAL before publication, Dr. Joyce Loper and Jason Cumbie for their valuable advice. Finally, we thank two anonymous reviewers for their helpful comments in improving this manuscript. This research was supported in part by General Research Funds to JHC and the National Research Initiative Competitive Grant no. 2008-35600-18783 from the USDA's National Institute of Food and Agriculture, Microbial Functional Genomics

Program to JHC, and by grants from the USDA CSREES Grass Seed Cropping Systems for Sustainable Agriculture Special Grant Program and from the OSU Agricultural Research Foundation to DJA and DIM.

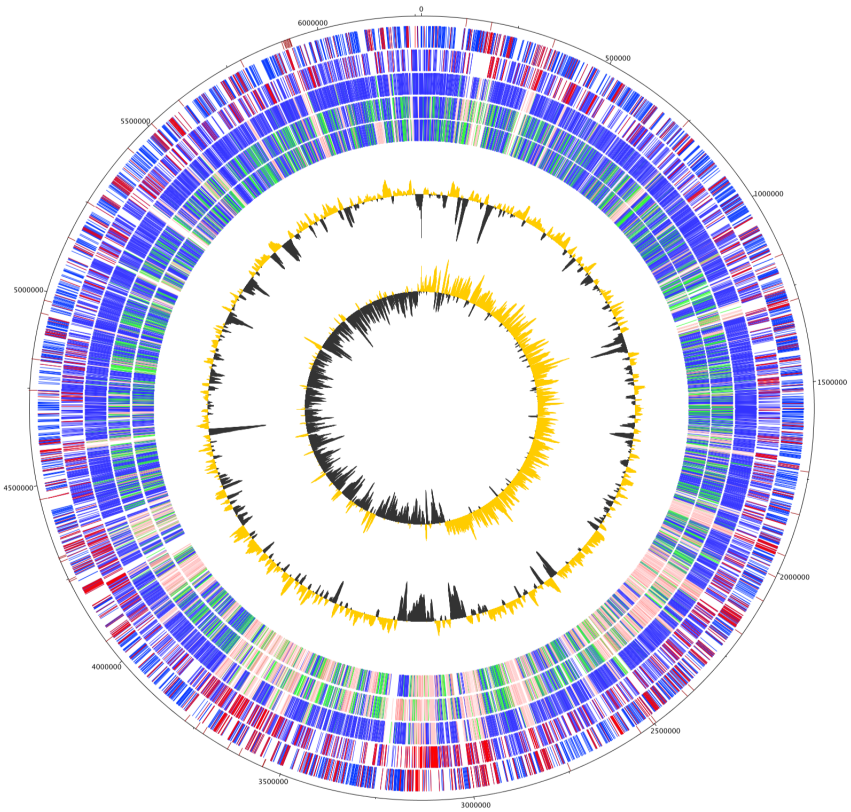



**Figure 3.1. Circular representation of the improved, high-quality draft genome sequence of WH6.**

The outer scale designates the coordinates in half million basepair increments. The red ticks indicate physical gaps. Circles 2 and 3 show the predicted coding regions of WH6 on the positive and negative strands, respectively. Coding regions are colored to highlight orthologous (blue) and 1567 unique (red) coding regions of WH6 relative to the other sequenced *P. fluorescens*. Circles 4, 5, and 6 show orthologs (BLASTP e-value ≤ 1 x $10^{-7}$) of SBW25, Pf-5, and Pf0-1, respectively. The extent of homology relative to WH6 is depicted using a heat map of arbitrarily chosen bins; dark blue: orthologs with greater than 80% homology over the length of the gene; green: orthologs with between 60-80% homology over the length of the gene; pink: orthologs with between 20-60% homology over the length of the gene; white: no homology (less than 20% homology over the length of the gene). The positions of loci of interest are also denoted (see corresponding text for more details). Circles 7 and 8 show GC% (gold >60.6% average; gray < 60.6% average) and GC-skew.

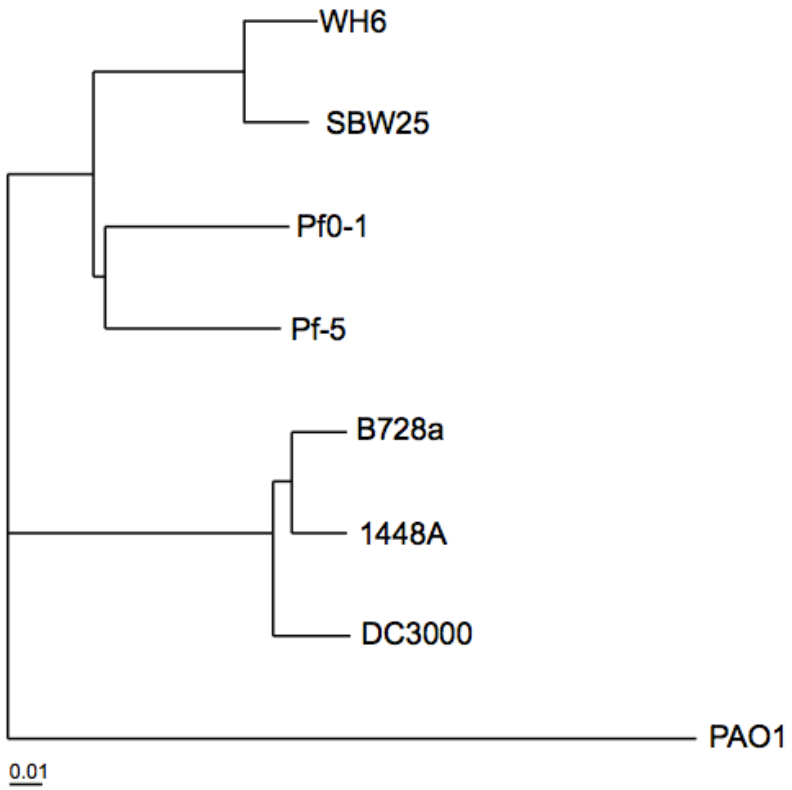



**Figure 3.2: Phylogenomic tree of eight *Pseudomonas* isolates based on a super alignment of 1966 translated sequences.**

*P. fluorescens* isolates: WH6, SBW25, Pf-5, and Pf0-1; *P. syringae* pathovars: *tomato* DC3000, *phaseolicola* 1448A, and *syringae* B728a; *P. aeruginosa* PAO1. Bootstrap support for nodes (r = 1000) were all 100. The scale bar indicates the number of amino acid substitutions per site.

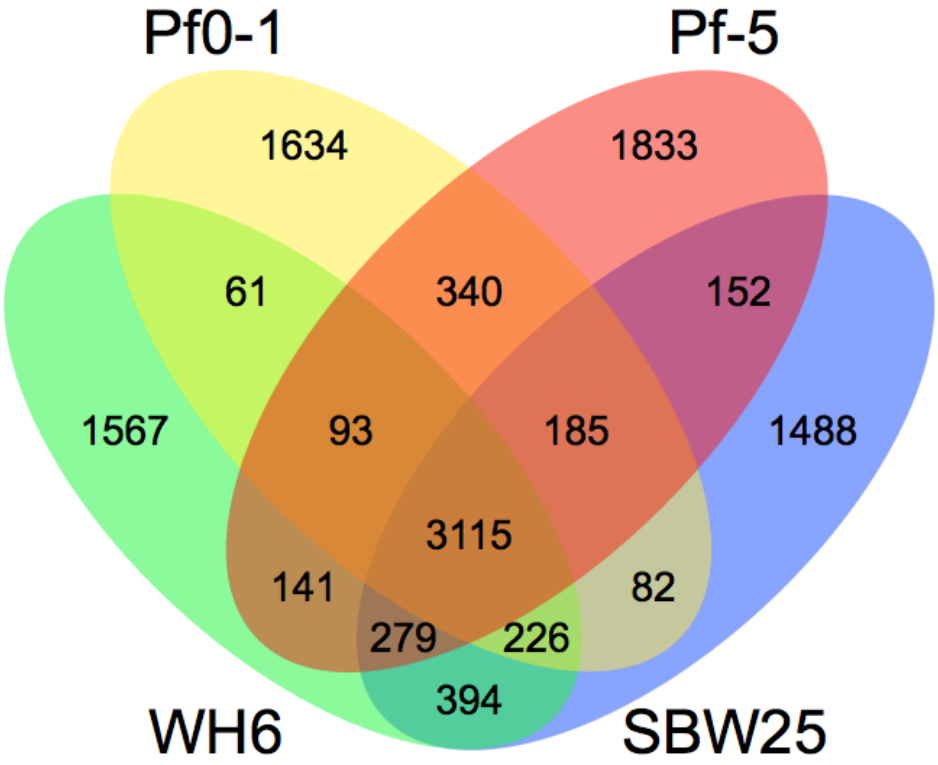



**Figure 3.3: Venn diagram comparing the gene inventories of four isolates of *P. fluorescens*.**

The numbers of shared and unique genes are shown. Comparisons were made using HAL.

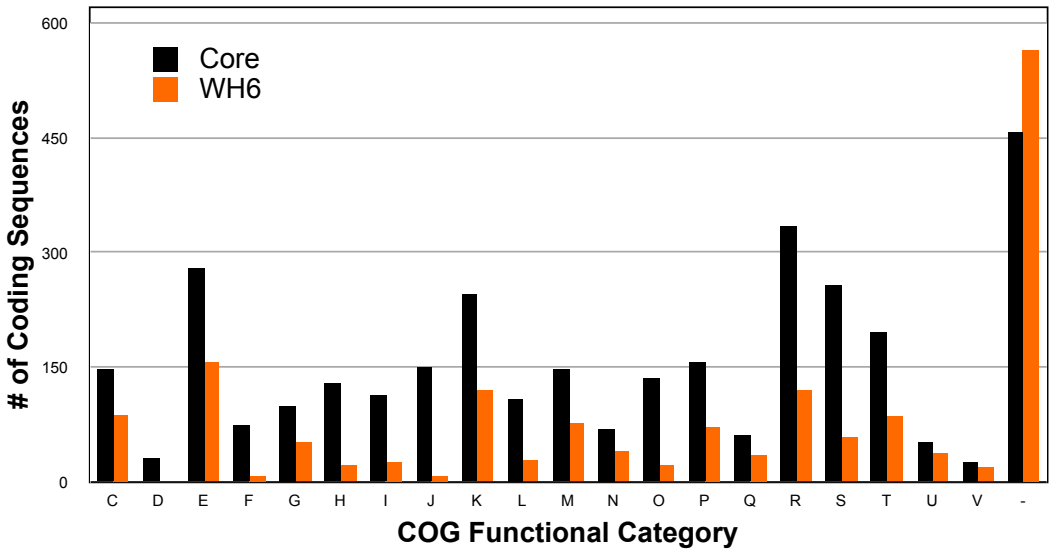



**Figure 3.4: Functional categories of the 3115 core genes of *P. fluorescens* and 1567 unique genes of WH6.**

The number of genes in each category are presented above each bar; black = core; orange = WH6. Categories: **C)** energy production and conversion; **D)** cell cycle control, cell division, chromosome partitioning; **E)** amino acid transport and metabolism; **F)** nucleotide transport and metabolism; **G)** carbohydrate transport and metabolism; **H)** coenzyme transport and metabolism; **I)** lipid transport and metabolism; **J)** translation, ribosomal structure and biogenesis; **K)** transcription; **L)** replication, recombination and repair; **M)** cell wall/membrane/envelope biogenesis; **N)** cell motility; **O)** posttranslational modification, protein turnover, chaperones; **P)** inorganic ion transport and metabolism; **Q)** secondary metabolites biosynthesis, transport and catabolism; **R)** general function; **S)** unknown function; **T)** signal transduction mechanisms; **U)** intracellular trafficking, secretion and vesicular transport; **V)** defense mechanisms; **-)** no COG designation.

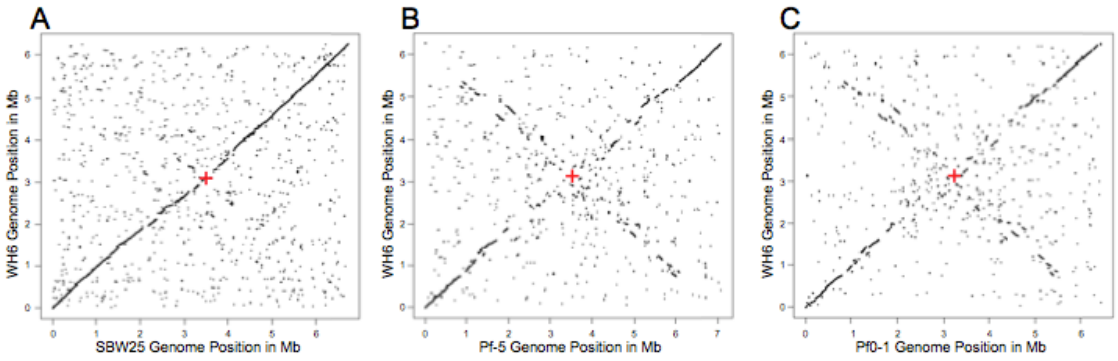



**Figure 3.5: Synteny plots comparing the organization of the WH6 genome to that of the three other isolates of *P. fluorescens*.**

Unique 25mers from *P. fluorescens* isolates SBW25 **(A)**, Pf-5 **(B)**, and Pf0-1 **(C)** were compared to the improved, high-quality draft genome sequence of WH6. The start positions of all matching pairs were plotted in an XY graph with the coordinates of the genomes of SBW25, Pf-5, and Pf0-1 along the x-axis and coordinates of the genome of WH6 along the y-axis. The termini are located at the approximate mid-way point for each comparison (red **+**). Genome scales are shown in one Mb increments.

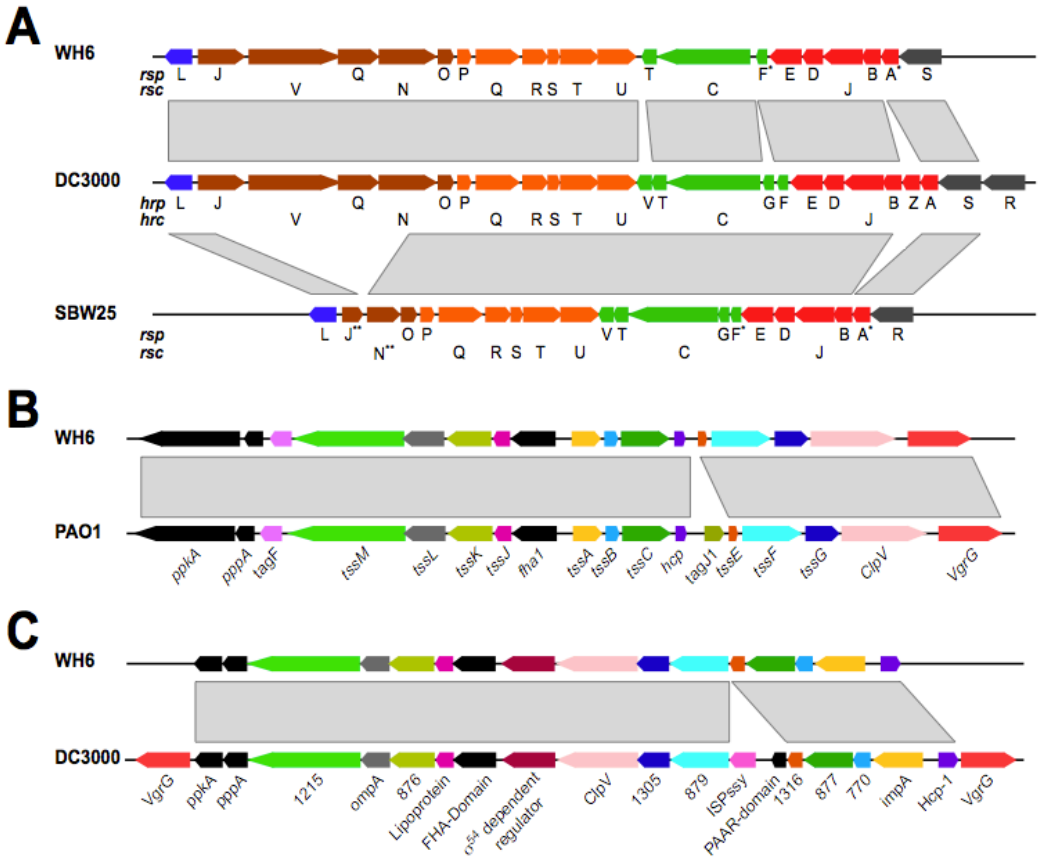



**Figure 3.6: Schematic Representations and Comparisons of Type III and Type VI Secretion Systems.**

**A)** The type III secretion system of WH6 (top) compared to that of *P. syringae* (middle), and SBW25 (bottom). Co-transcribed genes and orthologous transcriptional units are colored similarly. *No detectable homology but inferred based on location in the T3SS-encoding locus. **Truncated coding regions. **B)** The type VI secretion system-1 of WH6 (top) compared to HSI-I of *P. aeruginosa* PAO1. **C)** The type VI secretion system-2 of WH6 (top) compared to a candidate type VI secretion system of *P. syringae* pv *tomato* DC3000 (bottom). Orthologous transcriptional units are colored similarly. For **A-C**, the directions of transcription are represented. Gray boxes highlight homologous regions.

**Table 3.1. High-throughput sequencing statistics**

| Method | Total Reads | Reads Used | Theoretical Coverage* | # Contigs (>100bp)^ | Total size (Mb)[§] |
|---|---|---|---|---|---|
| Sanger | 178 | 178 | n/a | n/a | n/a |
| GAI 32 single | 16,852,820 | 9,298,356 | 83 | 5,884 | 6.06 |
| GAII 70 PE | 10,854,745 | 9,013,849 | 234 | 95 | 6.27 |
| 454 | 202,070 | 200,467 | 7 | 2,204 | 6.10 |
| All short reads | 38,764,380 | 23,742,926 | 316 | 256 | 6.26 |
| All reads | 38,764,558 | 23,810,966 | 316 | 115 | 6.27 |

*Based on approximated genome size of 6.5 Mb; ^Highest confident draft assemblies using Velvet 0.7.55 (Zerbino and Birney, 2008); ^Improved, high-quality draft including Sanger reads. n/a = not applicable; [§]Based on sum total of all contigs > 100 bp in length.

**Table 3.2. Comparison of *P. fluorescens* genome characteristics**

| Isolate | WH6* | SBW25^ | Pf-5[§] | Pf0-1^ |
|---|---|---|---|---|
| Genome Size | 6.27 Mb | 6.72 Mb | 7.07 Mb | 6.44 Mb |
| GC % | 60.6 | 60.5 | 63.3 | 60.5 |
| Coding regions | 5876 | 5921 | 6138 | 5736 |
| Avg. length of coding sequences | 951 | 1000 | 1020 | 1006 |
| Coding % | 89.2 | 88.1 | 88.5 | 89.6 |

*Improved, high-quality draft genome sequence ^(Silby et al., 2009); [§](Paulsen et al., 2005).

**Table 3.3. Candidate Host-association and virulence factors***

| WH6 Gene (PFWH6_) | Host-association or virulence factor | Reported function | Reference |
|---|---|---|---|
| 0718-0737 | T3SS-encoding region | Host-association; secretion apparatus | (Alfano et al., 2000; Preston et al., 2001) |
| 5796-5812 | T6SS-1-encoding region | Host-association; secretion apparatus | (Pukatzki et al., 2006; Bingle et al., 2008; Shrivastava and Mande, 2008) |
| 3251-3270 | T6SS-2-encoding region | Host-association; secretion apparatus | (Bingle et al., 2008; Shrivastava and Mande, 2008) |
| 0824-0827 | Betaine/choline uptake | Osmoprotection | (Chen and Beattie, 2007) |
| 5455 & 0252 | BCCT Transporter | Choline transport | (Chiliang Chen, 2008) |
| 5456-5458 | Choline to Betaine | Choline to Betaine conversion | (Tettelin et al., 2008) |
| 1723 & 2895 | *aprA*^ | Alkaline protease A; Insecticidal toxin | (Tan and Donovan, 2000) |
| 4097/98, 4099, 4100 | *tca-d* | Insecticidal toxin | (Bowen et al., 1999) |
| 5264, 5082, 3503 & 0070 | *katA* | Catalase A; $H_2O_2$ protection | (Ji-Sun Lee, 2005) |
| 0985 - 0996 | Synthesis of Alginate | Exopolysaccharide | (Yu et al., 1999) |
| 2199 | Synthesis of Levansucrase^ | Exopolysaccharide | (Hettwer et al., 1998; Koczan et al., 2009) |
| 4225 | *marR* | Transcriptional regulator virulence factors | (Ellison and Miller, 2006) |
| 3833 – 3843 | T2SS-encoding region | Secretion apparatus | (Johnson et al., 2006) |
| 0699 | *plc* lipase | Phospholipase C; virulence factor | (Meyers and Berk, 1990) |
| 0396 – 0398 | *TAT-encoding region | Secretion apparatus | (Bronstein et al., 2005; Caldelari et al., 2006) |
| 4428, 2727, 0116-0120 | Synthesis of mangotoxin | Antimetabolite toxin | (Arrebola et al., 2010) |
| 2331-2334 | Synthesis of hydrogen cyanide | Inhibitor of cytochrome c oxidase | (Gross and Loper, 2009) |

*Table derived from (Lindeberg et al., 2008); candidates were identified using BLASTP (e-value 1 x $10^{-7}$). There are 43 WH6 proteins with homology to candidate TAT-secreted proteins of *P. syringae* pv *tomato* DC3000 (Bronstein et al., 2005; Caldelari et al., 2006); ^no orthologs were detected in genomes of other *P. fluorescens*.

**Table 3.4. Putative *hrp*-boxes and candidate type III effector genes in WH6**

| Bit Score | *hrp*-box sequence | CDS* | Distance from *hrp*-box | Translocation Signal^ |
|---|---|---|---|---|
| 7.9 | TGGAACTGAATAGCCAGTACTGACCAC | *rspF* | 26 | - |
| 6.7 | AGGAACCGATTCCGACAGATGGGCCAC | 1942 | 77 | A, B |
| 5.3 | CGGAACCTTTACCGGCACCTGAACACT | 2917 | 248 | A, B, C |
| 5.1 | TGGAACGAAATCGTCGATCAAACCACT | 3173 | 58 | A, C |
| 5 | TGGAACCGTATTGCGTAAGACGTCACT | 1252 | 82 | - |
| 4.3 | GGGAACCGCATCGGTTGCCTTCCAAC | 3940 | 56 | - |
| 3.6 | TGAAACCGGCACGGCGTGCCTGACCCT | *rscR* | 324 | A |
| 1.2 | TGGAACCAGGTGGGCGGGGCTTGCCAC | *rspJ* | 33 | A, B |

Only coding sequences with no predicted homology or identifiable orthologs in Pf-5 or Pf0-1 are shown. *PFWH6_# unless otherwise noted; ^N-terminal translocation signal scores were assigned based on “A”; >10% serine in first 50 amino acids, “B”; absence of aliphatic amino acids in position 3 or 4, and/or “C”; absence of negatively-charged amino acids in the first 12 amino acids (Petnicki-Ocwieja et al., 2002).

# Evolutionary stasis of type III effector genes in mutualistic *Sinorhizobium fredii* and *Bradyrhizobium japonicum*

Jeffrey A. Kimbrel, William J. Thomas, Allison L. Creason, Caitlin A. Thireault,

Jeff H. Chang

**ABSTRACT**

The symbiotic relationships between nitrogen-fixing rhizobia and legume hosts are initiated following complex molecular dialogs that navigate multiple layers of specificity. One barrier that must be overcome by rhizobia and other bacteria is the plant immune system. Several species of rhizobia encode for a type III secretion system, which for phytopathogens, function to deliver collections of type III effector proteins directly into cells to perturb host defenses. As a first step towards determining whether type III effector proteins of mutualistic rhizobia play similar roles in suppressing host cell defenses, we cataloged type III effectors from eight strains distributed across the *Sinorhizobium fredii* and *Bradyrhizobium japonicum* species, selected based on phylogenetic divergence and demonstrable reliance on type III secretion systems for host infection. Draft and finished genome sequences were mined and candidate type III effector proteins were confirmed for direct delivery into plant cells using a heterologous delivery system. This is the largest and most extensive investigation into collections of type III effector genes from mutualistic bacteria. Over 300 type III effector-encoding genes, representing 60 different families, were confirmed. Each strain encoded for large collections, between approximately 15 and 35 type III effector genes, depending on the species. Interestingly, analyses of type III effector collections revealed extremely high conservation in content and nucleotide sequence with little evidence for diversity, suggesting an evolutionary stasis of type III effector genes in plant mutualistic bacteria.

## INTRODUCTION

The rhizobia are a diverse collection of α-proteobacteria, many of which are able to form nitrogen-fixing nodules in compatible host genotypes. This fixation of nitrogen is a critical component of the nitrogen cycle, establishing a crucial role for the rhizobia-legume relationship in both natural ecosystems and agricultural environments. The symbiotic relationship begins with a complex molecular dialog between the rhizobium and plant roots, culminating in the formation of root nodules. This conversation begins when detection of legume-specific flavonoids induces the expression of nodulation (*nod*) genes in compatible rhizobia (Peck et al., 2006; Jones et al., 2007). The products of the *nod* genes synthesize nod factors, lipo-chitooligosaccharides secreted during the initial stages of the plant-rhizobia interaction. Rhizobial nod factors are perceived by cognate receptors of compatible legume hosts, resulting in morphological changes to the host such as the formation of an infection thread and cortical cell activation (Spaink, 2000; Haney et al., 2011).

Host specificity is affected by other factors of both symbiont and host (Maria Lopez-Lara et al., 1995; Niehaus and Lagares, 1998; Jones et al., 2007; Hidalgo et al., 2010). The plant defense system could also contribute to affect the specificity of legume-rhizobial interactions (Soto et al., 2009; Zamioudis and Pieterse, 2012). The first layer of plant defense is the so-called PAMP- or Pattern-triggered immunity (PTI; (Jones and Dangl, 2006; Schwessinger and Zipfel, 2008)). In this layer, pattern recognition receptor (PRR) proteins detect conserved pathogen- or microbe- associated molecular patterns (PAMPs or MAMPs),

resulting in a battery of responses collectively necessary for defense (Zipfel, 2008; Boller and He, 2009; Postel and Kemmerling, 2009; Zipfel, 2009; Segonzac and Zipfel, 2011).

Many Gram-negative, host-associated pathogens use a type III secretion system (T3SS) to manipulate host cells, including the perturbation of host defenses to suppress or dampen PTI (He et al., 2004; Galán and Wolf-Watz, 2006; Grant et al., 2006; Galán, 2009). The T3SS is a conduit for the direct delivery of bacterial-encoded type III effector proteins (T3Es) into host cells. For plant pathogens, the T3Es are collectively necessary to dampen defense below a threshold required for effective resistance (Jones and Dangl, 2006; Cunnac et al., 2009; Lewis et al., 2009).

Individual T3Es of plant pathogens also have the potential to trigger an additional layer of defense. Plants encode for disease resistance proteins (R proteins) that perceive the presence or action of corresponding microbial effectors, resulting in effector-triggered immunity (ETI, aka “gene-for-gene resistance”; (Jones and Dangl, 2006; Cui et al., 2009; Mansfield, 2009; Rafiqi et al., 2009)). The outputs of ETI and PTI are similar but the former is associated with more robust and amplified responses and often a hypersensitive response (HR), a localized programmed cell death (Greenberg and Yao, 2004). R proteins often have nucleotide binding (NB) and leucine-rich repeat (LRR) motifs whereas effectors that elicit ETI, were classically designated as “avirulence proteins” (Avrs) because they rendered the pathogen avirulent.

For plant pathogens, the co-evolution of host defense and T3Es has been

modeled as an “evolutionary arms race” (Stavrinides et al., 2008). In this setting, collections of T3E genes are predicted to exhibit patterns of genetic variation that reflect rapid evolution. Indeed, analyses of plant pathogens have provided ample support for this model (Ma et al., 2006; Zhou et al., 2009; Baltrus et al., 2011; Jackson et al., 2011). In *Pseudomonas syringae*, for example, collections of T3E genes vary dramatically in size and content with few that would be considered “core” (Baltrus et al., 2011). Another aspect of pathogen T3E collections is that their robustness is ensured via redundancy so that any individual T3E gene is dispensable (Stavrinides et al., 2008; Cunnac et al., 2009; 2011). Thus, the loss and gain of genes contribute to the dynamic nature of pathogen T3E collections.

Studies of mutualistic bacteria have suggested that they and pathogens share many “virulence” mechanisms in common (Soto et al., 2009). Indeed, T3SS-encoding loci have been identified in rhizobia, including all examined strains of *Bradyrhizobium japonicum*, *Sinorhizobium fredii* strains USDA207, USDA257, USDA191, and *Rhizobium* NGR234 (which was reclassified as a *S. fredii*), as well as *Mesorhizobium loti* MAFF303099 (Bellato et al., 1996; Freiberg et al., 1997; Viprey et al., 1998; Kaneko et al., 2000; Mazurier et al., 2006). Characterization of the T3SS and candidate T3Es of rhizobia have provided evidence that mutualistic bacteria rely on this virulence mechanism to perturb PTI.

The T3SS genes are co-regulated with the *nod* genes, pointing to an early role, as is the case with plant pathogenic bacteria (Viprey et al., 1998; Krause et al., 2002; Wassem et al., 2008; Haapalainen et al., 2009). Flavonoid perception elicits the *nodD* regulatory cascade, which regulates the expression of *ttsI*, a two-

component regulator-like encoding protein (Krause et al., 2002; Marie et al., 2004). TtsI is proposed to bind an upstream *cis* regulatory element, the *tts*-box, that is found upstream of genes encoding for components of the T3SS, candidate T3Es, and other T3SS-associated accessory proteins (Zehner et al., 2008). Mutants compromised in the construction or expression of the T3SS mutants are compromised in establishing interactions with their otherwise compatible hosts (Meinhardt et al., 1993; de Lyra et al., 2006; Deng et al., 2009). In rhizobia even loss-of-function mutants of T3SS-secreted proteins and potential T3Es, NopT and NopJ (y4lO) of NGR234, are significantly compromised in nodulating *Crotalaria juncea* and *Vigna unguiculata*, respectively, and similar observations have been made for several other candidate T3Es of other rhizobia (Annapurna and Krishnan, 2003; Krishnan et al., 2003; Marie et al., 2003; Ausmees et al., 2004; Lorio et al., 2004; Saad et al., 2005; Rodrigues et al., 2007; Dai et al., 2008; López-Baena et al., 2008; Zehner et al., 2008; Hempel et al., 2009; Kambara et al., 2009; Yang et al., 2009; Schechter et al., 2010; Wenzel et al., 2010). In all, these observations indicate an important function for the T3SS for these strains of rhizobia.

T3E of mutualists, like their counterparts in pathogens, also affect bacterial host range (Fauvart and Michiels, 2008). Rhizobia mutants deleted of confirmed or putative T3E-encoding genes can evade detection and exhibit expanded host ranges (Bellato et al., 1997; Ausmees et al., 2004; Skorpil et al., 2005; Dai et al., 2008; Yang et al., 2009; Schechter et al., 2010; Wenzel et al., 2010). “Nodulation restriction”, i.e., R-avr interaction, was proposed to affect legume-rhizobium

interactions based on genetic and molecular data paralleling that of plant-pathogen interactions (Triplett and Sadowsky, 1992). Indeed, loci responsible for restricting soybean nodulation by certain strains of rhizobia encode for NB-LRR proteins and mapped to regions with clusters of other NB-LRR-encoding genes that are linked to resistance against phytopathogens (Kanazin et al., 1996; Graham et al., 2002a; 2002b). Together, these data provide tantalizing evidence for ETI between plants and mutualistic rhizobia.

Here, we describe the most extensive genome-enabled survey for T3Es from mutualistic bacteria. Three and five strains from *S. fredii* and *B. japonicum*, respectively, were selected for characterization based on phylogenetic diversity and/or their demonstrable T3SS-dependent polymorphic behavior in nodulating different host genotypes (Bellato et al., 1997; Pueppke and Broughton, 1999; Göttfert et al., 2001; Mazurier et al., 2006). Four finished genome sequences, some of which were completed during the course of our work, are available (Freiberg et al., 1997; Kaneko et al., 2000; 2002; Schmeisser et al., 2009; Kaneko et al., 2011; Margaret et al., 2011). In addition, we generated draft genome sequences for four additional strains. Candidate T3E-encoding genes were computationally identified based on association with the *tts*-box and experimentally confirmed based on their T3SS-dependent translocation into plant cells, using the plant pathogen, *P. syringae* pv *tomato* DC3000 (*Pto*DC3000).

Each of the characterized strains encoded extensive collections of T3E-encoding genes. However, in stark contrast to observations in plant pathogenic bacteria, analyses of T3Es of rhizobia yielded very little evidence in support for an

evolutionary arms race. Both *B. japonicum* and *S. fredii* had large cores of conserved T3E-encoding genes in which the large majority exhibited little sequence variation, suggestive of purifying selection. In total, our results suggest that T3E-encoding genes exhibit patterns indicative of a mutualistic environment wherein mutualist and host evolve to benefit their association with each other.

## RESULTS

### Draft genome assemblies for strains of T3SS - encoding rhizobia

We focused on *S. fredii* and *B. japonicum*, two species of rhizobia well characterized for requiring a T3SS for optimal infection of their host plants (Viprey et al., 1998; Mazurier et al., 2006). At the onset of our work, the only available finished genome sequence was from *B. japonicum* USDA110 (Kaneko et al., 2002). For a more species-level characterization of T3E-encoding genes and to facilitate the characterization of strains with demonstrable T3SS-dependent changes in host range, we used paired-end Illumina sequencing to generate draft genome sequences for several more strains of *S. fredii* and *B. japonicum*. We selected *S. fredii* NGR234, USDA207, and USDA257 as well as *B. japonicum* USDA6, USDA122, USDA123, and USDA124 (Mazurier et al., 2006). The paired-end reads were *de novo* assembled and the contigs were ordered using a suitable reference genome sequence. Since the finished genome sequence for NGR234 was published early in our study, we used it rather than the draft genome sequence (Freiberg et al., 1997; Schmeisser et al., 2009). Sequencing statistics are provided as supplemental information (Supplemental Table 1).

The assemblies were adequate for mining for T3E-encoding genes and

several, if judiciously used, were informative for genome comparisons (Pop and Salzberg, 2008; Klassen and Currie, 2012). The draft genome sequences varied in their quality, due in part to the different Illumina sequencing platforms that were used (Supplemental Table 4.1). Nevertheless, the estimated genome sizes and predicted numbers of open reading frames (ORFs) were similar within each species and to finished reference sequences (Table 4.1). With a stringent criterion of ≥90% identity, 85.1% and 96.1% of the translated ORF sequences annotated in the draft USDA207 and USDA6 genome sequence, respectively, were similarly annotated in their corresponding finished genome sequence (data not shown). Secondly, analysis of aligned conserved regions, such as the T3SS-encoding locus, further indicated the assemblies were of high quality (Supplemental Figure 4.1a). Finally, the draft genome sequences from USDA207 and USDA6 were similar in genome order to those of corresponding genome sequences that were finished subsequent to our efforts (Data not shown; (Kaneko et al., 2011; Margaret et al., 2011)).

**Phylogenomic comparisons reveals high orthology**

We next determined the extent of orthology for all pairwise combinations of strains. Within *S. fredii*, there was low orthology between the three strains. The percentage of orthologous genes ranged from only 54.6% to 66.3% between all of the pairwise comparisons (Figure 4.1A). Interestingly, USDA207 and USDA257 were more similar to NGR234 than they were to each other. This observation supports previous analyses of rhizobia that supported the reclassification of *Rhizobium* sp. NGR234 as a *Sinorhizobium* (Saldaña et al., 2003). However,

despite the relatively low percentage in overlap of genes, there were nonetheless high levels of identity between orthologous pairs, with the majority having more than 90% identity (Supplemental Figure 4.2A).

Within *B. japonicum*, the strains had a substantially higher percentage of orthologous genes, with a high of 83.6% between USDA110 and USDA122 and a low of 65.4% between USDA110 and USDA6 (Figure 4.1A). The percent orthology in the latter comparison was similar to that recently reported at 68.5% similarity between the genes of USDA6$^T$ and USDA110 (Kaneko et al., 2011). Additionally, relative to genes of USDA110, the majority of orthologous pairs shared more than 90% identity in sequence along at least 90% of the length of the nucleotide sequence, with USDA122 showing the highest amount of identity at 99% (Supplemental Figure 4.2B). Here again, this range was on the same level as previously reported based on comparisons of finished genome sequences of USDA6$^T$ and USDA110 (Kaneko et al., 2011).

Phylogenetic trees generated from different loci within the rrn operon of rhizobia have different topologies (van Berkum et al., 2003). To address this inconsistency, we used a whole genome phylogeny based on 624 translated orthologous sequences to determine the taxonomical relationship of the strains studied here (Figure 4.1B). We also included *Mesorhizobium loti* MAFF303099 and *Azorhizobium caulinodans* ORS571 (Kaneko et al., 2000; Lee et al., 2008). The former strain encodes for a T3SS whereas the latter does not. The topology of the tree was similar to those generated using 16S or 23S rRNA-encoding regions (van Berkum et al., 2003). *S. fredii* and *M. loti* share a common node but,

in addition to *B. japonicum,* clearly separated into different groups. The *B. japonicum* strains have much shorter branch lengths than the *S. fredii* strains, indicating either a slower substitution rate or a more recent divergence. The relationship of USDA6 and USDA123 contrasted to previous classifications based on the ITS region but were in line with topologies based on the 16S rRNA gene sequence (van Berkum and Fuhrmann, 2000; de Oliveira et al., 2006). Regardless, a tree based on whole genome comparisons supported, in general, previous inferences of strain diversity based on single gene sequences and results from reciprocal BLASTP analysis (Figure 4.1A).

**Mining genomes for candidate type III effector – encoding genes**

Few type III effector (T3E)-encoding genes have been identified from genome sequences of mutualistic bacteria. This is not surprising since few T3Es have sufficient sequence similarities between different genera of bacteria for comprehensive detection using homology-based approaches (Grant et al., 2006). We therefore searched for candidate T3E-encoding genes based on the presence of an upstream *tts*-box. We used sequences of 30 confirmed functional *tts*-boxes from *B. japonicum*, *S. fredii* and *M. loti* MAFF303099 to train a Hidden Markov Model (HMM; (Marie et al., 2004; Zehner et al., 2008; Sánchez et al., 2009). The score of significance was calibrated to 5.0 based on the identification of 11 functionally validated *tts*-boxes located on the pNGR234a megaplasmid (Marie et al., 2004).

We identified a total of 305 putative *tts*-boxes from the eight finished and draft genome sequences (Table 4.2). In *S. fredii*, 13~24 putative *tts*-boxes were

identified from the three genome sequences whereas in *B. japonicum*, upwards of 52 were found. The narrow variation in the number of putative *tts*-boxes within each of the two species of rhizobia was encouraging but nevertheless unexpected relative to the high number typically found using HMM-based searches for *hrp*-box sequences in genome sequences of pathogenic *P. syringae* (Ferreira et al., 2006). The difference in numbers of putative *tts*-boxes between rhizobia species, in contrast, was not surprising and likely reflects the difference in genome sizes.

In NGR234, other than the 11 *tts*-boxes previously found, we identified two additional sequences that were 1,345 bp and 1,119 basepairs (bp) upstream from NGR_a02270 (y4oB) and NGR_a00810 (NodD2), respectively (Marie et al., 2004). A fourteenth *tts*-box-like sequence was found 248 bp upstream of *fixC* (NGR_a01240), but it was not considered because of its low bitscore of 3.1. As previously reported, no other putative *tts*-boxes were found in the chromosome of NGR234 or pNGR234b (Schmeisser et al., 2009). A second T3SS-encoding locus is present on pNGR234b but its necessity in host infection has not been confirmed (Schmeisser et al., 2009).

In the finished genome sequence of USDA110, we found 52 *tts*-boxes with a bitscore of 5.0 or higher (Table 4.2). Of these 52, 29 were previously identified with 14 of these *tts*-boxes located upstream of thirteen genes (bll1862 has two upstream *tts*-boxes) that encode proteins that are secreted in a T3SS-dependent manner (Zehner et al., 2008). This was expected since these exact 14 *tts*-box sequences were used to train the HMM used in this study. However, our search failed to identify the other 10 found by Zehner, *et al.* (2008), of which none of the

downstream-encoded gene products were shown to be secreted in a T3SS-dependent manner.

To identify candidate T3E-encoding genes, we searched up to 10 kb downstream of the 305 *tts*-boxes. The reason for this relaxed criteria stems from observations that TtsI-regulated operons, such as the *nopB-rhcU* operon of NGR234, can be substantial in length (Perret et al., 2003). We identified a total of 403 ORFs but culled the list down to 277 by filtering out those with translated sequences homologous to components of the T3SS or proteins with functions atypical of T3Es (BLASTX e-value ≤ $1x10^{-7}$; table 4.2). Because the draft genome sequences were distributed across many contigs and T3E could be encoded in long operons, there was potential for physical gaps to uncouple ORFs from their *tts*-box. Use of BLASTN to identify homologs of the 277 candidate T3Es-encoding genes led to an additional 104 ORFs. Based on *in silico* analysis, we predict that ~60% of the T3E-encoding ORFs could be potentially transcribed from an operon.

The 381 candidate T3E-encoding genes clustered into 97 families. As expected of candidate T3Es and as a consequence of the filtering, 44% of the families had no members with matches to features in the NCBI conserved domain database (Marchler-Bauer et al., 2011). Surprisingly, there is a high degree of conservation in candidate T3Es within and across *S. fredii* and *B. japonicum* species. Nearly 80% of the candidate T3E families had an ortholog present in more than one strain and within *B. japonicum*, 47% of the families had an ortholog in all strains examined herein. Ten families had orthologous genes common to both species.

**T3SS-dependent translocation of type III effectors**

The T3SS-dependent translocation of a protein directly into host cells is the one defining characteristic of a T3E. Cya has been used to demonstrate T3SS-dependent translocation by rhizobia but assays required up to two weeks to complete (Casper-Lindley et al., 2002; Schechter et al., 2010; Wenzel et al., 2010). Due to the large number of candidates to be tested, we elected to use the ∆79AvrRpt2 reporter, which gives a more rapid response approximately 20 hours post inoculation (hpi) in the model plant, Arabidopsis Col-0 (Mudgett and Staskawicz, 1999; Guttman and Greenberg, 2001; Chang et al., 2005). Furthermore, we employed an additional criterion based on amino acid divergence to reduce the onerousness of testing, and selected a single candidate to test when amino acid identity was ≥ 90% for all members within a family.

The suitability of the ∆79AvrRpt2 reporter required validation since rhizobia do not infect Arabidopsis. We selected NopB and NopJ from NGR234 (NGR_a00680 and NGR_a02610, respectively) for testing. NopB is known to be secreted in a flavonoid- and T3SS-dependent manner, and NopJ is a member of the YopJ/HopZ T3E family (Ausmees et al., 2004; Lorio et al., 2004). The ORFs of *nopB* and *nopJ* were cloned downstream of a constitutive promoter and as translational fusions to ∆*79avrRpt2*. The gene fusions were mobilized into the γ-proteobacterium *Pseudomonas syringae* pv tomato DC3000 (*Pto*DC3000) and its T3SS-deficient mutant, ∆*hrcC*. Each of the strains were infiltrated into leaves of Col-0 and examined for an HR approximately 20 hpi.

*Pto*DC3000 carrying the positive control, a fusion between the full-length

*avrRpm1* gene and ∆*79avrRpt2*, elicited a robust HR 20 hpi (Figure 4.2A). Although Col-0 can elicit ETI in response to both AvrRpm1 and AvrRpt2, the HR we observed is known to be a consequence of perception of the latter by RPS2 since AvrRpt2 "interferes" with AvrRpm1 when co-delivered (Dangl et al., 1992). *Pto*DC3000 carrying a full-length *avrRpt2* also consistently elicited robust HRs (data not shown). In contrast, *Pto*DC3000 lacking fusions to ∆*79avrRpt2* failed to elicit an HR but showed tissue collapse approximately 28 hpi, indicative of disease symptoms (data not shown). Importantly, *Pto*DC3000 carrying *nopB*::∆*79avrRpt2* and *nopJ*::∆*79avrRpt2* fusions elicited robust HRs within the same time frame as the positive controls (Figure 4.2A). The ∆*hrcC* mutant, regardless of the gene it carried, failed to elicit any phenotype throughout the course of the study thereby confirming the T3SS-dependent delivery of T3Es. This is the first demonstration of heterologous T3SS-dependent delivery of rhizobial T3Es and the first validation of NopB and NopJ as T3Es. These data also validated the use of *Pto*DC3000 for the rapid characterization of candidate T3Es for T3SS-dependent translocation.

A total of 162 genes were tested for the 103 families and from these, 72 T3Es belonging to 60 families, were confirmed for T3SS-dependent translocation (table 4.2; figure 4.3). Twenty-eight of the families were either represented in only one strain or tested multiple times because of within-family diversity that was less than 90% amino acid identity. The NopB T3E family is presented as an example (Figure 4.2B). The translated sequences of *nopB* of USDA207 and USDA257 are 100% identical to each other and have 98% identity to its member from NGR234.

Within *B. japonicum*, the *nopB* members had ≥ 99% identity in all pairwise comparisons but only 32% identity to its family members from *S. fredii*. Regardless, as shown, each of the tested *nopB*::Δ*79avrRpt2* gene fusions were sufficient for *Pto*DC3000 to trigger an HR at 20hpi (Figure 4.2B). In contrast, no response was observed in plants infected with the Δ*hrcC* mutant carrying members of the *nopB* family (data not shown).

In total, 90% (54/60) of the families, all members, inferred based on ≥ 90% amino acid identity or confirmed via functional testing, were classified as T3Es. We only observed conflicting translocation results for six T3E families (NopP, NopZ, NopBB, NopBE, NopBG and NopBH). An additional three families had pseudogenes as determined based on the presence of premature termination codons relative to other members. The members of the other 37 families had no evidence for T3SS-dependent translocation and were not considered T3E based on the criteria used herein.

Prior to this study, a total of 22 candidate T3E families had been previously identified based on flavonoid-induced expression and T3SS-dependent *in vitro* secretion and three proteins have been confirmed as *bona fide* T3Es based on *in vivo* translocation using fusions to cya (Supplemental Table 2). Of the 25 total candidate T3E families, we identified 22 of which 20 were validated as T3Es (Figure 4.4). The use of *Pto*DC3000 to heterologously delivery proteins from rhizobia failed to confirm the translocation of NopA and NopT, previously shown to be secreted in a type III-dependent manner (Figure 4.3; (Deakin et al., 2005; Dai et al., 2008)). However, NopA may in fact be a structural component of the

T3SS rather than a T3E (Deakin et al., 2005). NopT, in contrast, is a member of the YopT/AvrPphB family and likely a *bona fide* T3E. The cytotoxic effects of NopT in Arabidopsis, could have caused misleading conclusions in the translocation assay (Dai et al., 2008). We did not test NopC, NopH or NopD for translocation because they were not identified or did not pass our filters (Deakin et al., 2005; Rodrigues et al., 2007; Hempel et al., 2009).

In total, 40 new T3E families were identified and were assigned the names NopY through NopBT, for Nodulation Outer Proteins previously proposed, and based according to rules previously developed for naming T3E of pathogenic bacteria (Supplemental Table 4.2) (Marie et al., 2001; Lindeberg et al., 2005). For seventeen families that were previously assigned a Nop name and confirmed for delivery in this study, we retained their assigned name. Other than those previously identified, none of the products of the T3E-encoding genes identified in this study have sufficient homology to proteins of known function (data not shown).

**Genetic diversity of rhizobial type III effectors**

Surprisingly, indications up to this point suggested that T3E collections are highly similar in content within the *S. fredii* and *B. japonicum* species (Figure 4.3). This observation could potentially suggest that T3E collections of rhizobia are under a different selective pressure than T3E collections of plant pathogens (McCann and Guttman, 2008; O'Brien et al., 2010). To examine this possibility, we calculated and plotted non-synonymous versus synonymous substitution rates (Ka/Ks) for the 46 confirmed T3E families that have more than one family member

(Figure 4.5A). A total of 201 pairwise comparisons were done, with 246 of the possible comparisons from 34 families excluded because the nucleotide sequences were identical for both family members under comparison.

In support of the possibility that T3E genes of rhizobia are under purifying selection, the vast majority of pairs that could be examined had Ka/Ks values below 1 (Figure 4.5A). We emphasize again that an additional 246 comparisons were excluded because of sequence identity and the evidence for purifying evidence is thus underestimated. Only 10 pairwise comparisons within five T3E families had Ka/Ks values above 1, suggestive of positive selection. For comparisons, we also calculated Ka/Ks values for the nod/fix genes, which given their functions in nodulation and nitrogen fixation, are predicted to be under purifying selection, which was indeed the case (Figure 4.5B). We also calculated Ka/Ks values for all orthologous genes encoded in the rhizobial genomes to determine how T3E families compared. Clustering of all predicted amino acid sequences from the eight strains resulted in 35,521 clusters, representing 65,694 orthologous pairs. Since 22,494 clusters only had a single representative, 13,027 pairwise comparisons were calculated (Figure 4.5C). As shown, most orthologous pairs had evidence for purifying selection. The great majority of comparisons between species were saturated, as expected (Figure 4.5C, red).

NGR234 had three genes (*nopP*, *nopZ* and *nopBB*) with membership in T3E families but when characterized using *Pto*DC3000, had no evidence for T3SS-dependent translocation. However, none of the Ka/Ks values between pairwise comparisons with these three genes were greater than 1.0 (Figure 4.5a,

green triangles). Additionally, three pairwise comparisons between family members of *nopBG*, *nopBH* and *nopBE* from USDA207 and USDA257 included those that differed in translocation potential when carried in *Pto*DC3000. The *nopBG* and *nopBH* genes have calculated Ka/Ks values of 2.13 and 0.025, respectively (Figure 4.5a, green triangles). The Ka/Ks for *nopBE* could not be calculated because each of the six observed nucleotide differences between the two family members resulted in non-synonymous amino acid substitutions and therefore a Ka of zero. Relative to whole-genome comparisons, we observed 168 pairwise comparisons having Ka/Ks values greater than 1.0. In all, data suggest that T3E-encoding genes of *S. fredii* and *B. japonicum* are similar to the vast majority of genes in their genomes with evidence for purifying selection.

**Mosaic genomes of *S. fredii* and *B. japonicum***

In pathogens, collections of T3Es are dynamic, likely a consequence of gain via horizontal gene transfer (HGT) and loss via mutation in response to selective pressures imposed by plant defenses (McCann and Guttman, 2008). We therefore examined rhizobial genomes for signatures of HGT and their correspondence or lack thereof to T3E-encoding loci. However, because of the challenges in using disjointed draft genome sequences to study HGT, we focused primarily on the finished *B. japonicum* USDA110 and USDA6$^T$ genomes. Furthermore, since all T3E-encoding genes of NGR234 are located on a plasmid, we decided against characterizing HGT in *S. fredii*. As previously described, we found very little evidence for loss of function via mutations; only three T3E families had pseudogenes.

The majority of the T3E-encoding genes clustered in the 680 kb symbiosis island that was previously identified in USDA110 based on differences in GC content relative to the rest of the genome ((Göttfert et al., 2001; Kaneko et al., 2002); Figure 4.6). It was previously suggested that the symbiosis island is from pieces of different origins, but unlike the *M. loti*, there is no additional evidence for an integration event and is likely a very "ancient" island (Sullivan and Ronson, 1998; Göttfert et al., 2001; Kaneko et al., 2002; 2011). A similar clustering of T3E-encoding genes to this symbiosis island was also observed in the genome of USDA6$^{T}$ (data not shown). The remaining T3E-encoding genes outside the island did not appear to associate to regions with evidence for HGT. In contrast, genes with evidence for positive selection tended to localize to regions with evidence for HGT (Figure 4.6). Finally, most of the conserved T3E-encoding genes, regardless of their location relative to the symbiosis island, shared similar genomic context between strains, potentially indicating their acquisition/presence in a common ancestor (Figure 4.7). A 25 kb genomic region of NGR234 and HH103 encoding genes of seven T3E families are entirely syntenous, with only one neighboring gene showing any signs of variability (Figure 4.7A). Similarly, a large genomic region shared between USDA110 and USDA6$^{T}$ are syntenous (Figure 4.7B). The one observed exception is *nopP*, which while seemingly syntenous within each of the *Bradyrhizobium* and *Sinorhizobium* species, does not appear syntenous between the two. This may indicate that *nopP* was inherited after species divergence. It therefore appears that most, if not all T3E genes are "core" to the *S. fredii* and *B. japonicum* genomes.

## DISCUSSION

Rhizobia are nitrogen-fixing bacteria comprised of nearly 100 different species. Many of these form symbiotic relationships with legumes that are crucial to the nitrogen cycle. The establishment of these symbioses is complex, with many factors contributing to the specificity of the host-microbe relationship. To identify additional factors that may influence host specificity, we cataloged collections of type III effector genes from eight different strains within the *S. fredii* and *B. japonicum* species. These two species and their corresponding strains were selected based on phylogenetic diversity as well as demonstrable T3SS-dependent host changes. We combined the use of next generation sequencing to generate draft genome sequences and computational and experimental methods to identify T3E candidates and validate their T3SS-dependent delivery into plant cells, respectively. This genome-enabled study provides the first insights into the content of and selective pressures that shape collections of T3Es of mutualistic bacteria.

### Role of type III effectors in mutualism

The T3SS is suggested to function during the early stages of nodulation, when rhizobia are potentially most vulnerable to host defense (Mithöfer, 2002; Gage, 2004; Soto et al., 2009). The ability of plants to recognize rhizobia and initiate PTI has been demonstrated. For example, *Mesorhizobium loti* elicits a response in *Lotus japonicus* similar to that elicited by the PAMP flg22, although the response to *M. loti* is less robust (Lopez-Gomez et al., 2012). Furthermore, the simultaneous treatment of plants with flg22 and *M. loti* dramatically reduced

nodulation efficiency, while flg22 treatment in the later stages of symbiosis had no effect, suggesting that PTI can interfere early in the nodulation process. If T3Es are indeed mechanisms for countering defense, these data would suggest a need for the early deployment of T3Es by mutualistic rhizobia to dampen defense responses.

Evidence has, in fact, been accumulating that suggests T3Es of rhizobia function to perturb defense. Characterization of NopL and NopT of NGR234, show these T3Es affect the MAPK defense signaling cascade and cause cytotoxic effects in transgenic plants, respectively (Yang et al., 2009; Zhang et al., 2011). Furthermore, candidate T3Es of rhizobia have been previously identified based on membership in families of T3Es of pathogens (Dai et al., 2008; Kambara et al., 2009). In this study, no additional homologs of characterized T3E families were discovered. Collectively, this body of evidence leads us to hypothesize that *S. fredii* and *B. japonicum* use their T3SSs to translocate collections of T3Es to manipulate host cells and dampen host PTI as an early step of colonization.

This hypothesis, however, is difficult to reconcile in light of the repeated observations that the T3SS is not essential for rhizobial infection of all plants. T3SS-deficient mutants of rhizobia are compromised in nodulating hosts that are compatible with their corresponding wild type strain. However, their mutants frequently gain what were previously incompatible plant species as hosts (Marie et al., 2003; Hubber et al., 2004; Skorpil et al., 2005; Dai et al., 2008). These data, in addition to the observation that not all rhizobia species encode for T3SS, suggest that the T3SS has more of an accessory or host-specific role rather than

a necessary function for host infection. An alternative explanation is discussed in the following section.

The *ttsI*-regulon includes genes with roles beyond effectors and secretion that may contribute to counteracting host defenses. In NGR234, *ttsI* is also required for the synthesis of rhamnose-rich polysaccharides (Marie et al., 2004). Polysaccharides can function to release bacteria from infection threads, protect rhizobia against plant defenses, and have been suggested to dampen host defenses (Graham et al., 1977; Dow et al., 2000; Gao et al., 2001; Marie et al., 2004). Our goal was to identify T3E-encoding genes and we disregarded many of the identified *tts*-boxes because the translated sequences of downstream genes were not indicative of a T3E (Table 4.2). Further investigation of their corresponding genes may provide a more complete picture of processes that these two species of rhizobia rely on for infecting their hosts.

Our analyses of the two species in this study contributed T3E gene sequences from 60 different families. T3E gene discovery relied on a bioinformatics-based screen for *tts*-boxes. Overall, we find fewer *tts*-boxes than similar *hrp*- or *pip*-box searches in genome sequences of *P. syringae* or xanthomonads, respectively (Ferreira et al., 2006; Kimbrel et al., 2011). This may reflect the conservative approach we took in developing the training set for the HMM model. We only included *tts*-box sequences that corresponded to ORFs that encode products with evidence for T3SS-dependent secretion. Alternatively, the lower number of *tts*-boxes may reflect a relative lack of sequence divergence or a comparatively higher number of *ttsI*-induced genes that are encoded in operons.

The latter arrangement allows for fewer regulatory *cis*-elements while still enabling the regulation of a large number of genes.

This is the first demonstration that *P. syringae* pv. *tomato* DC3000, a γ-proteobacteria, can deliver T3Es of α-proteobacteria rhizobia (Figure 4.2). The functionality of *Pto*DC3000 as a heterologous system for characterizing T3Es is not unprecedented given that it has been a workhorse for similar studies to deliver effectors from *Xanthomonas* as well as effectors of oomycete pathogens (Mudgett et al., 2000; Sohn et al., 2007). Interestingly however, we could not identify common similar N-terminal translocation patterns in the T3Es of rhizobia and *P. syringae* (data not shown; (Petnicki-Ocwieja et al., 2002; Greenberg and Vinatzer, 2003)). The genes we classified as T3Es produced a clear and robust HR and included nearly all candidate T3Es previously identified by others (Figure 4.4; Supplemental table 4.2). However, given the limits of using the AvrRpt2-elicited HR as a marker for translocation and heterologous effector translocation, we cannot exclude the possibility of false positives and false negatives. We previously developed the EtHAn T3SS-encoding bacterial system for this study (Thomas et al., 2009). For reasons as yet undetermined, EtHAn did not appear to translocate rhizobial T3Es (data not shown). Finally, some of the predicted T3Es could not be tested (Table 4.2). Some predicted T3E genes were distributed across multiple contigs, and recalcitrant to PCR amplification, which we infer as a result of local misassembly mistakes. For example, the candidate T3E, bll8244 from USDA110, was found in all *B. japonicum* strains as well as USDA207 and USDA257, but the short repeating sequences in this gene made it difficult to

assemble (Pop and Salzberg, 2008). Despite these few limitations, this study nonetheless represents a significant step towards understanding the functions and diversity of T3Es in mutualistic bacteria. The prevalence of T3SS in rhizobia is increasingly evident with the discovery of a potential T3SS-encoding locus in the genome sequence of *Mesorhizobium amorphae* (Hao et al., 2012). Similarly, use of a BLASTN search in the draft genome sequence of *Bradyrhizobium* sp. ORS 285 identified homologs for 32 of the T3E gene families identified in this study (data not shown; NCBI accession PRJNA80837).

**ETI and host range**

The ability of T3SS-deficient mutants of rhizobia to overcome nodulation restriction is a possible explanation for the observation that they gain previously incompatible plants as hosts. Evasion of ETI, however, would likely come with a cost; mutants will be compromised in their ability to suppress PTI and be less efficient at establishing symbiotic interactions. The T3SS and its delivered T3Es are thus still implied to be necessary for full infection. Further, given that conservation can be interpreted as an indicator of functional importance, it is noteworthy that homology-based surveys of *B. japonicum* show the T3SS-encoding loci to be highly conserved across all soybean-associated strains (Mazurier et al., 2006).

*Rfg1* and *Rj2* are nodulation restriction genes of soybean that affect symbioses with *S. fredii* USDA257 and *B. japonicum* USDA122 (Yang et al., 2010). Molecular characterization showed that both genes encode for NB-LRR motifs characteristic of R proteins, pointing to rhizobial T3Es as negative host

range determinants through elicitation of ETI. Given the ability of *S. fredii* USDA207 to nodulate *Rfg1*-expressing plants, likely candidate Avrs are those polymorphic between the two *S. fredii* strains, i.e., NopBI, NopBJ, and NopBT. It is possible that USDA207 encodes additional T3Es that interfere with ETI, similar to what has been observed between *avrPphF* and *avrPphC* of *P. syringae* pv *phaseolicola* (Tsiamis et al., 2000). Whatever the case may be, our cataloging and comparisons of T3E collections within the *S. fredii* and *B. japonicum* strains provides a very short list of avr candidates for the molecular demonstration of ETI in legume-rhizobia interactions (Figure 4.3).

An *R* gene-mediated defense response that limits nodulation by restricting a presumably mutualistic symbiosis seems counter-intuitive. One possible explanation is that rhizobia vary in their nitrogen fixing ability and are not always beneficial. Plants may therefore rely on ETI as a mechanism for host-sanction since T3Es are detectable in plant nodules (Schechter et al., 2010; Wenzel et al., 2010). This is unlikely, however, as ETI-eliciting bacteria inevitably will compromise the nodulation ability of non-ETI-eliciting bacteria. For example, co-inoculation of wild-type USDA257 with its T3SS-deficient mutant negated the ability of the mutant to nodulate McCall soybean cultivar, suggesting that ETI will affect all infecting rhizobia (Meinhardt et al., 1993). In this study, the relatively large size of T3E collectives suggests redundancy in function, such as that observed in collections of T3Es of *P. syringae* (Table 4.2; (Baltrus et al., 2011)). Redundancy lends to rapid diversification such that T3Es can be easily jettisoned to avoid recognition at a minimal cost in the fitness of the microbe in its host. The

observed conservation of T3Es in rhizobia argues against an avoidance of ETI as a selective pressure.

Another possible explanation for restriction of nodulation is that ETI against rhizobia is simply due to convergence of pathogen and mutualist T3Es on defense proteins that are guarded by common R proteins. This so-called guard hypothesis predicts that R proteins elicit ETI upon perception of changes to a guarded host protein (Jones and Dangl, 2006). One classic example is the RPS2-AvrRpt2 system used in this study. AvrRpt2 is a cysteine protease that targets and cleaves the Arabidopsis protein, RIN4 (Axtell and Staskawicz, 2003; Mackey et al., 2003). The cleavage and subsequent degradation of RIN4 alleviates inhibition of RPS2 to elicit ETI. Therefore, ETI against rhizobia is thus potentially an unfortunate consequence of rhizobia having to target guarded host proteins. Therefore, the selective pressure for the host is to maintain resistance against pathogens, rather than against beneficial rhizobium. Supporting this is the observation that nodulation restriction loci, such as *Rfg1*, are often clustered with large families of NB-LRR-encoding genes and genetically linked to resistance specificities against agricultural pests (Graham et al., 2002b).

### Evolution of type III effectors

The T3E collections of *S. fredii* and *B. japonicum* are highly conserved in both content and sequence, with very little evidence of diversifying selection or acquisition via horizontal gene transfer (Figures 4.3, 4.5 and 4.6). Among the translocated T3E families of rhizobia, 12 of 20 and 31 of 45 T3E were “core” to the *S. fredii* and *B. japonicum* species. Furthermore, of these core T3Es, five

families (*nopY*, *nopB*, *nopU*, *nopAA* and *nopM1*) are distributed across both species (Figure 4.3). The observed evolutionary stasis of T3E collections indicates an evolutionary framework different than that proposed for T3Es of pathogens.

Pathogen virulence proteins have been modeled according to an "evolutionary arms race" with multiple studies of their T3Es showing collections to be rapidly evolving with dramatic variation in content and sequence (McCann and Guttman, 2008; Stavrinides et al., 2008; Baltrus et al., 2011). In this context, novelty potentiates evasion of detection by the host, e.g., loss of *avrs* as an avoidance of ETI (Sachs et al., 2011a). By contrast, novelty in symbiosis can result in instability. For this reason, mutualists may be under more pressure to limit diversification (Sachs et al., 2011b). In this mutualism context, hosts select for the most frequent genotype of microbe and the more common genotype of microbe is more likely to find a suitable host.

It could also be argued that the apparent conservation of T3E genes in rhizobia simply mirrors the low diversity of their genomes. However, in pathogens, in spite of high genome conservation, their T3E collections are rapidly diversifying. The genome diversity within the examined strains of *S. fredii* is far more extensive than that of most plant pathogens while the diversity within the examined strains of *B. japonicum* is similar to the observed diversity of plant pathogens (Figure 4.1A). Pairwise comparisons of pathovars of *P. syringae* have upwards of 85% orthology, while *Xanthomonas* is greater than 70%, and *P. fluorescens* >60% (Silby et al., 2009; Kimbrel et al., 2010; Baltrus et al., 2011;

Kimbrel et al., 2011). Even across 19 phylogenetically different *P. syringae* pathovars, ~60% of the genes can be considered “core” (Baltrus et al., 2011). Despite this high conservation, two different pairs of *P. syringae* strains that infect a common host, the closely-related tomato pathovars DC3000 and T1, and the distantly-related cucumber pathovars, *lachrymans* 106 and 107, share ~80% and ~70% orthology, respectively, yet have less than 50% conservation in their collections of T3Es (Almeida et al., 2009; Baltrus et al., 2011). Taken together, the conservation in T3E collection content, the low number of unique genes, and evidence for purifying selection argue against an evolutionary arms race for T3Es of mutualistic rhizobia. Instead, these data support the possibility of a “mutualistic environment”, where co-evolution of plant and rhizobia favor maintaining symbiosis (Law and Lewis, 2008; Sachs et al., 2011a).

Many of the T3E-encoding genes of rhizobia are potentially members of operons that include genes encoding structural components of the T3SS. Furthermore, the majority of the highly conserved T3E-encoding genes are located in the symbiosis island where the majority of *nod* and *fix* genes are also located. The conservation of the T3E-encoding genes may result from their association with functions essential to mutualism. However, a similar situation of association is found in *P. syringae* with a different outcome. The *hopM1* T3E-encoding gene is linked with the T3SS-encoding region but despite the family being represented in all 19 examined strains, its members have pronounced nucleotide and functional diversity, indicative of the evolutionary arms race (Baltrus et al., 2011). In contrast, even the rhizobial T3E-encoding genes distal to

the symbiosis islands appear as static as those within the island with little evidence for HGT or sequence divergence (Figure 4.6).

T3SS-encoding loci have also been detected in other mutualists such as insect endosymbionts *Buchnera aphidicola*, *Hamiltonella defensa*, *Sodalis glossinidius*, and SZPE, as well as commensal bacteria such as *Pseudomonas fluorescens* (Rainey, 1999; Shigenobu et al., 2000; Dale et al., 2001; Preston et al., 2001; Dale et al., 2002; Moran et al., 2005; Kimbrel et al., 2010). Loss-of-function mutants of the insect mutualist *S. glossinidius*, for example, are non-invasive and when microinjected directly into female flies, not transmitted to progeny (Dale et al., 2001). It will be interesting to determine whether the T3E collections of these mutualists exhibit similar patterns as observed in rhizobia.

It is becoming evident that the T3SS is an important mechanism for bacteria to establish symbioses with their hosts, regardless of the outcome of the interactions. However, the outcome may influence collections of T3Es, with pathogens in an “evolutionary arms race” and mutualistic rhizobia in an “evolutionary armistice” punctuated by infrequent “skirmishes” with their hosts.

## ACKNOWLEDGEMENTS

We would like to thank an outstanding group of undergraduate researchers for their assistance (alphabetical order): Andres Alvarez, Philip Hillebrand, Denise Hruda, Stanley Lee, Ryan Lilley, Meesha Pena, Liz Stoener, Jayme Stout, and David Swader-Hines. We would also like to thank Mark Dasenko and Chris Sullivan in the CGRB for high-throughput genome sequencing and data

preparation as well as the Soybean Genomics and Improvement Laboratory for the rhizobial strains. Finally, we thank Dr. Jeffery Dangl for his guidance, wonderful mentorship, and generosity in providing the space and resources to initiate this project. This work was supported by the National Research Initiative Competitive Grants Program Grant no. 2008-35600-04691 and the Agriculture Research Foundation.

## MATERIALS AND METHODS

### Bacterial strains and plasmids

Bacterial strains used in this study were: *B. japonicum* strains USDA6, USDA110, USDA122, USDA123, and USDA124; *S. fredii* strains USDA207 and USDA257; *Rhizobium* sp. NGR234, *Pseudomonas syringae* pv. *tomato* DC3000 (*Pto*DC3000), its T3SS-deficient mutant (Δ*hrcC*), and *Escherichia coli* DH5α. Rhizobia strains and *P. syringae* were grown in modified arabinose gluconate media (MAG) or King's B (KB) media, respectively, at 28°. *E. coli* DH5α was grown in Luria-Bertani (LB) media at 37°C. Antibiotics were used at the following concentrations: 50 µg/ml rifampicin (*Pto*DC3000), 30 µg/ml kanamycin (all bacterial strains), 50 µg/ml chloramphenicol (*B. japonicum* strains), and 25 µg/ml gentamycin (*E. coli*). Plasmids used in this work were pDONR207 (Invitrogen, Carlsbad, CA), the Gateway destination vector pDD62-Δ79AvrRpt2 (Mudgett et al., 2000), and the pRK2013 conjugation helper plasmid (Helinski, 1979).

### Genome sequencing

Genomic DNA was extracted from *B. japonicum* strains USDA6, USDA122, USDA123, and USDA124, and *S. fredii* strains NGR234, USDA207

and USDA257 using osmotic shock, followed by alkaline lysis and phenol-chloroform extraction. We prepared 5 µg of DNA from each strain according to the instructions provided by the manufacturer (Illumina, San Diego, CA). Libraries were sequenced using the paired-end cycle sequencing kit on the Illumina (Supplemental Table 4.1). Sequencing was done by the Center for Genome Research and Biocomputing Core Labs (CGRB; Oregon State University, Corvallis, OR).

**Genome assembly and annotation**

We used Velvet 0.7.55 to *de novo* assemble the genomes (Table 1; (Zerbino and Birney, 2008)). Multiple assembles, using different parameters, were produced for each genome and the highest quality assembly was identified using methods described previously (Kimbrel et al., 2010). The Mauve Aligner 2.3 (default settings) program was used to order the contigs greater than 1 kb in length into scaffolds based on a closely related reference genome sequence (Rissman et al., 2009). The reference sequence used for *B. japonicum* strains was USDA110, while the symbiotic plasmid of *Rhizobium* sp. NGR234 was used as a reference for the *S. fredii* strains (Freiberg et al., 1997; Kaneko et al., 2002). Genome assemblies were annotated using Xbase and open reading frames (ORFs) annotations were further refined using the NCBI conserved domain database (CDD; (Altschul et al., 1997; Lowe, 1997; Kurtz et al., 2004; Delcher et al., 2007; Lagesen et al., 2007)).

**Bioinformatic analyses**

We used HAL (default settings) to generate whole-genome phylogenies of the six draft genome sequences and representative finished genome sequences

(Robbertse et al., 2011); NGR234 (NC_000914, NC_012586 and NC_012587; (Freiberg et al., 1997; Schmeisser et al., 2009)), USDA110 (NC_004463; (Kaneko et al., 2002)) and *Azorhizobium caulinodans* ORS571 (NC_009937.1; (Lee et al., 2008)).

For visualization of genomic regions, blast atlases were generated using the Gview Server (Petkau et al., 2010). USDA207 and USDA257 were compared against pNGR234a (NC_000914), and USDA6, USDA122, USDA123 and USDA124 were compared against USDA110 (NC_004463). ORFs longer than 100bp from pNGR234a or USDA110 (Supplemental Figure 3, black vertical lines) were used as queries against each genome, and hits with greater than 80% identity and an e-value ≤ $1e^{-15}$ were mapped against the genome position of the query sequence. BLAST Atlases were obtained through the Gview Server using BLASTN with the following settings: e-value ≤ $1 \times 10^{-15}$, alignment length ≥ 100 bp, and percent identity ≥ 80% (Petkau et al., 2010). The Circos was plot was generated using the Circos Table Viewer (Krzywinski et al., 2009).

Non-synonymous (Ka) and synonymous (Ks) substitution rates were determined by first clustering all of the translated sequences using CD-HIT with (-c .75 -n 5 -T 6 -s 0.7) settings (Li and Godzik, 2006). We used *ad hoc* Perl shell scripts to call ClustalW2 to align the translated sequences, PAL2NAL to construct codon alignments from the ClustalW2 output, and KaKs_Calculator 1.2 with the 'YN' approximate method to calculate Ka/Ks values for all pairwise comparisons (Yang and Nielsen, 2000; Suyama et al., 2006; Zhang et al., 2006; Larkin et al., 2007).

The Sanger Institute's Alien Hunter (default settings) was used to analyze genome sequences for potential horizontal gene transfer (HGT) events. Alien Hunter uses Interpolated Variable Order Motifs (IVOMs) to identify compositional biases and localize boundaries of predicted HGT regions (Vernikos and Parkhill, 2006).

**T3E candidate discovery**

We used the sequences of 30 *tts*-boxes from genes previously identified as TtsI-regulated as input for HMMer to generate a Hidden Markov Model for rhizobial *tts*-boxes (HMM; (Eddy, 1998; Marie et al., 2004; Zehner et al., 2008)). An *ad hoc* Perl shell script was used to identify *tts*-boxes with a HMMer bit score of 5.0 or higher. We next identified ORFs on the same strand as the *tts*-box, either up to 10 kb downstream or until another ORF on the opposite strand was encountered. We used BLASTX (e-value ≤ 1 x $10^{-15}$) to filter out ORFs with translated sequences homologous to components of the T3SS, proteins encoded by organisms that lack a T3SS, or proteins with general housekeeping functions. We used BLASTN and sequences of candidate T3E-encoding genes to identify homologs from each of the eight genome sequences (e-value cutoff ≤1x$10^{-15}$).

T3Es were grouped into families based on BLASTP scores ≤ 1x$10^{-15}$ across ≥50% the length of the protein. When all members of a family had amino acid identity ≥90% as determined using clustalw a single representative family member was chosen for testing (Larkin et al., 2007). In families of < 90% amino acid identity, members representative of the diversity were tested.

**T3E candidate cloning**

Oligonucleotide primers were designed for candidate T3E-encoding genes

to include a partial B1 and B2 sequence to the top and bottom oligonucleotide primers, respectively (sequences available upon request; Gateway® system (Invitrogen, Carlsbad, CA; (Chang et al., 2005)). Gene-specific primers and 1 ul of genomic DNA were used in two-step PCR as previously described (Kimbrel et al., 2011). PCR products were cloned into pDONR207 using BP clonase according to the instructions of the manufacturer (Invitrogen, Carlsbad, CA). PCR products were cloned into the destination vector pDD62-Δ79AvrRpt2 using LR clonase according to the instructions of the manufacturer (Mudgett et al., 2000). Plasmids were transformed into *E. coli* DH5α cells and mated into *Pto*DC3000 or Δ*hrcC* via triparental mating.

### *In planta* assay

*Arabidopsis thaliana* Col-0 plants were grown in a controlled growth chamber environment (15-hour day at 22°C followed by 9-hour night at 20°C). *Pto*DC3000 cells were grown overnight in KB with appropriate antibiotics. Cells were washed and re-suspended in 10mM $MgCl_2$ at a final concentration of $OD_{600}$=0.1. These inocula were infiltrated into the abaxial side of leaves of ~6-week-old plants using 1-ml needle-less syringes. The hypersensitive response (HR) was scored approximately 20-24 hpi based on comparisons to *Pto*DC3000 carrying an empty vector or full-length *avrRpt2*, negative and positive controls, respectively. Disease symptoms were scored approximately 28 hpi. Experiments were replicated a minimum of three times.

**Table 4.1. Statistics for draft genome assemblies.**

| Strain* | Size (Mb)[†] | # of Contigs[‡] | N50 (kb)[§] | Predicted ORFs[ʃ] |
|---|---|---|---|---|
| *S. fredii* NGR234* | 6.9 | - | - | 6322 |
| *S. fredii* USDA207* | 6.5 | 291 | 27.8 | 6995 |
| *S. fredii* USDA257 | 7.0 | 384 | 26.7 | 6723 |
| *B. japonicum* USDA6* | 8.7 | 788 | 19.6 | 7960 |
| *B. japonicum* USDA110 | 9.1 | - | - | 8317 |
| *B. japonicum* USDA122 | 8.9 | 186 | 31 | 8107 |
| *B. japonicum* USDA123 | 9.1 | 815 | 21.6 | 8361 |
| *B. japonicum* USDA124 | 8.9 | 1285 | 19.3 | 7943 |

*The genome sequences of the (*) indicated strains were finished during various stages of this study. [†]Sizes were estimated based on Velvet calculated values and included contigs from chromosomes and plasmids. [‡]Only contigs greater than 1 kb in length were included; "-" = the finished genome sequences were used in this study. [§]The contigs were ranked according to length and the contigs necessary to represent 50% of the estimated total size of the genome were identified. The size in kilobase (kb) of the smallest contig of the subset necessary to represent 50% of the genome is presented. [ʃ]Draft genome sequences were annotated using XBase and the predicted numbers of ORFs larger than 150 bp in length are presented.

**Table 4.2: Statistics for genome mining for T3E-encoding genes.**

| Strain* | # identified *tts*-boxes[†] | # ORFs post filtering[‡] | # ORFs post homology search[§] | T3Es[ʃ] | Not tested[#] |
|---|---|---|---|---|---|
| *S. fredii* NGR234* | 13 | 24 | 24 | 12 | 2 |
| *S. fredii* USDA207* | 21 | 19 | 23 | 15 | 3 |
| *S. fredii* USDA257 | 24 | 21 | 27 | 14 | 3 |
| *B. japonicum* USDA6* | 46 | 41 | 58 | 32 | 9 |
| *B. japonicum* USDA110 | 52 | 51 | 66 | 36 | 10 |
| *B. japonicum* USDA122 | 50 | 40 | 61 | 30 | 10 |
| *B. japonicum* USDA123 | 47 | 39 | 60 | 36 | 7 |
| *B. japonicum* USDA124 | 52 | 42 | 62 | 33 | 7 |

*The genome sequences of the (*) indicated strains were finished during various stages of this study. [†]A trained Hidden Markov Model (HMM) was used to identify candidate *tts*-boxes; the number of sequences with bit-scores ≥ 5.0 are presented. [‡]ORFs within 10 kb and encoded on the same strand as the predicted *tts*-box were identified and filtered based on results from BLASTX searches (e-value ≤ 1 x $10^{-15}$). [§]The number of candidates after using BLASTN searches to identify additional homologies (e-value ≤ 1 x $10^{-15}$). [ʃ]T3E-encoding genes based on T3SS-dependent elicitation of HR by *Pto*DC3000 in *Arabidopsis thaliana* Col-0. [#]The number of candidate T3E-encoding genes that were recalcitrant to cloning.

**Supplemental Table 4.1. Detailed statistics for genome sequencing.**

| Strain* | Sequencing technology† | # of channels used | Read length | Total # of PE reads |
|---|---|---|---|---|
| *S. fredii* USDA207* | GAI | 2 | 32mer | 22,136,363 |
| *S. fredii* USDA257 | GAI | 2 | 32mer | 27,603,487 |
| *B. japonicum* USDA6* | GAI | 3 | 32mer | 20,995,355 |
| *B. japonicum* USDA122 | GAII | 0.5 | 72mer | 19,430,723 |
| *B. japonicum* USDA123 | GAI | 3 | 32mer | 27,393,134 |
| *B. japonicum* USDA124 | GAI | 3 | 32mer | 14,219,973 |

*The genome sequences of the (*) indicated strains were finished during various stages of this study. †All genomes were sequenced using an Illumina

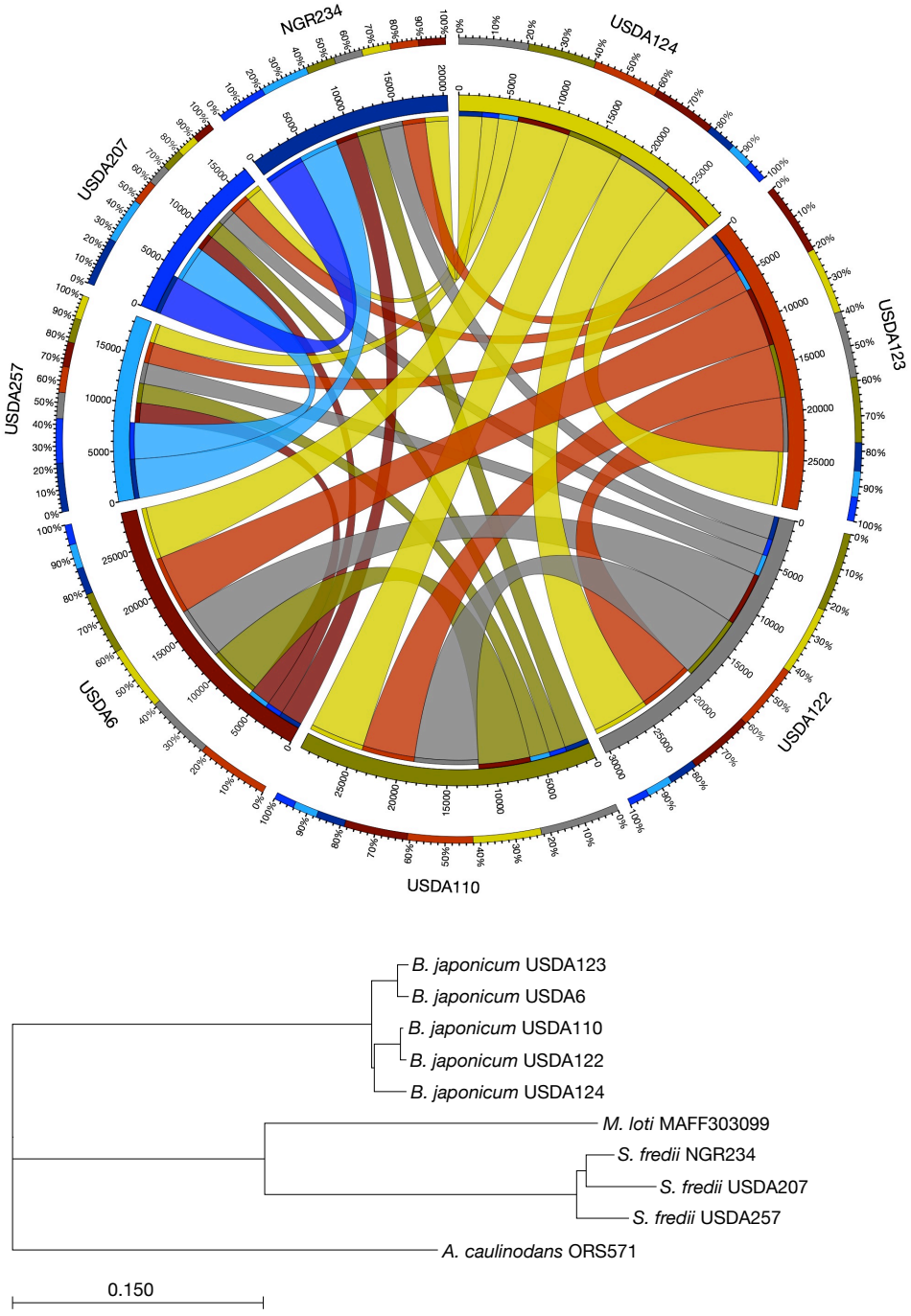



**Figure 4.1. Within and between genetic diversity of *S. fredii* and *B. japonicum* strains.**

Circos visualization of rhizobial genome sequences (A). The most outer track, based on the length of the different colored bars, represents the amount of

orthology other genomes have to the indicated genome. Genomes are ranked according to highest percent orthology (starting close to 0%) to lowest (ending closer to 100%). The inner track represents the amount of orthology the indicated genome has to the other genomes with interior ribbons connecting genomes, and variation in width depicting the extent of orthology. Genomes were assigned arbitrary colors (in a counter clockwise direction: NGR234 (dark blue); USDA207 (blue); USDA257 (cyan); USDA6 (maroon); USDA110 (olive); USDA122 (gray); USDA123 (orange); and USDA124 (yellow). Orthology was determined using reciprocal BLASTP of translated sequences between the eight rhizobia strains. Neighbor-joining phylogenomic tree of representative strains of rhizobia (B). The tree is based on a super alignment of 624 translated amino acid sequences. *A. caulinodans* ORS571 was included as the out group. Bootstrap values for each node are 100 (not shown); scale bar represents the number of substitutions per site.

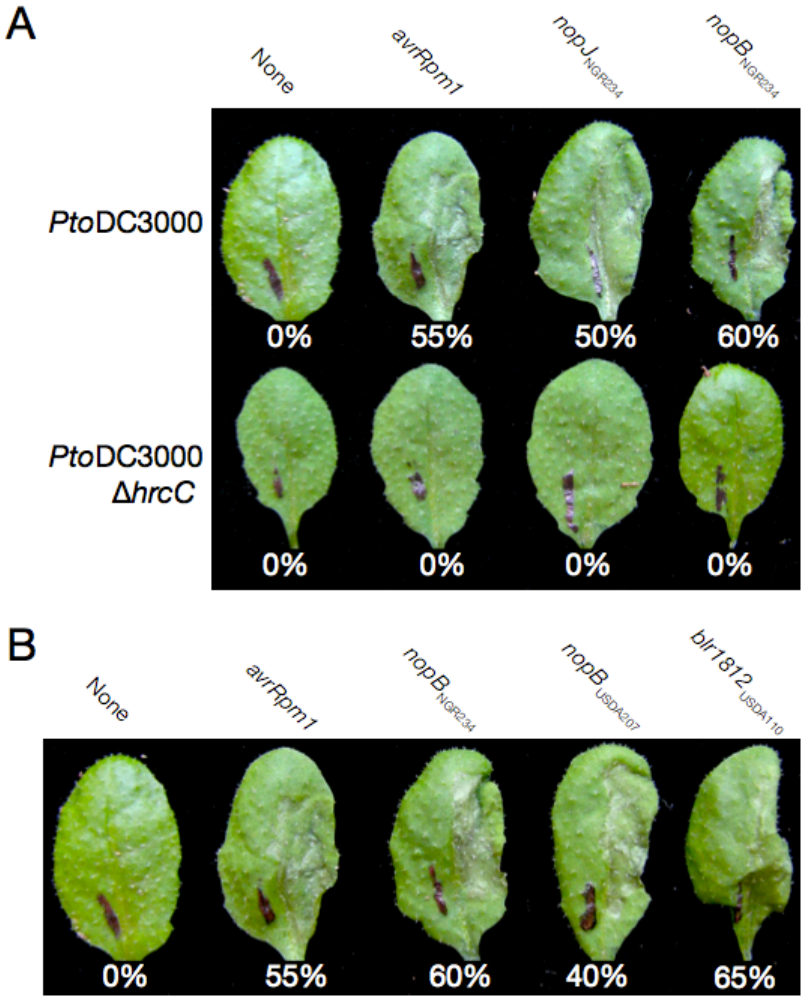



**Figure 4.2. *Pto*DC3000 delivers T3Es of rhizobia in a T3SS-dependent manner.**

Leaves of Arabidopsis Col-0 (*Rps2*/*Rps2*) were infiltrated with *Pto*DC3000 (top row) and its T3SS-deficient mutant, *hrcC* (bottom row) carrying no gene fusion to Δ79*avrRpt2* or gene fusions to *P. syringae* T3E gene *avrRpm1* or NGR234 candidate T3E genes, *nopJ* or *nopB* (A). Members of the NopB T3E gene family all encode for functional T3Es (B). Leaves of Arabidopsis Col-0 (*Rps2*/*Rps2*) were infiltrated with *Pto*DC3000 carrying no gene fusion to Δ79*avrRpt2* or gene fusions to *P. syringae* T3E gene *avrRpm1* or *nopB* genes from NGR234, USDA207, or USDA110. Leaves did not respond to infiltrations of *hrcC* (data not shown). In all experiments, leaves were scored for the HR ~20 hpi and the percent of responding leaves are presented (at least 20 leaves infiltrated). Experiments were repeated at least three times.

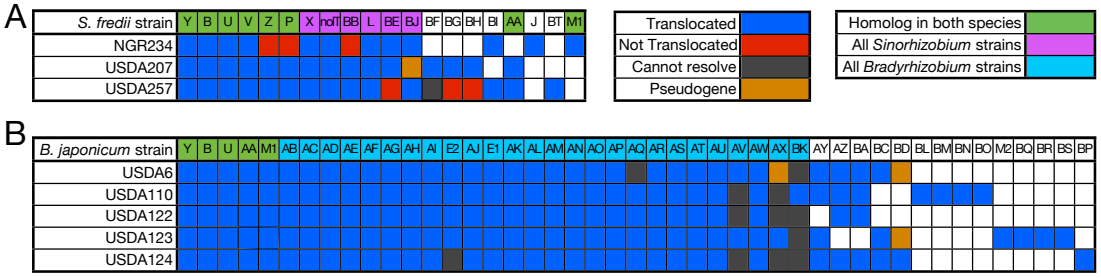



**Figure 4.3. Distribution and conservation of T3E families in rhizobia.**

The T3Es are listed across the top with strains of *S. fredii* (A) and *B. japonicum* (B) listed down the side. The conserved T3E-encoding genes are color-coded: between species (green) and within all three *S. fredii* (purple) or all five *B. japonicum* strains (cyan). The boxes are color-coded according to: functional T3E-encoding genes (blue); no evidence for T3SS-dependent delivery (red); homolog present but sequence could not be resolved (gray); homolog present with premature termination codon relative to other family members (brown); no detectable homolog (white).

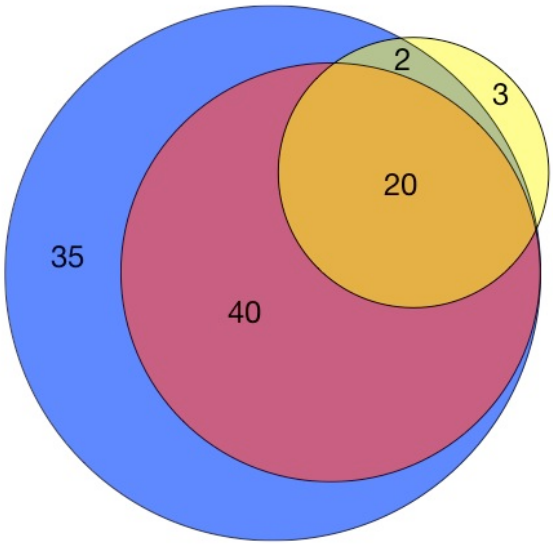



**Figure 4.4. Area-proportional Venn diagram of candidate and confirmed rhizobial T3Es.**
The yellow circle represents the union of 22 families of secreted and three families of delivered T3Es, previously discovered from rhizobia. Red and blue circles represent the 97 and 60 predicted and confirmed T3E families, respectively, identified in this study.

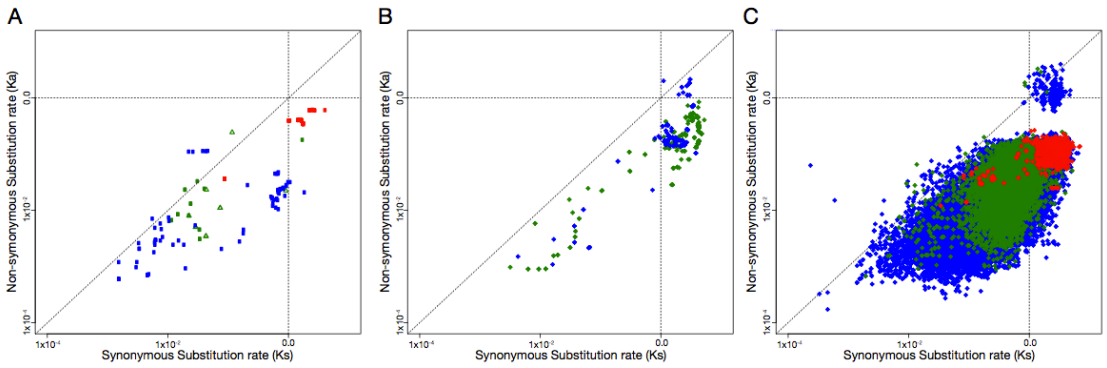



**Figure 4.5. The majority of rhizobia T3Es have low Ka/Ks scores.**

Synonymous (Ks) and non-synonymous (Ka) rates were plotted along the x- and y-axes, respectively, for all possible pairwise comparisons of rhizobia T3E-encoding genes (A), nod/fix genes (blue and green, respectively) (B), and all genes (C). Green, blue and red data points represent pairwise comparisons of genes within S*. fredii, B. japonicum*, or between species, respectively. For panel (A), squares and triangles represent family members that had both or only one member translocated, respectively. The dotted diagonal line indicates a ratio of 1. The dotted vertical and horizontal lines identify the boundaries for saturation of Ks and Ka, respectively. Comparisons between genes with identical nucleotide sequences or genes encoding potential pseudogenes were excluded from analyses.

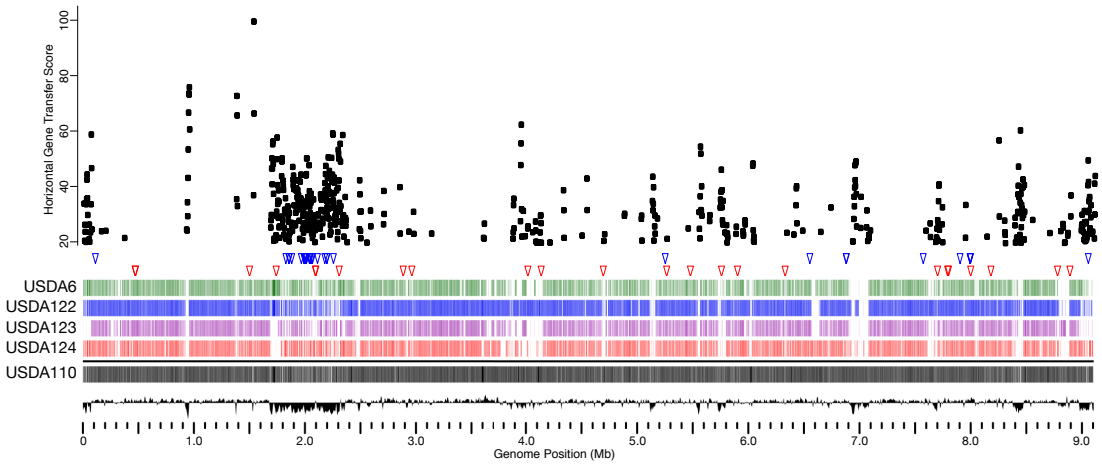



**Figure 4.6. Analysis of *B. japonicum* genomes for evidence of HGT events.**

(Top) The HGT scores as determined by Alien Hunter, are shown along the y-axis, only scores >20 are shown. The location of genes confirmed to encode T3E and those with Ka/Ks ratios > 1 of USDA110 are mapped according to the genome coordinates of USDA110 (blue and red triangles, respectively). (Bottom) BLAST Atlas showing orthology of *B. japonicum* genes > 100 bp in length (vertical lines) and plotted according to the genome coordinates of USDA110. Orthology was determined using BLASTN (e-value ≤ $1x10^{15}$; > 80% nucleotide identity). Bottom histogram depicts average GC% along a sliding window of 10 kb. The genome coordinates for USDA110 are presented along the x-axis in 100 kb increments; major ticks indicate megabase (Mb) increments.

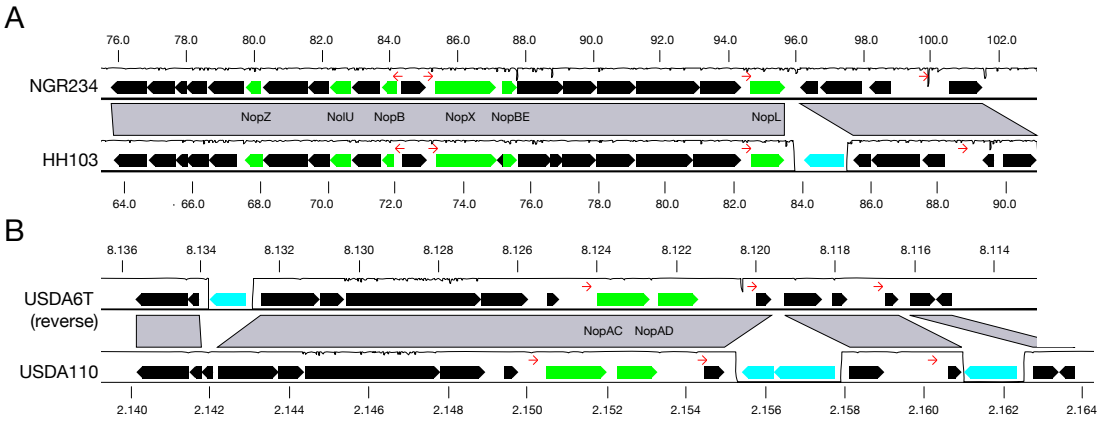
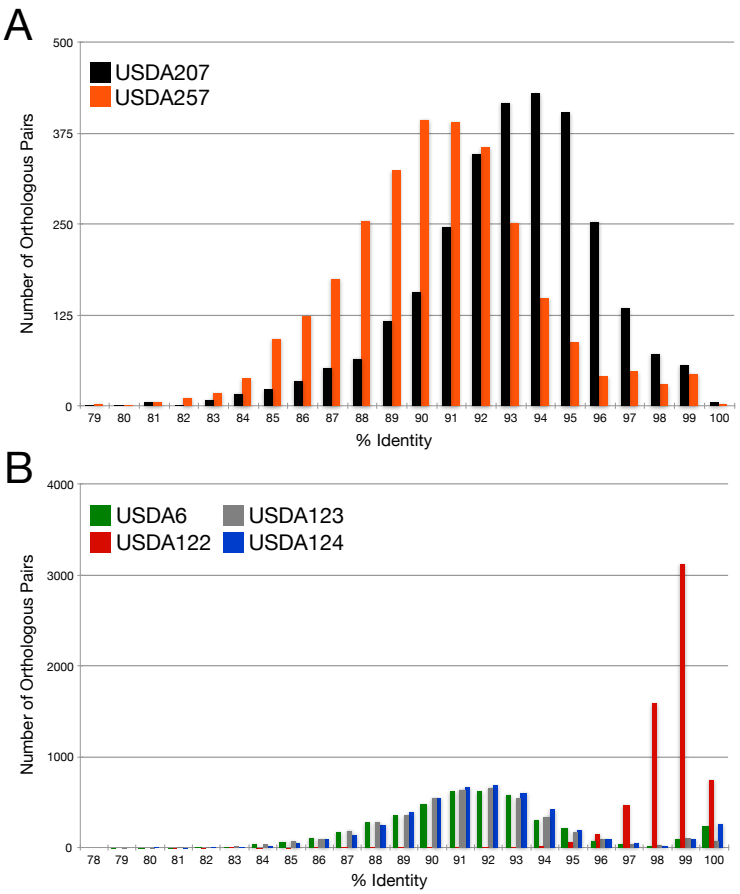



**Figure 4.7. Synteny of representative conserved T3E-encoding genes.**

A 25 kb region syntenous between pNGR234a and pSHH103d (A) and between USDA110 and USDA6$^{T}$ (B). Syntenous blocks are highlighted as gray blocks with traces representing the amount of synteny as determined using Mauve. Thick arrows represent ORFs with direction indicating expression from the leading or lagging strand. Labeled and thick colored arrows depict candidate ORFs identified based on location relative to predicted *tts*-box (thin red arrows). Thick green arrows encode confirmed T3E. Thick black arrows represent syntenous ORFs, and thick cyan arrows represent non-syntenous ORFs. Numbers at the top and bottom indicate genome coordinates in kb or Mb for *S. fredii* and *B. japonicum*, respectively.

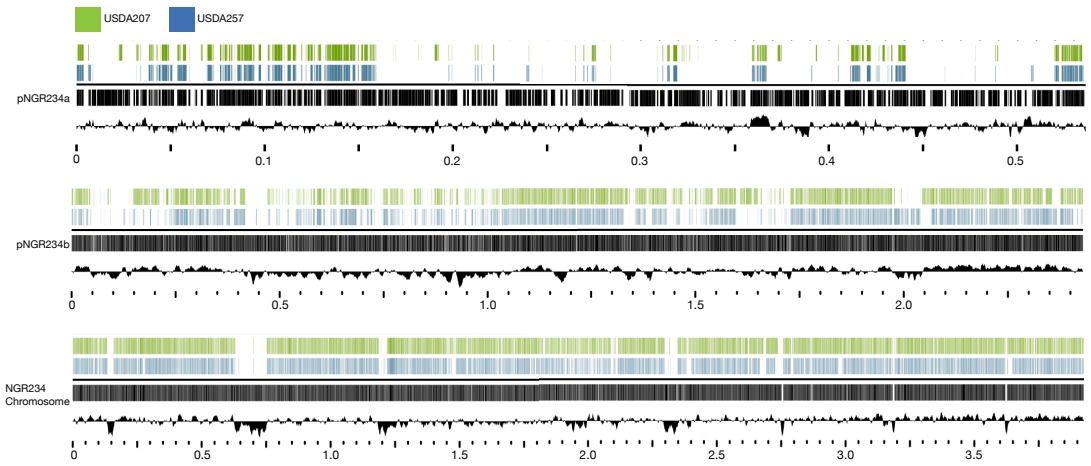

**Supplemental Figure 4.1. Screenshot of Mauve alignment of T3SS-encoding loci of the eight strains of rhizobia.**

Illumina short reads were *de novo* assembled and contigs were ordered using Mauve Aligner and NGR234 or USDA110 as a reference sequence (Freiberg et al,. 1997; Kaneko et al., 2002). The T3SS-encoding loci were aligned using the outmost located T3SS-associated gene, NGR_a00520 (y4yS) as the landmark (thick blue line). Large blocks represent syntenous regions and are further indicated by the thin long connecting vertical lines. The extent of synteny within each block is depicted by the height of the colored regions. Colors are arbitrarily assigned by Mauve Aligner. Shorter red lines show contig breaks. Coding sequences are depicted by open boxes for NGR234 and USDA110. Coordinates are for each of the regions depicted (bp).




**Supplemental Figure 4.2. Histogram of orthologs based on percent nucleotide identity.**

Orthologs between *S. fredii* (A) and *B. japonicum* (B) were identified using BLASTN (e-value ≤ $1x10^{30}$; > 90% the length of the coding sequence). For *S. fredii* ORFs from USDA207 and USDA257 were compared to the annotated ORFs in pNGR234a, pNGR234b and NGR234 chromosome (Freiberg et al., 1997; Schmeisser et al., 2009). For *B. japonicum*, the ORFs from the draft genomes of USDA6, USDA122, USDA123 and USDA124 were compared to the annotated ORFs of USDA110 (Kaneko et al., 2002). The nucleotide identity was determined for each orthologous pair, binned based on percent identity, and the numbers of each pair (y-axis) were plotted within each bin (x-axis).




**Supplemental Figure 4.3. BLAST Atlas of *S. fredii* genes.**

Orthologs of > 100 bp in length (vertical lines) were determined using BLASTN (e-value ≤ $1x10^{15}$; > 80% nucleotide identity) and plotted according to the pNGR234a, pNGR234b, and chromosome coordinates (along the bottom). USDA207, USDA257, and NGR234 ORFs are represented by green, blue, and black, respectively, vertical lines. Histograms depict average GC% along a sliding window of 10 kb. Coordinates in Mb increments are presented

# Conclusions and Future Directions

Jeffrey A. Kimbrel and Jeff H. Chang

Many species of plant-associated bacteria utilize a type III secretion system (T3SS) to inject collections of type III effector proteins (T3Es) into host cells to modulate defense. The T3SS was once considered exclusive to pathogens, but the identification of T3SS-encoding loci and demonstrable use by strains of non-pathogenic species such as the commensal *Pseudomonas fluorescens* and mutualistic *Sinorhizobium, Bradyrhizobium and Mesorhizobium* as well as those that interact with non-plant hosts, has instigated a conceptual shift for the role of the T3SS. In many of these cases, the T3SS is functional such that loss-of-function mutations compromise the abilities of the microbe to associate with their hosts. Undoubtedly the T3SS, regardless of the outcome of the host-microbe interaction, functions to deliver T3Es to manipulate the host cell.

The work presented in this thesis contributed significantly to increase the inventory and diversity of T3Es and challenged the dogma of T3E evolution. I presented the first genome sequence from a representative strain of a poorly characterized clade of xanthomonads (chapter 2). From this work, we identified a T3SS-encoding cluster and many candidate T3Es. Results from this work have direct application to agriculture for molecular diagnostics of crop plants and logical avenues to pursue for the development of resistant crop lines. Further, we identified new candidate T3E families for a commensal *P. fluorescens* and provided genomic resources to facilitate the development of methods useful for sustainable agriculture practices (chapter 3). Finally, we developed and executed a genome-enabled search for T3E-encoding genes from two species of mutualistic rhizobia (chapter 4). The pathogen-centric based investigations of

T3Es have led to the dogma that their collections are under diversifying selection as a consequence of the co-evolutionary arms race. Work presented herein offers an alternative view in which T3E collections of mutualists are seemingly static, with little diversity in content and sequence variation, likely a consequence of the mutualistic environment.

The next challenges are to characterize effectors for their molecular functions, to identify their host targets, and to understand how the functions of entire collections of T3Es are coordinated to benefit the microbe.

We sequenced the genome of *Xanthomonas hortorum* pv. *carotae* to discover molecular markers to detect *Xhc* contamination in carrot seed lots. With this work, we have shown that draft genome sequencing is sufficient to develop molecular markers for the rapid and specific typing of important agricultural pathogens. Additionally, we identified several novel T3Es, as well as genes encoding for two avirulence proteins, *xopQ* and *avrBs2*. This finding could be used to develop new control measures in an effort to reduce epiphytic populations of *Xhc* on carrots.

We sequenced the genome of *P. fluorescens* WH6, and subsequently identified one of the most complete T3SS-encoding loci of any *P. fluorescens* isolates. WH6 is a commensal organism used for biocontrol against important agricultural pests, both grassy weeds, and *Erwinia amylovora*, the causal agent of fire blight. As far as we know, the T3SS is unrelated to either of these biocontrol applications. Another *P. fluorescens* isolate, KD, has been shown to control the plant pathogen oomycete *Phythium ultimum* on cucumber in a T3SS-dependent

manner. A similar role for the WH6 T3SS in interacting with other organisms is at best speculative, nevertheless its characterization presents exciting opportunities. We identified over 30 new T3E families from several *Sinorhizobium fredii* and *Bradyrhizobium japonicum* strains. There is high conservation of the T3E families among rhizobia species, indicating these species have reached an evolutionary stasis and effective collection of T3Es to modulate host defense. The potential future directions from this ambitious study are significant and far-reaching. First, the T3SS of rhizobia has both a positive and negative influence on host range, and additional knowledge of T3Es will help in identifying how they trigger ETI in host legumes. Second, learning how rhizobia are able to modulate host defense is critical to understanding this unique relationship that has co-evolved. Third, commercial rhizobia inoculants have difficulty competing with indigenous rhizobia, and it is therefore critical to develop rhizobia inoculants with a competitive advantage over indigenous strains less efficient at fixing nitrogen.

# Bibliography

Abu Khweek, A., Fetherston, J.D., and Perry, R.D. (2010). Analysis of HmsH and its role in plague biofilm formation. Microbiology *156*, 1424–1438.

Alfano, J.R., Bauer, D.W., Milos, T.M., and Collmer, A. (2003). Analysis of the role of the Pseudomonas syringae pv. syringae HrpZ harpin in elicitation of the hypersensitive response in tobacco using functionally non-polar hrpZ deletion mutations, truncated HrpZ fragments, and hrmA mutations. Mol Microbiol *19*, 715–728.

Alfano, J.R., Charkowski, A.O., Deng, W.-L., Badel, J.L., Petnicki-Ocwieja, T., van Dijk, K., and Collmer, A. (2000). The Pseudomonas syringae Hrp pathogenicity island has a tripartite mosaic structure composed of a cluster of type III secretion genes bounded by exchangeable effector and conserved effector loci that contribute to parasitic fitness and pathogenicity in plants. Proc Natl Acad Sci USA *97*, 4856–4861.

Almeida, N.F., Yan, S., Lindeberg, M., Studholme, D.J., Schneider, D.J., Condon, B., Liu, H., Viana, C.J., Warren, A., Evans, C., et al. (2009). A draft genome sequence of Pseudomonas syringae pv. tomato T1 reveals a type III effector repertoire significantly divergent from that of Pseudomonas syringae pv. tomato DC3000. Mol Plant Microbe Interact *22*, 52–62.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389–3402.

Alvarez-Martinez, C.E., and Christie, P.J. (2009). Biological diversity of prokaryotic type IV secretion systems. Microbiol Mol Biol Rev *73*, 775–808.

Annapurna, K., and Krishnan, H.B. (2003). Molecular aspects of soybean cultivar-specific nodulation by Sinorhizobium fredii USDA257. Indian J. Exp. Biol. *41*, 1114–1123.

Arakawa, K., and Tomita, M. (2007). The GC skew index: a measure of genomic compositional asymmetry and the degree of replicational selection. Evol. Bioinform. Online *3*, 159–168.

Armstrong, D.J., Azevedo, M., Mills, D., Bailey, B., Russell, B., Groenig, A., Halgren, A., Banowetz, G.M., and Mcphail, K. (2009). Germination-Arrest Factor (GAF): 3. Determination that the herbicidal activity of GAF is associated with a ninhydrin-reactive compound and counteracted by selected amino acids. Biological Control *51*, 181–190.

Arrebola, E., Cazorla, F.M., Codina, J.C., Gutiérrez-Barranquero, J.A., Pérez-García, A., and de Vicente, A. (2010). Contribution of mangotoxin to the virulence and epiphytic fitness of Pseudomonas syringae pv. syringae. Int Microbiol *12*, 87–95.

Aslam, S.N., Newman, M.-A., Erbs, G., Morrissey, K.L., Chinchilla, D., Boller, T., Jensen, T.T., De Castro, C., Ierano, T., Molinaro, A., et al. (2008). Bacterial Polysaccharides Suppress Induced Innate Immunity by Calcium Chelation. Current Biology *18*, 1078–1083.

Augustin, D.K., Song, Y., Baek, M.S., Sawa, Y., Singh, G., Taylor, B., Rubio-Mills, A., Flanagan, J.L., Wiener-Kronish, J.P., and Lynch, S.V. (2007). Presence or absence of lipopolysaccharide O antigens affects type III secretion by Pseudomonas aeruginosa. J Bacteriol *189*, 2203–2209.

Ausmees, N., Kobayashi, H., Deakin, W.J., Marie, C., Krishnan, H.B., Broughton, W.J., and Perret, X. (2004). Characterization of NopP, a type III secreted effector of *Rhizobium* sp. strain NGR234. J Bacteriol *186*, 4774–4780.

Axtell, M.J., and Staskawicz, B.J. (2003). Initiation of RPS2-specified disease resistance in Arabidopsis is coupled to the AvrRpt2-directed elimination of RIN4. Cell *112*, 369–377.

Baltrus, D.A., Nishimura, M.T., Romanchuk, A., Chang, J.H., Mukhtar, M.S., Cherkis, K., Roach, J., Grant, S.R., Jones, C.D., and Dangl, J.L. (2011). Dynamic Evolution of Pathogenicity Revealed by Sequencing and Comparative Genomics of 19 Pseudomonas syringae Isolates. PLoS Pathog *7*, e1002132.

Banowetz, G.M., Azevedo, M.D., Armstrong, D.J., and Mills, D.I. (2009). Germination arrest factor (GAF): Part 2. Physical and chemical properties of a novel, naturally occurring herbicide produced by Pseudomonas fluorescens strain WH6. Biological Control *50*, 103–110.

Banowetz, G.M., Azevedo, M.D., Armstrong, D.J., Halgren, A.B., and Mills, D.I. (2008). Germination-Arrest Factor (GAF): Biological properties of a novel, naturally-occurring herbicide produced by selected isolates of rhizosphere bacteria. Biological Control *46*, 380–390.

Baudouin, E., Pieuchot, L., Engler, G., Pauly, N., and Puppo, A. (2006). Nitric oxide is formed in Medicago truncatula-Sinorhizobium meliloti functional nodules. Mol Plant Microbe Interact *19*, 970–975.

Bellato, C., Krishnan, H.B., Cubo, T., Temprano, F., and Pueppke, S.G. (1997). The soybean cultivar specificity gene nolX is present, expressed in a nodD-dependent manner, and of symbiotic significance in cultivar-nonspecific strains of Rhizobium (Sinorhizobium) fredii. Microbiology (Reading, Engl) *143 ( Pt 4)*, 1381–1388.

Bellato, C.M., Balatti, P.A., Pueppke, S.G., and Krishnan, H.B. (1996). Proteins from cells of Rhizobium fredii bind to DNA sequences precedingnolX, a flavonoid-inducible nod gene that is not associated with a nod box. Mol Plant Microbe Interact *9*, 457–463.

Bingle, L.E., Bailey, C.M., and Pallen, M.J. (2008). Type VI secretion: a beginner's guide. Curr Opin Microbiol *11*, 3–8.

Blanvillain, S., Meyer, D., Boulanger, A., Lautier, M., Guynet, C., Denancé, N., Vasse, J., Lauber, E., and Arlat, M. (2007). Plant carbohydrate scavenging through tonB-dependent receptors: a feature shared by phytopathogenic and aquatic bacteria. PLoS ONE *2*, e224.

Blom, J., Albaum, S.P., Doppmeier, D., Pühler, A., Vorhölter, F.-J., Zakrzewski, M., and Goesmann, A. (2009). EDGAR: a software framework for the comparative analysis of prokaryotic genomes. BMC Bioinformatics *10*, 154.

Bobik, C., Meilhoc, E., and Batut, J. (2006). FixJ: a major regulator of the oxygen limitation response and late symbiotic functions of Sinorhizobium meliloti. J Bacteriol *188*, 4890–4902.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. Science *326*, 1509–1512.

Bogdanove, A.J., Beer, S.V., Bonas, U., Boucher, C.A., Collmer, A., Coplin, D.L., Cornelis, G.R., Huang, H.-C., Hutcheson, S.W., Panopoulos, N.J., et al. (1996). Unified nomenclature for broadly conserved hrp genes of phytopathogenic bacteria. Mol Microbiol *20*, 681–683.

Boller, T. (2005). Peptide signalling in plant development and self/non-self perception. Curr. Opin. Cell Biol. *17*, 116–122.

Boller, T., and He, S.Y. (2009). Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. Science *324*, 742–744.

Bowen, D., Blackburn, M., Rocheleau, T.A., Andreev, O., and Golubeva, E. (1999). Insecticidal toxins from the bacterium Photorhabdus luminescens: gene cloning and toxin histopathology. Pesticide Science *55*, 666–668.

Brewin, N.J. (2004). Plant Cell Wall Remodelling in the Rhizobium–Legume Symbiosis. Critical Reviews in Plant Sciences *23*, 293–316.

Bronstein, P.A., Marrichi, M., Cartinhour, S., Schneider, D.J., and DeLisa, M.P. (2005). Identification of a twin-arginine translocation system in Pseudomonas syringae pv. tomato DC3000 and its contribution to pathogenicity and fitness. J

Bacteriol *187*, 8450.

Buell, C.R., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I.T., Gwinn, M.L., Dodson, R.J., Deboy, R.T., Durkin, A.S., Kolonay, J.F., et al. (2003). The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000. Proc Natl Acad Sci USA *100*, 10181–10186.

Burger, M., Woods, R.G., McCarthy, C., and Beacham, I.R. (2000). Temperature regulation of protease in Pseudomonas fluorescens LS107d2 by an ECF sigma factor and a transmembrane activator. Microbiology (Reading, Engl) *146 Pt 12*, 3149–3155.

Büttner, D., and Bonas, U. (2009). Regulation and secretion of Xanthomonas virulence factors. FEMS Microbiol Rev.

C B Harley, R.P.R. (1987). Analysis of E. coli promoter sequences. Nucleic Acids Res *15*, 2343.

CABI (2010). *Crop Protection Compendium.* Wallingford, Oxfordshire: CAB International.

Caldelari, I., Mann, S., Crooks, C., and Palmer, T. (2006). The Tat pathway of the plant pathogen Pseudomonas syringae is required for optimal virulence. Mol Plant Microbe Interact *19*, 200–212.

Cambronne, E.D., and Roy, C.R. (2006). Recognition and delivery of effector proteins into eukaryotic cells by bacterial secretion systems. Traffic *7*, 929–939.

Carver, T., Thomson, N., Bleasby, A., Berriman, M., and Parkhill, J. (2009). DNAPlotter: circular and linear interactive genome visualization. Bioinformatics *25*, 119–120.

Casper-Lindley, C., Dahlbeck, D., Clark, E.T., and Staskawicz, B.J. (2002). Direct biochemical evidence for type III secretion-dependent translocation of the AvrBs2 effector protein into plant cells. Proc Natl Acad Sci USA *99*, 8336–8341.

Chain, P.S.G., Grafham, D.V., Fulton, R.S., Fitzgerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C., et al. (2009). Genomics. Genome project standards in a new era of sequencing. Science *326*, 236–237.

Chang, J.H., Urbach, J.M., Law, T.F., Arnold, L.W., Hu, A., Gombar, S., Grant, S.R., Ausubel, F.M., and Dangl, J.L. (2005). A high-throughput, near-saturating screen for type III effector genes from Pseudomonas syringae. Proc Natl Acad Sci USA *102*, 2549–2554.

Chen, C., and Beattie, G.A. (2007). Characterization of the osmoprotectant

transporter OpuC from Pseudomonas syringae and demonstration that cystathionine-beta-synthase domains are required for its osmoregulatory function. J Bacteriol *189*, 6901–6912.

Chiliang Chen, G.A.B. (2008). Pseudomonas syringae BetT Is a Low-Affinity Choline Transporter That Is Responsible for Superior Osmoprotection by Choline over Glycine Betaine. J Bacteriol *190*, 2717.

Chisholm, S.T., Dahlbeck, D., Krishnamurthy, N., Day, B., Sjolander, K., and Staskawicz, B.J. (2005). Molecular characterization of proteolytic cleavage sites of the Pseudomonas syringae effector AvrRpt2. Proc Natl Acad Sci USA *102*, 2087–2092.

Cooper, J.E. (2004). Multiple responses of rhizobia to flavonoids during legume root infection (Elsevier).

Cui, H., Xiang, T., and Zhou, J.-M. (2009). Plant immunity: a lesson from pathogenic bacterial effector proteins. Cell Microbiol *11*, 1453–1461.

Cunnac, S., Chakravarthy, S., Kvitko, B.H., Russell, A.B., Martin, G.B., and Collmer, A. (2011). Genetic disassembly and combinatorial reassembly identify a minimal functional repertoire of type III effectors in Pseudomonas syringae. Proc Natl Acad Sci USA *108*, 2975–2980.

Cunnac, S., Lindeberg, M., and Collmer, A. (2009). Pseudomonas syringae type III secretion system effectors: repertoires in search of functions. Curr Opin Microbiol *12*, 53–60.

Cunnac, S., Occhialini, A., Barberis, P., Boucher, C., and Genin, S. (2004). Inventory and functional analysis of the large Hrp regulon in Ralstonia solanacearum: identification of novel effector proteins translocated to plant host cells through the type III secretion system. Mol Microbiol *53*, 115–128.

Cuthbertson, L., Powers, J., and Whitfield, C. (2005). The C-terminal domain of the nucleotide-binding domain protein Wzt determines substrate specificity in the ATP-binding cassette transporter for the lipopolysaccharide O-antigens in Escherichia coli serotypes O8 and O9a. J Biol Chem *280*, 30310–30319.

Da Silva, A.C.R., Ferro, J.A., Reinach, F.C., Farah, C.S., Furlan, L.R., Quaggio, R.B., Monteiro-Vitorello, C.B., van Sluys, M.A., Almeida, N.F., Alves, L.M.C., et al. (2002). Comparison of the genomes of two Xanthomonas pathogens with differing host specificities. Nature *417*, 459–463.

Dai, W.-J., Zeng, Y., Xie, Z.-P., and Staehelin, C. (2008). Symbiosis-promoting and deleterious effects of NopT, a novel type 3 effector of *Rhizobium* sp. strain NGR234. J Bacteriol *190*, 5101–5110.

Dale, C., Plague, G.R., Wang, B., Ochman, H., and Moran, N.A. (2002). Type III secretion systems and the evolution of mutualistic endosymbiosis. Proc Natl Acad Sci USA *99*, 12397–12402.

Dale, C., Young, S.A., Haydon, D.T., and Welburn, S.C. (2001). The insect endosymbiont Sodalis glossinidius utilizes a type III secretion system for cell invasion. Proc Natl Acad Sci USA *98*, 1883–1888.

Dangl, J.L., and Jones, J.D. (2001). Plant pathogens and integrated defence responses to infection. Nature *411*, 826–833.

Dangl, J.L., Ritter, C., Gibbon, M.J., Mur, L.A., Wood, J.R., Goss, S., Mansfield, J., Taylor, J.D., and Vivian, A. (1992). Functional homologs of the Arabidopsis RPM1 disease resistance gene in bean and pea. Plant Cell *4*, 1359–1369.

Das, A., Rangaraj, N., and Sonti, R.V. (2009). Multiple adhesin-like functions of Xanthomonas oryzae pv. oryzae are involved in promoting leaf attachment, entry, and virulence on rice. Mol Plant Microbe Interact *22*, 73–85.

Davies, B.W., and Walker, G.C. (2007). Disruption of sitA compromises Sinorhizobium meliloti for manganese uptake required for protection against oxidative stress. J Bacteriol *189*, 2101–2109.

de Lyra, M.D.C.C.P., Lopez-Baena, F.J., Madinabeitia, N., Vinardell, J.M., Espuny, M.D.R., Cubo, M.T., Bellogín, R.A., Ruiz-Sainz, J.E., and Ollero, F.J. (2006). Inactivation of the Sinorhizobium fredii HH103 rhcJ gene abolishes nodulation outer proteins (Nops) secretion and decreases the symbiotic capacity with soybean. Int Microbiol *9*, 125–133.

de Oliveira, V.M., Manfio, G.P., Da Costa Coutinho, H.L., Keijzer-Wolters, A.C., and van Elsas, J.D. (2006). A ribosomal RNA gene intergenic spacer based PCR and DGGE fingerprinting method for the analysis of specific rhizobial communities in soil. J Microbiol Methods *64*, 366–379.

Deakin, W.J., and Broughton, W.J. (2009). Symbiotic use of pathogenic strategies: rhizobial protein secretion systems. Nat Rev Microbiol *7*, 312–320.

Deakin, W.J., Marie, C., Saad, M.M., Krishnan, H.B., and Broughton, W.J. (2005). NopA is associated with cell surface appendages produced by the type III secretion system of *Rhizobium* sp. strain NGR234. Mol Plant Microbe Interact *18*, 499–507.

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

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. (1999).

Improved microbial gene identification with GLIMMER. Nucleic Acids Res *27*, 4636–4641.

Deng, W.L., Preston, G., Collmer, A., Chang, C.J., and Huang, H.C. (1998). Characterization of the hrpC and hrpRS operons of Pseudomonas syringae pathovars syringae, tomato, and glycinea and analysis of the ability of hrpF, hrpG, hrcC, hrpT, and hrpV mutants to elicit the hypersensitive response and disease in plants. J Bacteriol *180*, 4523–4531.

Deng, X., Xiao, Y., Lan, L., Zhou, J.-M., and Tang, X. (2009). Pseudomonas syringae pv. phaseolicola Mutants Compromised for type III secretion system gene induction. Mol Plant Microbe Interact *22*, 964–976.

Dodds, P., and Lawrence, G. (2001). Six amino acid changes confined to the leucine-rich repeat ß-strand/ß-turn motif determine the difference between the P and P2 rust resistance specificities in flax. The Plant Cell Online.

Dow, M., Newman, M.-A., and Roepenack, von, E. (2000). The Induction and Modulation of Plant Defense Responses by Bacterial Lipopolysaccharides. Annual Review of Phytopathology *38*, 241–261.

Dresler-Nurmi, A., Fewer, D., Räsänen, L., and Lindström, K. (2009). The diversity and evolution of rhizobia. Prokaryotic Symbionts in Plants 3–41.

Eddy, S. (1998). Profile hidden Markov models. Bioinformatics.

Ellis, J., and Dodds, P. (2007). Flax rust resistance gene specificity is based on direct resistance-avirulence protein interactions. Annu Rev Phytopathol.

Ellison, D.W., and Miller, V.L. (2006). Regulation of virulence by members of the MarR/SlyA family. Curr Opin Microbiol *9*, 153–159.

Fahlgren, N., Sullivan, C.M., Kasschau, K.D., Chapman, E.J., Cumbie, J.S., Montgomery, T.A., Gilbert, S.D., Dasenko, M., Backman, T.W.H., Givan, S.A., et al. (2009). Computational and analytical framework for small RNA profiling by high-throughput sequencing. Rna *15*, 992–1002.

Fauvart, M., and Michiels, J. (2008). Rhizobial secreted proteins as determinants of host specificity in the rhizobium-legume symbiosis. FEMS Microbiol Lett.

Feil, H., Feil, W.S., Chain, P., Larimer, F., DiBartolo, G., Copeland, A., Lykidis, A., Trong, S., Nolan, M., Goltsman, E., et al. (2005). Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000. Proc Natl Acad Sci USA *102*, 11064–11069.

Felix, G., Duran, J.D., Volko, S., and Boller, T. (1999). Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. Plant J *18*,

265–276.

Feltman, H., Schulert, G., Khan, S., Jain, M., Peterson, L., and Hauser, A.R. (2001). Prevalence of type III secretion genes in clinical and environmental isolates of Pseudomonas aeruginosa. Microbiology (Reading, Engl) *147*, 2659.

Fenselau, S., and Bonas, U. (1995). Sequence and expression analysis of the hrpB pathogenicity operon of Xanthomonas campestris pv. vesicatoria which encodes eight proteins with similarity to components of the Hrp, Ysc, Spa, and Fli secretion systems. Mol Plant Microbe Interact *8*, 845–854.

Ferrarini, A., de Stefano, M., Baudouin, E., Pucciariello, C., Polverari, A., Puppo, A., and Delledonne, M. (2008). Expression of Medicago truncatula genes responsive to nitric oxide in pathogenic and symbiotic conditions. Mol Plant Microbe Interact *21*, 781–790.

Ferreira, A.O., Myers, C.R., Gordon, J.S., Martin, G.B., Vencato, M., Collmer, A., Wehling, M.D., Alfano, J.R., Moreno-Hagelsieb, G., Lamboy, W.F., et al. (2006). Whole-genome expression profiling defines the HrpL regulon of Pseudomonas syringae pv. tomato DC3000, allows de novo reconstruction of the Hrp cis clement, and identifies novel coregulated genes. Mol Plant Microbe Interact *19*, 1167–1179.

Flor, H. (1956). The Complementary Genic Systems in Flax and Flax Rust 10.1016/S0065-2660(08)60498-8 : Advances in Genetics | ScienceDirect.com. Advances in Genetics.

Flor, H.H. (1971). Current Status of the Gene-For-Gene Concept. Annual Review of Phytopathology *9*, 275–296.

Flores-Vargas, R.D., and O'Hara, G.W. (2006). Isolation and characterization of rhizosphere bacteria with potential for biological control of weeds in vineyards. J Appl Microbiol *100*, 946–954.

Fouts, D.E., Abramovitch, R.B., Alfano, J.R., Baldo, A.M., Buell, C.R., Cartinhour, S., Chatterjee, A.K., D'Ascenzo, M., Gwinn, M.L., Lazarowitz, S.G., et al. (2002). Genomewide identification of Pseudomonas syringae pv. tomato DC3000 promoters controlled by the HrpL alternative sigma factor. Proc Natl Acad Sci USA *99*, 2275–2280.

Freiberg, C., Fellay, R., Bairoch, A., Broughton, W.J., Rosenthal, A., and Perret, X. (1997). Molecular basis of symbiosis between Rhizobium and legumes. Nature *387*, 394–401.

Fu, Z.Q., Guo, M., Jeong, B.-R., Tian, F., Elthon, T.E., Cerny, R.L., Staiger, D., and Alfano, J.R. (2007). A type III effector ADP-ribosylates RNA-binding proteins and quells plant immunity. Nature *447*, 284–288.

Gage, D. (2002). Analysis of Infection Thread Development Using Gfp- and DsRed-Expressing Sinorhizobium meliloti. J Bacteriol.

Gage, D.J. (2004). Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumes. Microbiology and Molecular Biology Reviews *68*, 280.

Gail Preston, W.-L.D.H.-C.H.A.C. (1998). Negative Regulation of hrp Genes in Pseudomonas syringae by HrpV. J Bacteriol *180*, 4532.

Galán, J. (2009). Common themes in the design and function of bacterial effectors. Cell Host Microbe.

Galán, J.E., and Wolf-Watz, H. (2006). Protein delivery into eukaryotic cells by type III secretion machines. Nature *444*, 567–573.

Gao, M., D'Haeze, W., De Rycke, R., Wolucka, B., and Holsters, M. (2001). Knockout of an Azorhizobial dTDP-L-Rhamnose Synthase Affects Lipopolysaccharide and Extracellular Polysaccharide Production and Disables Symbiosis with Sesbania rostrata. Mol Plant Microbe Interact *14*, 857–866.

Gassmann, W., Dahlbeck, D., Chesnokova, O., Minsavage, G.V., Jones, J.B., and Staskawicz, B.J. (2000). Molecular evolution of virulence in natural field strains of Xanthomonas campestris pv. vesicatoria. J Bacteriol *182*, 7053–7059.

Giovannoni, S.J., Hayakawa, D.H., Tripp, H.J., Stingl, U., Givan, S.A., Cho, J.-C., Oh, H.-M., Kitner, J.B., Vergin, K.L., and Rappé, M.S. (2008). The small genome of an abundant coastal ocean methylotroph. Environ Microbiol *10*, 1771–1782.

Gómez-Gómez, L., Felix, G., and Boller, T. (1999). A single locus determines sensitivity to bacterial flagellin in Arabidopsis thaliana. Plant J *18*, 277–284.

Göttfert, M., Röthlisberger, S., Kündig, C., Beck, C., Marty, R., and Hennecke, H. (2001). Potential symbiosis-specific genes uncovered by sequencing a 410-kilobase DNA region of the Bradyrhizobium japonicum chromosome. J Bacteriol *183*, 1405–1412.

Graham, M., Marek, L., and Shoemaker, R. (2002a). PCR Sampling of disease resistance-like sequences from a disease resistance gene cluster in soybean. TAG Theoretical and Applied Genetics *105*, 50–57.

Graham, M.A., Marek, L.F., and Shoemaker, R.C. (2002b). Organization, expression and evolution of a disease resistance gene cluster in soybean. Genetics *162*, 1961.

Graham, T.L., Sequeira, L., and Huang, T.S. (1977). Bacterial lipopolysaccharides as inducers of disease resistance in tobacco. Appl Environ Microbiol *34*, 424–432.

Grant, S.R., Fisher, E.J., Chang, J.H., Mole, B.M., and Dangl, J.L. (2006). Subterfuge and manipulation: type III effector proteins of phytopathogenic bacteria. Annu Rev Microbiol *60*, 425–449.

Greenberg, J.T., and Vinatzer, B.A. (2003). Identifying type III effectors of plant pathogens and analyzing their interaction with plant cells. Curr Opin Microbiol *6*, 20–28.

Greenberg, J.T., and Yao, N. (2004). The role and regulation of programmed cell death in plant-pathogen interactions. Cell Microbiol *6*, 201–211.

Gross, H., and Loper, J.E. (2009). Genomics of secondary metabolite production by Pseudomonas spp. Nat Prod Rep *26*, 1408–1446.

Gudesblat, G.E., Torres, P.S., and Vojnov, A.A. (2009). Xanthomonas campestris overcomes Arabidopsis stomatal innate immunity through a DSF cell-to-cell signal-regulated virulence factor. Plant Physiol *149*, 1017–1027.

Guttman, D.S., and Greenberg, J.T. (2001). Functional analysis of the type III effectors AvrRpt2 and AvrRpm1 of Pseudomonas syringae with the use of a single-copy genomic integration system. Mol Plant Microbe Interact *14*, 145–155.

Guttman, D.S., Vinatzer, B.A., Sarkar, S.F., Ranall, M.V., Kettler, G., and Greenberg, J.T. (2002). A functional screen for the type III (Hrp) secretome of the plant pathogen Pseudomonas syringae. Science *295*, 1722–1726.

Haapalainen, M., van Gestel, K., Pirhonen, M., and Taira, S. (2009). Soluble plant cell signals induce the expression of the type III secretion system of Pseudomonas syringae and upregulate the production of pilus protein HrpA. Mol Plant Microbe Interact *22*, 282–290.

Haas, D., and Défago, G. (2005). Biological control of soil-borne pathogens by fluorescent pseudomonads. Nat Rev Microbiol *3*, 307–319.

Haney, C.H., Riely, B.K., Tricoli, D.M., Cook, D.R., Ehrhardt, D.W., and Long, S.R. (2011). Symbiotic rhizobia bacteria trigger a change in localization and dynamics of the Medicago truncatula receptor kinase LYK3. The Plant Cell Online *23*, 2774–2787.

Hao, X., Lin, Y., Johnstone, L., Baltrus, D.A., Miller, S.J., Wei, G., and Rensing, C. (2012). Draft Genome Sequence of Plant Growth-Promoting Rhizobium Mesorhizobium amorphae, Isolated from Zinc-Lead Mine Tailings. J Bacteriol *194*, 736–737.

He, S.Y., Nomura, K., and Whittam, T.S. (2004). Type III protein secretion mechanism in mammalian and plant pathogens. Biochim Biophys Acta *1694*, 181–206.

Helinski, D. (1979). Replication of an origin-containing derivative of plasmid RK2 dependent on a plasmid function provided in trans. In Proceedings of the National ….

Hempel, J., Zehner, S., Göttfert, M., and Patschkowski, T. (2009). Analysis of the secretome of the soybean symbiont Bradyrhizobium japonicum. J Biotechnol *140*, 51–58.

Hettwer, U., Jaeckel, F.R., Boch, J., Meyer, M., Rudolph, K., and Ullrich, M.S. (1998). Cloning, Nucleotide Sequence, and Expression in Escherichia coli of Levansucrase Genes from the Plant Pathogens Pseudomonas syringae pv. glycinea and P. syringae pv. phaseolicola. Appl Environ Microbiol *64*, 3180.

Hidalgo, A., Margaret, I., Crespo-Rivas, J.C., Parada, M., Murdoch, P.D.S., Lopez, A., Buendia-Claveria, A.M., Moreno, J., Albareda, M., Gil-Serrano, A.M., et al. (2010). The rkpU gene of Sinorhizobium fredii HH103 is required for bacterial K-antigen polysaccharide production and for efficient nodulation with soybean but not with cowpea. Microbiology *156*, 3398–3411.

Holland, I.B., Schmitt, L., and Young, J. (2005). Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway (review). Mol. Membr. Biol. *22*, 29–39.

Hood, R.D., Singh, P., Hsu, F., Güvener, T., Carl, M.A., Trinidad, R.R.S., Silverman, J.M., Ohlson, B.B., Hicks, K.G., Plemel, R.L., et al. (2010). A type VI secretion system of Pseudomonas aeruginosa targets a toxin to bacteria. Cell Host Microbe *7*, 25–37.

Hsu, F., Schwarz, S., and Mougous, J.D. (2009). TagR promotes PpkA-catalysed type VI secretion activation in Pseudomonas aeruginosa. Mol Microbiol *72*, 1111–1125.

Hubber, A., Vergunst, A.C., Sullivan, J.T., Hooykaas, P.J.J., and Ronson, C.W. (2004). Symbiotic phenotypes and translocated effector proteins of the Mesorhizobium loti strain R7A VirB/D4 type IV secretion system. Mol Microbiol *54*, 561–574.

Hueck, C.J. (1998). Type III protein secretion systems in bacterial pathogens of animals and plants. Microbiol Mol Biol Rev *62*, 379–433.

Hughes, K.T., and Mathee, K. (1998). The anti-sigma factors. Annu. Rev. Microbiol. *52*, 231–286.

Hutcheson, S.W., Bretz, J., Sussan, T., Jin, S., and Pak, K. (2001). Enhancer-binding proteins HrpR and HrpS interact to regulate hrp-encoded type III protein secretion in Pseudomonas syringae strains. J Bacteriol *183*, 5589–5598.

Ideses, D., Gophna, U., Paitan, Y., Chaudhuri, R.R., Pallen, M.J., and Ron, E.Z. (2005). A degenerate type III secretion system from septicemic Escherichia coli contributes to pathogenesis. J Bacteriol *187*, 8164–8171.

Innes, R.W., Bent, A.F., Kunkel, B.N., Bisgrove, S.R., and Staskawicz, B.J. (1993). Molecular analysis of avirulence gene avrRpt2 and identification of a putative regulatory sequence common to all known Pseudomonas syringae avirulence genes. J Bacteriol *175*, 4859–4869.

Jackson, R.W., Preston, G.M., and Rainey, P.B. (2005). Genetic characterization of Pseudomonas fluorescens SBW25 rsp gene expression in the phytosphere and in vitro. J Bacteriol *187*, 8477–8488.

Jackson, R.W., Vinatzer, B., Arnold, D.L., Dorus, S., and Murillo, J. (2011). The influence of the accessory genome on bacterial pathogen evolution. Mob Genet Elements *1*, 55–65.

Jamet, A., Kiss, E., Batut, J., Puppo, A., and Hérouart, D. (2005). The katA catalase gene is regulated by OxyR in both free-living and symbiotic Sinorhizobium meliloti. J Bacteriol *187*, 376–381.

Jamet, A., Sigaud, S., Van de Sype, G., Puppo, A., and Hérouart, D. (2003). Expression of the bacterial catalase genes during Sinorhizobium meliloti-Medicago sativa symbiosis and their crucial role during the infection process. Mol Plant Microbe Interact *16*, 217–225.

Jelenska, J., Yao, N., Vinatzer, B.A., Wright, C.M., Brodsky, J.L., and Greenberg, J.T. (2007). A J domain virulence effector of Pseudomonas syringae remodels host chloroplasts and suppresses defenses. Curr Biol *17*, 499–508.

Ji-Sun Lee, Y.-J.H.J.K.L.Y.-H.C. (2005). KatA, the Major Catalase, Is Critical for Osmoprotection and Virulence in Pseudomonas aeruginosa PA14. Infect Immun *73*, 4399.

Jia, Y., McAdams, S.A., Bryan, G.T., Hershey, H.P., and Valent, B. (2000). Direct interaction of resistance gene and avirulence gene products confers rice blast resistance. Embo J *19*, 4004–4014.

Joardar, V., Lindeberg, M., and Jackson, R. (2005). Whole-Genome Sequence Analysis of Pseudomonas syringae pv. phaseolicola 1448A Reveals Divergence among Pathovars in Genes Involved in Virulence and Transposition. Journal of ….

Johnson, T.L., Abendroth, J., Hol, W.G.J., and Sandkvist, M. (2006). Type II secretion: from structure to function. FEMS Microbiol Lett *255*, 175–186.

Jones, J.D.G., and Dangl, J.L. (2006). The plant immune system. Nature *444*,

323–329.

Jones, K.M., Kobayashi, H., Davies, B.W., Taga, M.E., and Walker, G.C. (2007). How rhizobial symbionts invade plants: the Sinorhizobium-Medicago model. Nat Rev Microbiol *5*, 619–633.

Kambara, K., Ardissone, S., Kobayashi, H., Saad, M.M., Schumpp, O., Broughton, W.J., and Deakin, W.J. (2009). Rhizobia utilize pathogen-like effector proteins during symbiosis. Mol Microbiol *71*, 92–106.

Kanazin, V., Marek, L.F., and Shoemaker, R.C. (1996). Resistance gene analogs are conserved and clustered in soybean. Proc Natl Acad Sci USA *93*, 11746–11750.

Kaneko, T., Maito, H., Hirakawa, H., Uchiike, N., Minamisawa, K., Watanabe, A., and Sato, S. (2011). Complete Genome Sequence of the Soybean Symbiont *Bradyrhizobium japonicum* Strain USDA6$^{T}$. Genes.

Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A., Idesawa, K., Ishikawa, A., Kawashima, K., et al. (2000). Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti. DNA Res *7*, 331–338.

Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., Watanabe, A., Idesawa, K., Iriguchi, M., Kawashima, K., et al. (2002). Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110. DNA Res *9*, 189–197.

Katzen, F., Ferreiro, D.U., Oddo, C.G., Ielmini, M.V., Becker, A., Pühler, A., and Ielpi, L. (1998). Xanthomonas campestris pv. campestris gum mutants: effects on xanthan biosynthesis and plant virulence. J Bacteriol *180*, 1607–1617.

Kearney, B., and Staskawicz, B.J. (1990). Widespread distribution and fitness contribution of Xanthomonas campestris avirulence gene avrBs2. Nature *346*, 385–386.

Kereszt, A., Mergaert, P., Maroti, G., and Kondorosi, E. (2011). Innate immunity effectors and virulence factors in symbiosis. Curr Opin Microbiol *14*, 76–81.

Kimbrel, J.A., Givan, S.A., Halgren, A.B., Creason, A.L., Mills, D.I., Banowetz, G.M., Armstrong, D.J., and Chang, J.H. (2010). An improved, high-quality draft genome sequence of the Germination-Arrest Factor-producing Pseudomonas fluorescens WH6. BMC Genomics *11*, 522.

Kimbrel, J.A., Givan, S.A., Temple, T.N., Johnson, K.B., and Chang, J.H. (2011). Genome sequencing and comparative analysis of the carrot bacterial blight pathogen, Xanthomonas hortorum pv. carotae M081, for insights into

pathogenicity and applications in molecular diagnostics. Molecular Plant Pathology *12*, 580–594.

Klassen, J.L., and Currie, C.R. (2012). Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. BMC Genomics *13*, 14.

Knodler, L.A., Vallance, B.A., Hensel, M., Jäckel, D., Finlay, B.B., and Steele-Mortimer, O. (2004). Salmonella type III effectors PipB and PipB2 are targeted to detergent-resistant microdomains on internal host cell membranes. Mol Microbiol *49*, 685–704.

Koczan, J.M., McGrath, M.J., Zhao, Y., and Sundin, G.W. (2009). Contribution of Erwinia amylovoraExopolysaccharides Amylovoran and Levan to Biofilm Formation: Implications in Pathogenicity. Phytopathology *99*, 1237–1244.

Koebnik, R., Krüger, A., Thieme, F., Urban, A., and Bonas, U. (2006). Specific binding of the Xanthomonas campestris pv. vesicatoria AraC-type transcriptional activator HrpX to plant-inducible promoter boxes. J Bacteriol *188*, 7652–7660.

Kouchi, H., Shimomura, K., Hata, S., Hirota, A., Wu, G.-J., Kumagai, H., Tajima, S., Suganuma, N., Suzuki, A., Aoki, T., et al. (2004). Large-scale analysis of gene expression profiles during early stages of root nodule formation in a model legume, Lotus japonicus. DNA Res *11*, 263–274.

Krasileva, K.V., Dahlbeck, D., and Staskawicz, B.J. (2010). Activation of an Arabidopsis resistance protein is specified by the in planta association of its leucine-rich repeat domain with the cognate oomycete effector. The Plant Cell Online *22*, 2444–2458.

Krause, A., Doerfel, A., and Göttfert, M. (2002). Mutational and transcriptional analysis of the type III secretion system of Bradyrhizobium japonicum. Mol Plant Microbe Interact *15*, 1228–1235.

Krishnan, H.B., Lorio, J., Kim, W.S., Jiang, G., Kim, K.Y., DeBoer, M., and Pueppke, S.G. (2003). Extracellular proteins involved in soybean cultivar-specific nodulation are associated with pilus-like surface appendages and exported by a type III protein secretion system in *Sinorhizobium fredii* USDA257. Mol Plant Microbe Interact *16*, 617–625.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639–1645.

Kunze, G., Zipfel, C., Robatzek, S., Niehaus, K., Boller, T., and Felix, G. (2004). The N terminus of bacterial elongation factor Tu elicits innate immunity in Arabidopsis plants. Plant Cell *16*, 3496–3507.

Kurtz, S., Phillippy, A., Delcher, A., and Smoot, M. (2004). Versatile and open software for comparing large genomes. Genome ….

Lagesen, K., Hallin, P., Rodland, E., and Staerfeldt, H. (2007). RNammer: consistent annotation of rRNA genes in genomic sequences (Nucleic Acids Res).

Lang, J.M., Hamilton, J.P., Diaz, M.G.Q., van Sluys, M.-A., Burgos, M.R.G., Vera-Cruz, C.M., Buell, C.R., Tisserat, N.A., and Leach, J.E. (2010). Genomics-Based Diagnostic Marker Development for Xanthomonas oryzaepv. oryzaeand X. oryzaepv. oryzicola. Plant Disease *94*, 311–319.

Larkin, M., Blackshields, G., and Brown, N. (2007). Clustal W and Clustal X version 2.0. ….

Law, C.J., Maloney, P.C., and Wang, D.-N. Ins and Outs of Major Facilitator Superfamily Antiporters. Http://Dx.Doi.org.Proxy.Library.Oregonstate.Edu/10.1146/Annurev.Micro.61.080706.093329.

Law, R., and Lewis, D. (2008). Biotic environments and the maintenance of sex-some evidence from mutualistic symbioses. Biological Journal of the Linnean Society *20*, 249–276.

Leach, J., Cruz, C.V., and Bai, J. (2001). Pathogen fitness penalty as a predictor of durability of disease resistance genes. Annual Review of ….

Lee, K.-B., De Backer, P., Aono, T., Liu, C.-T., Suzuki, S., Suzuki, T., Kaneko, T., Yamada, M., Tabata, S., Kupfer, D.M., et al. (2008). The genome of the versatile nitrogen fixer Azorhizobium caulinodans ORS571. BMC Genomics *9*, 271.

Lewis, J., Guttman, D., and Desveaux, D. (2009). The targeting of plant cellular systems by injected type III effector proteins. Semin Cell Dev Biol.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Li, X., Lin, H., Zhang, W., Zou, Y., Zhang, J., Tang, X., and Zhou, J.-M. (2005). Flagellin induces innate immunity in nonhost interactions that is suppressed by Pseudomonas syringae effectors. Proc Natl Acad Sci USA *102*, 12990–12995.

Li, Y., Sun, Z., Zhuang, X., Xu, L., and Chen, S. (2003). Research progress on microbial herbicides. Crop Protection.

Lindeberg, M., Myers, C.R., Collmer, A., and Schneider, D.J. (2008). Roadmap to new virulence determinants in Pseudomonas syringae: insights from comparative genomics and genome organization. Mol Plant Microbe Interact *21*, 685–700.

Lindeberg, M., Stavrinides, J., Chang, J.H., Alfano, J.R., Collmer, A., Dangl, J.L., Greenberg, J.T., Mansfield, J.W., and Guttman, D.S. (2005). Proposed guidelines for a unified nomenclature and phylogenetic analysis of type III Hop effector proteins in the plant pathogen Pseudomonas syringae. Mol Plant Microbe Interact *18*, 275–282.

Lindgren, P.B., Peet, R.C., and Panopoulos, N.J. (1986). Gene cluster of Pseudomonas syringae pv. “phaseolicola” controls pathogenicity of bean plants and hypersensitivity of nonhost plants. J Bacteriol *168*, 512–522.

Lopez-Gomez, M., Sandal, N., Stougaard, J., and Boller, T. (2012). Interplay of flg22-induced defence responses and nodulation in Lotus japonicus. J Exp Bot *63*, 393–401.

Lorio, J.C., Kim, W.S., and Krishnan, H.B. (2004). NopB, a soybean cultivar-specificity protein from *Sinorhizobium fredii* USDA257, is a type III secreted protein. Mol Plant Microbe Interact *17*, 1259–1268.

Lowe, T. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res.

López-Baena, F.J., Vinardell, J.M., Pérez-Montaño, F., Crespo-Rivas, J.C., Bellogín, R.A., Espuny, M.D.R., and Ollero, F.J. (2008). Regulation and symbiotic significance of nodulation outer proteins secretion in Sinorhizobium fredii HH103. Microbiology (Reading, Engl) *154*, 1825–1836.

Lu, H., Patil, P., van Sluys, M.-A., White, F.F., Ryan, R.P., Dow, J.M., Rabinowicz, P., Salzberg, S.L., Leach, J.E., Sonti, R., et al. (2008). Acquisition and evolution of plant pathogenesis-associated gene clusters and candidate determinants of tissue-specificity in xanthomonas. PLoS ONE *3*, e3828.

Luna, E., Pastor, V., Robert, J., Flors, V., Mauch-Mani, B., and Ton, J. (2011). Callose Deposition: A Multifaceted Plant Defense Response. Mol Plant Microbe Interact *24*, 183–193.

Ma, W. (2011). Roles of Ca2+ and cyclic nucleotide gated channel in plant innate immunity. Plant Sci. *181*, 342–346.

Ma, W., Dong, F.F.T., Stavrinides, J., and Guttman, D.S. (2006). Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. PLoS Genet *2*, e209.

Mackey, D., Belkhadir, Y., Alonso, J.M., Ecker, J.R., and Dangl, J.L. (2003). Arabidopsis RIN4 is a target of the type III virulence effector AvrRpt2 and modulates RPS2-mediated resistance. Cell *112*, 379–389.

Mansfield, J.W. (2009). From bacterial avirulence genes to effector functions via

the hrp delivery system: an overview of 25 years of progress in our understanding of plant innate immunity. Molecular Plant Pathology *10*, 721–734.

Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., et al. (2009). CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res *37*, D205–10.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res *39*, D225–9.

Margaret, I., Becker, A., Blom, J., Bonilla, I., Goesmann, A., Göttfert, M., Lloret, J., Mittard-Runte, V., Rückert, C., Ruiz-Sainz, J.E., et al. (2011). Symbiotic properties and first analyses of the genomic sequence of the fast growing model strain Sinorhizobium fredii HH103 nodulating soybean. J Biotechnol *155*, 11–19.

Maria Lopez-Lara, I., Orgambide, G., Dazzo, F.B., Olivares, J., and Toro, N. (1995). Surface polysaccharide mutants of Rhizobium sp. (Acacia) strain GRH2: major requirement of lipopolysaccharide for successful invasion of Acacia nodules and host range determination. Microbiology *141*, 573–581.

Marie, C., Broughton, W.J., and Deakin, W.J. (2001). Rhizobium type III secretion systems: legume charmers or alarmers? Curr Opin Plant Biol *4*, 336–342.

Marie, C., Deakin, W.J., Ojanen-Reuhs, T., Diallo, E., Reuhs, B., Broughton, W.J., and Perret, X. (2004). TtsI, a key regulator of Rhizobium species NGR234 is required for type III-dependent protein secretion and synthesis of rhamnose-rich polysaccharides. Mol Plant Microbe Interact *17*, 958–966.

Marie, C., Deakin, W.J., Viprey, V., Kopciñska, J., Golinowski, W., Krishnan, H.B., Perret, X., and Broughton, W.J. (2003). Characterization of Nops, nodulation outer proteins, secreted via the type III secretion system of NGR234. Mol Plant Microbe Interact *16*, 743–751.

Masson-Boivin, C., Giraud, E., Perret, X., and Batut, J. (2009). Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? Trends Microbiol *17*, 458–466.

Matthysse, A.G., Stretton, S., Dandie, C., McClure, N.C., and Goodman, A.E. (1996). Construction of GFP vectors for use in Gram-negative bacteria other than Escherichia coli. FEMS Microbiol Lett *145*, 87–94.

Mazurier, S., Lemunier, M., Hartmann, A., Siblot, S., and Lemanceau, P. (2006). Conservation of type III secretion system genes in Bradyrhizobium isolated from soybean. FEMS Microbiol Lett *259*, 317–325.

McCann, H.C., and Guttman, D.S. (2008). Evolution of the type III secretion system and its effectors in plant-microbe interactions. New Phytol *177*, 33–47.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. Curr. Opin. Genet. Dev. *15*, 589–594.

Meinhardt, L.W., Krishnan, H.B., Balatti, P.A., and Pueppke, S.G. (1993). Molecular cloning and characterization of a sym plasmid locus that regulates cultivar-specific nodulation of soybean by Rhizobium fredii USDA257. Mol Microbiol *9*, 17–29.

Meng, X., Umesh, K., Davis, R., and Gilbertson, R. (2004). Development of PCR-based assays for detecting Xanthomonas campestris pv. carotae, the carrot bacterial leaf blight pathogen, from different substrates. Plant Disease *88*, 1226–1234.

Merhej, V., Royer-Carenzi, M., Pontarotti, P., and Raoult, D. (2009). Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol Direct *4*, 13.

Meyers, D.J., and Berk, R.S. (1990). Characterization of phospholipase C from Pseudomonas aeruginosa as a potent inflammatory agent. Infect Immun *58*, 659–666.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., and Marshall, D. (2010). Tablet--next generation sequence assembly visualization. Bioinformatics *26*, 401–402.

Mithöfer, A. (2002). Suppression of plant defence in rhizobia-legume symbiosis. Trends Plant Sci *7*, 440–444.

Mitra, R.M., Shaw, S.L., and Long, S.R. (2004). Six nonnodulating plant mutants defective for Nod factor-induced transcriptional changes associated with the legume-rhizobia symbiosis. Proc Natl Acad Sci USA *101*, 10217–10222.

Moran, N.A., Degnan, P.H., Santos, S.R., Dunbar, H.E., and Ochman, H. (2005). The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. Proc Natl Acad Sci USA *102*, 16919–16926.

Moreira, L.M., Almeida, N.F., Potnis, N., Digiampietri, L.A., Adi, S.S., Bortolossi, J.C., Da Silva, A.C., Da Silva, A.M., de Moraes, F.E., de Oliveira, J.C., et al. (2010). Novel insights into the genomic basis of citrus canker based on the genome sequences of two strains of Xanthomonas fuscans subsp. aurantifolii. BMC Genomics *11*, 238.

Mori, Y., and Notomi, T. (2009). Loop-mediated isothermal amplification (LAMP): a rapid, accurate, and cost-effective diagnostic method for infectious diseases. J.

Infect. Chemother. *15*, 62–69.

Moscou, M.J., and Bogdanove, A.J. (2009). A simple cipher governs DNA recognition by TAL effectors. Science *326*, 1501.

Mougous, J.D., Cuff, M.E., Raunser, S., Shen, A., Zhou, M., Gifford, C.A., Goodman, A.L., Joachimiak, G., Ordoñez, C.L., Lory, S., et al. (2006). A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus. Science *312*, 1526–1530.

Mudgett, M.B., and Staskawicz, B.J. (1999). Characterization of the Pseudomonas syringae pv. tomato AvrRpt2 protein: demonstration of secretion and processing during bacterial pathogenesis. Mol Microbiol *32*, 927–941.

Mudgett, M.B., Chesnokova, O., Dahlbeck, D., Clark, E.T., Rossier, O., Bonas, U., and Staskawicz, B.J. (2000). Molecular signals required for type III secretion and translocation of the Xanthomonas campestris AvrBs2 protein to pepper plants. Proc Natl Acad Sci USA *97*, 13324–13329.

Mukaihara, T., Tamura, N., Murata, Y., and Iwabuchi, M. (2004). Genetic screening of Hrp type III-related pathogenicity genes controlled by the HrpB transcriptional activator in Ralstonia solanacearum. Mol Microbiol *54*, 863–875.

Nakagawa, T., Kurose, T., Hino, T., Tanaka, K., Kawamukai, M., Niwa, Y., Toyooka, K., Matsuoka, K., Jinbo, T., and Kimura, T. (2007). Development of series of gateway binary vectors, pGWBs, for realizing efficient construction of fusion genes for plant transformation. J. Biosci. Bioeng. *104*, 34–41.

Niehaus, K., and Lagares, A. (1998). A Sinorhizobium meliloti lipopolysaccharide mutant induces effective nodules on the host plant Medicago sativa (alfalfa) but fails to establish a symbiosis with …. Molecular Plant-Microbe ….

Niepold, F., Anderson, D., and Mills, D. (1985). Cloning determinants of pathogenesis from Pseudomonas syringae pathovar syringae. Proc Natl Acad Sci USA *82*, 406–410.

Nissan, G., Manulis, S., Weinthal, D.M., Sessa, G., and Barash, I. (2005). Analysis of promoters recognized by HrpL, an alternative sigma-factor protein from Pantoea agglomerans pv. gypsophilae. Mol Plant Microbe Interact *18*, 634–643.

Nissinen, R.M., Ytterberg, A.J., Bogdanove, A.J., van Wijk, K.J., and Beer, S.V. (2007). Analyses of the secretomes of Erwinia amylovora and selected hrp mutants reveal novel type III secreted proteins and an effect of HrpJ on extracellular harpin levels. Molecular Plant Pathology *8*, 55–67.

O'Brien, H.E., Thakur, S., and Guttman, D.S. (2010). Evolution of Plant

Pathogenesis in Pseudomonas syringae: A Genomics Perspective. Annual Review of Phytopathology.

Ortiz-Martín, I., Thwaites, R., Mansfield, J.W., and Beuzón, C.R. (2010). Negative Regulation of the Hrp Type III Secretion System in Pseudomonas syringaepv. phaseolicola. Mol Plant Microbe Interact *23*, 682–701.

Pallen, M.J., Beatson, S.A., and Bailey, C.M. (2005). Bioinformatics, genomics and evolution of non-flagellar type-III secretion systems: a Darwinian perspective. FEMS Microbiol Rev *29*, 201–229.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., Holden, M.T.G., Churcher, C.M., Bentley, S.D., Mungall, K.L., et al. (2003). Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nat Genet *35*, 32–40.

Parkinson, N., Aritua, V., Heeney, J., Cowie, C., Bew, J., and Stead, D. (2007). Phylogenetic analysis of Xanthomonas species by comparison of partial gyrase B gene sequences. Int J Syst Evol Microbiol *57*, 2881–2887.

Parkinson, N., Cowie, C., Heeney, J., and Stead, D. (2009). Phylogenetic structure of Xanthomonas determined by comparison of gyrB sequences. Int J Syst Evol Microbiol *59*, 264–274.

Paulsen, I.T., Press, C.M., Ravel, J., Kobayashi, D.Y., Myers, G.S.A., Mavrodi, D.V., Deboy, R.T., Seshadri, R., Ren, Q., Madupu, R., et al. (2005). Complete genome sequence of the plant commensal Pseudomonas fluorescens Pf-5. Nat Biotechnol *23*, 873–878.

Peck, M.C., Fisher, R.F., and Long, S.R. (2006). Diverse flavonoids stimulate NodD1 binding to nod gene promoters in Sinorhizobium meliloti. J Bacteriol *188*, 5417–5427.

Perret, X., Kobayashi, H., and Collado-Vides, J. (2003). Regulation of expression of symbiotic genes in Rhizobium sp. NGR234. Indian J. Exp. Biol. *41*, 1101–1113.

Petkau, A., Stuart-Edwards, M., Stothard, P., and Van Domselaar, G. (2010). Interactive microbial genome visualization with GView. Bioinformatics *26*, 3125–3126.

Petnicki-Ocwieja, T., Schneider, D.J., Tam, V.C., Chancey, S.T., Shan, L., Jamir, Y., Schechter, L.M., Janes, M.D., Buell, C.R., Tang, X., et al. (2002). Genomewide identification of proteins secreted by the Hrp type III protein secretion system of Pseudomonas syringae pv. tomato DC3000. Proc Natl Acad Sci USA *99*, 7652–7657.

Pop, M., and Salzberg, S.L. (2008). Bioinformatics challenges of new sequencing

technology. Trends Genet *24*, 142–149.

Poplawsky, A., Robles, L., Chun, W., and Derie, M. (2004). Identification of a Xanthomonas pathogen of coriander from Oregon USA. Phytopathology.

Postel, S., and Kemmerling, B. (2009). Plant systems for recognition of pathogen-associated molecular patterns. Semin Cell Dev Biol *20*, 1025–1031.

Potvin, E., Sanschagrin, F., and Levesque, R.C. (2008). Sigma factors in Pseudomonas aeruginosa. FEMS Microbiol Rev *32*, 38–55.

Preston, G.M., Bertrand, N., and Rainey, P.B. (2001). Type III secretion in plant growth-promoting Pseudomonas fluorescens SBW25. Mol Microbiol *41*, 999–1014.

Pueppke, S., Bolanos-Vasquez, M., Werner, D., Bec-Ferte, M., Prome, J., and Krishnan, H. (1998). Release of flavonoids by the soybean cultivars McCall and peking and their perception as signals by the nitrogen-fixing symbiont *Sinorhizobium fredii*. Plant Physiol *117*, 599–606.

Pueppke, S.G., and Broughton, W.J. (1999). Rhizobium sp. strain NGR234 and R. fredii USDA257 share exceptionally broad, nested host ranges. Mol Plant Microbe Interact *12*, 293–318.

Pukatzki, S., Ma, A.T., Revel, A.T., Sturtevant, D., and Mekalanos, J.J. (2007). Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin. Proc Natl Acad Sci USA *104*, 15508–15513.

Pukatzki, S., Ma, A.T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W.C., Heidelberg, J.F., and Mekalanos, J.J. (2006). Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system. Proc Natl Acad Sci USA *103*, 1528–1533.

R Development Core Team. R: A Language and Environment for Statistical Computing. http://www.R-Project.org.

Rafiqi, M., Bernoux, M., Ellis, J.G., and Dodds, P.N. (2009). In the trenches of plant pathogen recognition: Role of NB-LRR proteins. Semin Cell Dev Biol *20*, 1017–1024.

Rainey, P.B. (1999). Adaptation of Pseudomonas fluorescens to the plant rhizosphere. Environ Microbiol *1*, 243–257.

Ramos, A.R., Morello, J.E., Ravindran, S., Deng, W.-L., Huang, H.-C., and Collmer, A. (2007). Identification of Pseudomonas syringae pv. syringae 61 type III secretion system Hrp proteins that can travel the type III pathway and contribute to the translocation of effector proteins into plant cells. J Bacteriol *189*,

5773–5778.

Rezzonico, F., Binder, C., Défago, G., and Moënne-Loccoz, Y. (2005). The type III secretion system of biocontrol Pseudomonas fluorescens KD targets the phytopathogenic Chromista Pythium ultimum and promotes cucumber protection. Mol Plant Microbe Interact *18*, 991–1001.

Rezzonico, F., Défago, G., and Moënne-Loccoz, Y. (2004). Comparison of ATPase-encoding type III secretion system hrcN genes in biocontrol fluorescent Pseudomonads and in phytopathogenic proteobacteria. Appl Environ Microbiol *70*, 5119–5131.

Rissman, A.I., Mau, B., Biehl, B.S., Darling, A.E., Glasner, J.D., and Perna, N.T. (2009). Reordering contigs of draft genomes using the Mauve aligner. Bioinformatics *25*, 2071–2073.

Robbertse, B., Reeves, J.B., Schoch, C.L., and Spatafora, J.W. (2006). A phylogenomic analysis of the Ascomycota. Fungal Genet Biol *43*, 715–725.

Robbertse, B., Yoder, R.J., Boyd, A., Reeves, J., and Spatafora, J.W. (2011). Hal: an Automated Pipeline for Phylogenetic Analyses of Genomic Data. PLoS Curr *3*, RRN1213.

Rocchetta, H.L., and Lam, J.S. (1997). Identification and functional characterization of an ABC transport system involved in polysaccharide export of A-band lipopolysaccharide in Pseudomonas aeruginosa. J Bacteriol *179*, 4713–4724.

Rocha, E.P.C. (2004). The replication-related organization of bacterial genomes. Microbiology *150*, 1609–1627.

Rodrigues, J.A., López-Baena, F.J., Ollero, F.J., Vinardell, J.M., Espuny, M.D.R., Bellogín, R.A., Ruiz-Sainz, J.E., Thomas, J.R., Sumpton, D., Ault, J., et al. (2007). NopM and NopD are rhizobial nodulation outer proteins: identification using LC-MALDI and LC-ESI with a monolithic capillary column. J. Proteome Res. *6*, 1029–1037.

Saad, M.M., Kobayashi, H., Marie, C., Brown, I.R., Mansfield, J.W., Broughton, W.J., and Deakin, W.J. (2005). NopB, a type III secreted protein of Rhizobium sp. strain NGR234, is associated with pilus-like surface appendages. J Bacteriol *187*, 1173–1181.

Sachs, J.L., Essenberg, C.J., and Turcotte, M.M. (2011a). New paradigms for the evolution of beneficial infections. Trends Ecol. Evol. (Amst.) *26*, 202–209.

Sachs, J.L., Russell, J.E., and Hollowell, A.C. (2011b). Evolutionary instability of symbiotic function in Bradyrhizobium japonicum. PLoS ONE *6*, e26370.

Saldaña, G., Martinez-Alcántara, V., Vinardell, J.M., Bellogín, R., Ruiz-Sainz, J.E., and Balatti, P.A. (2003). Genetic diversity of fast-growing rhizobia that nodulate soybean ( Glycine max L. Merr). Arch. Microbiol. *180*, 45–52.

Salzberg, S.L., Sommer, D.D., Schatz, M.C., Phillippy, A.M., Rabinowicz, P.D., Tsuge, S., Furutani, A., Ochiai, H., Delcher, A.L., Kelley, D., et al. (2008). Genome sequence and rapid evolution of the rice pathogen Xanthomonas oryzae pv. oryzae PXO99A. BMC Genomics *9*, 204.

Santos, R., Hérouart, D., Sigaud, S., Touati, D., and Puppo, A. (2001). Oxidative burst in alfalfa-Sinorhizobium meliloti symbiotic interaction. Mol Plant Microbe Interact *14*, 86–89.

Sánchez, C., Iannino, F., Deakin, W.J., Ugalde, R.A., and Lepek, V.C. (2009). Characterization of the Mesorhizobium loti MAFF303099 type-three protein secretion system. Mol Plant Microbe Interact *22*, 519–528.

Schaad, N., and Stall, R. (1988). Xanthomonas. In: Laboratory Guide for Identification of Plant Pathogenic Bacteria, (Laboratory guide for identification of plant pathogenic bacteria).

Schechter, L.M., Guenther, J., Olcay, E.A., Jang, S., and Krishnan, H.B. (2010). Translocation of NopP by *Sinorhizobium fredii* USDA257 into *Vigna unguiculata* root nodules. Appl Environ Microbiol *76*, 3758–3761.

Schechter, L.M., Vencato, M., Jordan, K.L., Schneider, S.E., Schneider, D.J., and Collmer, A. (2006). Multiple approaches to a complete inventory of Pseudomonas syringae pv. tomato DC3000 type III secretion system effector proteins. Mol Plant Microbe Interact *19*, 1180–1192.

Schmeisser, C., Liesegang, H., Krysciak, D., Bakkou, N., Le Quéré, A., Wollherr, A., Heinemeyer, I., Morgenstern, B., Pommerening-Röser, A., Flores, M., et al. (2009). Rhizobium sp. strain NGR234 possesses a remarkable number of secretion systems. Appl Environ Microbiol *75*, 4035–4045.

Schwessinger, B., and Zipfel, C. (2008). News from the frontline: recent insights into PAMP-triggered immunity in plants. Curr Opin Plant Biol *11*, 389–395.

Scofield, S., Tobias, C., Rathjen, J., Chang, J., Lavelle, D., Michelmore, R., and Staskawicz, B. (1996). Molecular Basis of Gene-for-Gene Specificity in Bacterial Speck Disease of Tomato. Science *274*, 2063–2065.

Segonzac, C., and Zipfel, C. (2011). Activation of plant pattern-recognition receptors by bacteria. Curr Opin Microbiol *14*, 54–61.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp.

APS. Nature *407*, 81–86.

Shrivastava, S., and Mande, S.S. (2008). Identification and functional characterization of gene components of Type VI Secretion system in bacterial genomes. PLoS ONE *3*, e2955.

Sigaud, S., Becquet, V., Frendo, P., Puppo, A., and Hérouart, D. (1999). Differential regulation of two divergent Sinorhizobium meliloti genes for HPII-like catalases during free-living growth and protective role of both catalases during symbiosis. J Bacteriol *181*, 2634–2639.

Silby, M., Cerdeño-Tárraga, A.M., Vernikos, G., Giddens, S., Jackson, R., Preston, G., Zhang, X., Moon, C., Gehrig, S., Godfrey, S., et al. (2009). Genomic and genetic analyses of diversity and plant interactions of Pseudomonas fluorescens. Genome Biol *10*, R51.

Skorpil, P., Saad, M.M., Boukli, N.M., Kobayashi, H., Ares-Orpel, F., Broughton, W.J., and Deakin, W.J. (2005). NopP, a phosphorylated effector of Rhizobium sp. strain NGR234, is a major determinant of nodulation of the tropical legumes Flemingia congesta and Tephrosia vogelii. Mol Microbiol *57*, 1304–1317.

Sohn, K., Lei, R., and Nemri, A. (2007). The Downy Mildew Effector Proteins ATR1 and ATR13 Promote Disease Susceptibility in Arabidopsis thaliana. The Plant Cell Online.

Soto, M.J., Domínguez-Ferreras, A., Pérez-Mendoza, D., Sanjuan, J., and Olivares, J. (2009). Mutualism versus pathogenesis: the give-and-take in plant-bacteria interactions. Cell Microbiol *11*, 381–388.

Spaink, H.P. (2000). Root nodulation and infection factors produced by rhizobial bacteria. Annu Rev Microbiol *54*, 257–288.

Staroń, A., Sofia, H.J., Dietrich, S., Ulrich, L.E., Liesegang, H., and Mascher, T. (2009). The third pillar of bacterial signal transduction: classification of the extracytoplasmic function (ECF) sigma factor protein family. Mol Microbiol *74*, 557–581.

Stavrinides, J., McCann, H.C., and Guttman, D.S. (2008). Host-pathogen interplay and the evolution of bacterial effectors. Cell Microbiol *10*, 285–292.

Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, S.D., Warrener, P., Hickey, M.J., Brinkman, F.S., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., et al. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature *406*, 959–964.

Studholme, D.J., Ibanez, S.G., MacLean, D., Dangl, J.L., Chang, J.H., and Rathjen, J.P. (2009). A draft genome sequence and functional screen reveals the

repertoire of type III secreted proteins of Pseudomonas syringae pathovar tabaci 11528. BMC Genomics *10*, 395.

Sullivan, J.T.J., and Ronson, C.W.C. (1998). Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. Proc Natl Acad Sci USA *95*, 5145–5149.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res *34*, W609–12.

Swords, K.M., Dahlbeck, D., Kearney, B., Roy, M., and Staskawicz, B.J. (1996). Spontaneous and induced mutations in a single open reading frame alter both virulence and avirulence in Xanthomonas campestris pv. vesicatoria avrBs2. J Bacteriol *178*, 4661–4669.

Tai, T.H., Dahlbeck, D., Clark, E.T., Gajiwala, P., Pasion, R., Whalen, M.C., Stall, R.E., and Staskawicz, B.J. (1999). Expression of the Bs2 pepper gene confers resistance to bacterial spot disease in tomato. Proc Natl Acad Sci USA *96*, 14153–14158.

Tan, Y., and Donovan, W.P. (2000). Deletion of aprA and nprA genes for alkaline protease A and neutral protease A from Bacillus thuringiensis: effect on insecticidal crystal proteins. J Biotechnol *84*, 67–72.

Tang, X., Frederick, R., Zhou, J., Halterman, D., Jia, Y., and Martin, G. (1996). Initiation of Plant Disease Resistance by Physical Interaction of AvrPto and Pto Kinase. Science *274*, 2060–2063.

Temple, T., and KB, J. (2009). Detection of Xanthomonas hortorum pv. carotae on and in carrot with loop-mediated isothermal amplification (LAMP) (Abstr.). Phytopathology *99*, S186.

Tena, G., Boudsocq, M., and Sheen, J. (2011). Protein kinase signaling networks in plant innate immunity. Curr Opin Plant Biol *14*, 519–529.

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol *11*, 472–477.

Thieme, F., Koebnik, R., Bekel, T., Berger, C., Boch, J., Büttner, D., Caldana, C., Gaigalat, L., Goesmann, A., Kay, S., et al. (2005). Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium Xanthomonas campestris pv. vesicatoria revealed by the complete genome sequence. J Bacteriol *187*, 7254–7266.

Thomas, W.J., Thireault, C.A., (null), and Chang, J.H. (2009). Recombineering and stable integration of the Pseudomonas syringae pv. syringae 61 hrp/hrc

cluster into the genome of the soil bacterium Pseudomonas fluorescens Pf0-1. Plant J *60*, 919–928.

Tock, M.R., and Dryden, D.T.F. (2005). The biology of restriction and anti-restriction. Curr Opin Microbiol *8*, 466–472.

Toh, H., Weiss, B.L., Perkin, S.A.H., Yamashita, A., Oshima, K., Hattori, M., and Aksoy, S. (2006). Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of Sodalis glossinidius in the tsetse host. Genome Res *16*, 149–156.

Toit, du, L., Crowe, F., Derie, M., Simmons, R., and Pelter, G. (2005). Bacterial blight in carrot seed crops in the Pacific Northwest. Plant Disease *89*, 896–907.

Torres, M.A., Jones, J.D.G., and Dangl, J.L. (2006). Reactive oxygen species signaling in response to pathogens. Plant Physiol *141*, 373–378.

Triplett, E.W., and Sadowsky, M.J. (1992). Genetics of Competition for Nodulation of Legumes. Annu Rev Microbiol *46*, 399–422.

Tsiamis, G., Mansfield, J.W., Hockenhull, R., Jackson, R.W., Sesma, A., Athanassopoulos, E., Bennett, M.A., Stevens, C., Vivian, A., Taylor, J.D., et al. (2000). Cultivar-specific avirulence and virulence functions assigned to avrPphF in Pseudomonas syringae pv. phaseolicola, the cause of bean halo-blight disease. Embo J *19*, 3204–3214.

Tsuge, S., Terashima, S., and Furutani, A. (2005). Effects on promoter activity of base substitutions in the cis-acting regulatory element of HrpXo regulons in Xanthomonas oryzae pv. oryzae. Journal of ….

van Berkum, P., and Fuhrmann, J.J. (2000). Evolutionary relationships among the soybean bradyrhizobia reconstructed from 16S rRNA gene and internally transcribed spacer region sequence divergence. Int J Syst Evol Microbiol *50 Pt 6*, 2165–2172.

van Berkum, P., Terefework, Z., Paulin, L., Suomalainen, S., Lindstrom, K., and Eardly, B.D. (2003). Discordant phylogenies within the rrn loci of Rhizobia. J Bacteriol *185*, 2988–2998.

van der Biezen, E.A., and Jones, J.D. (1998). Plant disease-resistance proteins and the gene-for-gene concept. Trends Biochem Sci *23*, 454–456.

Vasse, J., Billy, F., and Truchet, G. (1993). Abortion of infection during the Rhizobium meliloti—alfalfa symbiotic interaction is accompanied by a hypersensitive reaction. Plant J *4*, 555–566.

Vauterin, L., Hoste, B., Kersters, K., and Swings, J. (1995). Reclassification of

Xanthomonas. International Journal of Systematic Bacteriology *45*, 472–489.

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

Vinatzer, B.A., Jelenska, J., and Greenberg, J.T. (2005). Bioinformatics correctly identifies many type III secretion substrates in the plant pathogen Pseudomonas syringae and the biocontrol isolate P. fluorescens SBW25. Mol Plant Microbe Interact *18*, 877–888.

Viprey, V., del Greco, A., Golinowski, W., Broughton, W.J., and Perret, X. (1998). Symbiotic implications of type III protein secretion machinery in Rhizobium. Mol Microbiol *28*, 1381–1389.

Wassem, R., Kobayashi, H., Kambara, K., Le Quéré, A., Walker, G.C., Broughton, W.J., and Deakin, W.J. (2008). TtsI regulates symbiotic genes in Rhizobium species NGR234 by binding to tts boxes. Mol Microbiol *68*, 736–748.

Wei, C.-F., Kvitko, B.H., Shimizu, R., Crabill, E., Alfano, J.R., Lin, N.-C., Martin, G.B., Huang, H.-C., and Collmer, A. (2007). A Pseudomonas syringae pv. tomato DC3000 mutant lacking the type III effector HopQ1-1 is able to cause disease in the model plant Nicotiana benthamiana. Plant J *51*, 32–46.

Wei, Z., Kim, J.F., and Beer, S.V. (2000). Regulation of hrp genes and type III protein secretion in Erwinia amylovora by HrpX/HrpY, a novel two-component system, and HrpS. Mol Plant Microbe Interact *13*, 1251–1262.

Wenzel, M., Friedrich, L., Göttfert, M., and Zehner, S. (2010). The Type III-Secreted Protein NopE1 Affects Symbiosis and Exhibits a Calcium-Dependent Autocleavage Activity. Mol Plant Microbe Interact *23*, 124–129.

White, F.F., Potnis, N., Jones, J.B., and Koebnik, R. (2009). The type III effectors of Xanthomonas. Molecular Plant Pathology *10*, 749–766.

Wilkinson, B., and Micklefield, J. (2009). Chapter 14 Biosynthesis of Nonribosomal Peptide Precursors (Elsevier).

Williford, R., and Schaad, N. (1984). Agar medium for selective isolation of Xanthomonas campestris pv. carotae from carrot seeds (Phytopathology).

Wolfgang, M.C., Kulasekara, B.R., Liang, X., Boyd, D., Wu, K., Yang, Q., Miyada, C.G., and Lory, S. (2003). Conservation of genome content and virulence determinants among clinical and environmental isolates of Pseudomonas aeruginosa. Proc Natl Acad Sci USA *100*, 8484–8489.

Wu, H.-Y., Chung, P.-C., Shih, H.-W., Wen, S.-R., and Lai, E.-M. (2008).

Secretome analysis uncovers an Hcp-family protein secreted via a type VI secretion system in Agrobacterium tumefaciens. J Bacteriol *190*, 2841–2850.

Yamamoto, S., Kasai, H., Arnold, D.L., Jackson, R.W., Vivian, A., and Harayama, S. (2000). Phylogeny of the genus Pseudomonas: intrageneric structure reconstructed from the nucleotide sequences of gyrB and rpoD genes. Microbiology (Reading, Engl) *146 ( Pt 10)*, 2385–2394.

Yang, F.-J., Cheng, L.-L., Zhang, L., Dai, W.-J., Liu, Z., Yao, N., Xie, Z.-P., and Staehelin, C. (2009). Y4lO of Rhizobium sp. strain NGR234 is a symbiotic determinant required for symbiosome differentiation. J Bacteriol *191*, 735–746.

Yang, S., Tang, F., Gao, M., Krishnan, H.B., and Zhu, H. (2010). R gene-controlled host specificity in the legume-rhizobia symbiosis. Proc Natl Acad Sci USA *107*, 18735–18740.

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol *17*, 32–43.

Yoshimura, T., Jhee, K.H., and Soda, K. (1996). Stereospecificity for the hydrogen transfer and molecular evolution of pyridoxal enzymes. Biosci. Biotechnol. Biochem. *60*, 181–187.

Young, J.M., Park, D.-C., Shearman, H.M., and Fargier, E. (2008). A multilocus sequence analysis of the genus Xanthomonas. Syst Appl Microbiol *31*, 366–377.

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.

Zamioudis, C., and Pieterse, C.M.J. (2012). Modulation of host immunity by beneficial microbes. Mol Plant Microbe Interact *25*, 139–150.

Zehner, S., Schober, G., Wenzel, M., Lang, K., and Göttfert, M. (2008). Expression of the Bradyrhizobium japonicum Type III Secretion System in Legume Nodules and Analysis of the Associated tts box Promoter. Mol Plant Microbe Interact *21*, 1087–1093.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res *18*, 821–829.

Zhang, L., Chen, X.-J., Lu, H.-B., Xie, Z.-P., and Staehelin, C. (2011). Functional analysis of the type 3 effector nodulation outer protein L (NopL) from Rhizobium sp. NGR234: symbiotic effects, phosphorylation, and interference with mitogen-activated protein kinase signaling. J Biol Chem *286*, 32178–32187.

Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G.K.-S., and Yu, J. (2006).

KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics *4*, 259–263.

Zhou, H., Morgan, R.L., Guttman, D.S., and Ma, W. (2009). Allelic variants of the Pseudomonas syringae type III effector HopZ1 are differentially recognized by plant resistance systems. Mol Plant Microbe Interact *22*, 176–189.

Zipfel, C. (2008). Pattern-recognition receptors in plant innate immunity. Curr Opin Immunol *20*, 10–16.

Zipfel, C. (2009). Early molecular events in PAMP-triggered immunity. Curr Opin Plant Biol.

Zmasek, C.M., and Eddy, S.R. (2001). ATV: display and manipulation of annotated phylogenetic trees. Bioinformatics *17*, 383–384.

# Appendix

# Appendix I: Recombineering and stable integration of the *Pseudomonas syringae* pv *syringae* 61 *hrp/hrc* cluster into the genome of the soil bacterium *Pseudomonas fluorescens* Pf0-1

William J. Thomas, Caitlin A. Thireault, Jeffrey A. Kimbrel[1], and Jeff H. Chang

The Plant Journal
2009, vol 60 (5), pp. 919-928
PMID: 19682294

**SUMMARY**

Many Gram-negative bacteria use a type III secretion system (T3SS) to establish associations with their hosts. The T3SS is a conduit for direct injection of type III effector proteins into host cells where they manipulate the host for the benefit of the infecting bacterium. For plant-associated pathogens, the variations in number and amino acid sequences of type III effectors as well as their functional redundancy make studying type III effectors challenging. To mitigate this challenge, we developed a stable delivery system for individual or defined sets of type III effectors into plant cells. We used recombineering and Tn5-mediated transposition to clone and stably integrate, respectively, the complete *hrp/hrc* region from *Pseudomonas syringae* pv *syringae* 61 into the genome of the soil bacterium *Pseudomonas fluorescens* Pf0-1. We describe our development of Effector-to-Host Analyzer (EtHAn) and demonstrate its utility for studying effectors for their *in planta* functions.

## INTRODUCTION

Pathogens must overcome plant defenses in order to successfully infect their hosts. One defense mechanism that plant pathogens encounter is basal defense, or PAMP-triggered Immunity (PTI). PTI relies on pattern recognition receptors (PRRs) to perceive microbes by their conserved molecular patterns (pathogen- or microbial-associated molecular patterns; PAMPs or MAMPs, respectively, Ausubel, 2005; Jones and Dangl, 2006; Schwessinger and Zipfel, 2008). Perception results in the induction of plant responses that include the deposition of callose into the cell walls (Schwessinger and Zipfel, 2008). From hereafter for the sake of simplicity, we use the terms PAMPs in re ference to both PAMPs and MAMPs and PTI in reference to basal defense regardless of the source of the molecular pattern being perceived by the plant.

Many Gram-negative phytopathogenic bacteria deliver type III effectors into host cells through a type III secretion system (T3SS) as countermeasures against PTI (Cunnac *et al.*, 2009). The more than 20 genes encoding the regulatory elements and structural components of the T3SS are often clustered together in bacterial genomes (Galan and Wolf-Watz, 2006). In *Pseudomonas syringae* the T3SS-encoding genes of the *hrp*/*hrc* region are located in a 26-kilobase pathogenicity island (Huang *et al.*, 1988; Alfano *et al.*, 2000; Oh *et al.*, 2007). Central to its expression is the alternative sigma factor HrpL (Xiao *et al.*, 1994). HrpL is a member of the extracytoplasmic function (ECF) family of sigma factors and activates expression of the T3SS-encoding

genes and type III effector genes by recognizing a cis-regulatory element, the *hrp*-box (Fouts *et al.*, 2002; Innes *et al.*, 1993). Their expression is usually low or repressed when cells are grown in rich media but induced to high levels when grown *in planta* or *in vitro* in *hrp*-inducing media (Huynh *et al.*, 1989; Rahme *et al.*, 1992; Xiao *et al.*, 1992).

Type III effectors of phytopathogens are collectively necessary for the bacteria to cause disease. A T3SS-deficient mutant is incapable of causing disease on its normally compatible host (Lindgren *et al.*, 1986; Niepold *et al.*, 1985). Moreover, studies of transgenic plants expressing individual type III effectors provide strong evidence that their functions are to dampen host PTI (Hauck *et al.*, 2003; Nomura *et al.*, 2006; Fu *et al.*, 2007; Underwood *et al.*, 2007; Shan *et al.*, 2008; Xiang *et al.*, 2008).

Some type III effectors can also betray the presence of the infecting bacterium to the host. Plants encode disease resistance proteins (R) that can perceive a single or limited number of cognate type III effector proteins (DeYoung and Innes, 2006; Jones and Dangl, 2006). Perception triggers a defense response that can be viewed as amplified PTI. This effector-triggered immunity (ETI) is frequently associated with visual evidence of a programmed cell death response called a hypersensitive response (HR; Greenberg and Yao, 2004).

Germane to the work described herein are three ETI-elicitors, AvrRpt2, AvrRpm1, and HopQ1-1. AvrRpt2 and AvrRpm1 are type III effectors of *P.*

*syringae* that perturb the *Arabidopsis* protein RIN4. The R proteins RPS2 and RPM1, respectively, perceive their corresponding modifications to RIN4 to elicit ETI (Mackey *et al.*, 2002; Axtell and Staskawicz, 2003; Mackey *et al.*, 2003). HopQ1-1 of *P. syringae* pv *tomato* DC3000 (*Pto*DC3000) elicits ETI in tobacco and is a negative host range determinant; its deletion from the genome enables *Pto*DC3000 to grow significantly more than the wild-type strain in tobacco (Wei *et al.*, 2007).

Type III effectors are challenging to study for many reasons. There is a great diversity of effectors even between strains of the same species. A draft genome sequence of the pathogen *P. syringae* pv *tomato* T1 (*Pto*T1) was completed and compared to the completed genome sequence of *Pto*DC3000 (Almeida *et al.*, 2009). Both are pathogens of tomato and their inventories of genes had a high degree of conservation. In striking contrast, the number of homologous type III effector genes common to both strains was much lower. This diversity in type III effector collections is likely in response to the selective pressures host defenses impose on pathogens (Jones and Dangl, 2006).

Deletion or overexpression of type III effector genes in pathogens does not often result in changes in phenotype (Chang *et al.*, 2004). This reflects the observation that pathogens have collections of type III effectors with overlapping functions (Kvitko *et al.*, 2009). The overlap in functions of homologous and especially non-homologous type III effectors is thus a significant impediment in genetic-based studies of their functions.

*P. fluorescens* 55 carrying the pHIR11 cosmid or its derivatives, has been a workhorse heterologous delivery system and a valuable resource for studying type III effectors (Kim *et al.*, 2002; Jamir *et al.*, 2004; Schechter *et al.*, 2004; Fujikawa *et al.*, 2006). This cosmid clone has the entire T3SS-encoding region of *P. syringae* pv *syringae* 61 (*Psy*61) and also a chaperone and its cognate type III effector gene, *shcA* and *hopA1* (aka *hopPsyA*) as well as 1.6 kb of the type III effector gene *avrE1* (Huang *et al.*, 1988). HopA1 elicits effector-triggered immunity (ETI) in tobacco and the Ws-0 cultivar of *Arabidopsis* (Jamir *et al.*, 2004; Gassmann, 2005). The presence of *hopA1* complicates the characterization of defined sets of type III effectors and it has been disrupted via marker-exchange mutagenesis (Fouts *et al.*, 2003; Jamir *et al.*, 2004).

One limitation to pHIR11 or its variants is that bacterial cells do not stably maintain them in the absence of antibiotic selection. After 24 hours of growth *in planta*, only 50% of a derivative of *Pto*DC3000 still carried pCPP2071, a derivative of pHIR11 (Fouts *et al.*, 2003). By four days, only 4% of the cells still had the cosmid. We have observed ~90% loss from *P. fluorescens* grown overnight in culture in the absence of selection (Unrath and Chang, unpublished). This rapid loss of the T3SS-encoding locus potentially limits the utility of pHIR11 and its derivatives for studying type III effectors *in planta*.

To address the limitations in stability and potential conflict with antibiotic resistance markers, we have developed a new system to deliver individual or defined sets of type III effector proteins directly into host cells. We used recombineering to transfer the *hrp*/*hrc* region from *Psy*61 cloned in a cosmid vector, into a mini-*Tn5* Tc plasmid (de Lorenzo *et al.*, 1990; Court *et al.*, 2002; Jamir *et al.*, 2004; Oh *et al.*, 2007). Recombineering is catalyzed by bacteriophage-encoded systems that inhibit bacteria RecBCD nucleases from degrading double-stranded linear DNA fragments and can be used to precisely delete, insert, or alter a DNA sequence (Court *et al.*, 2002; Sharan *et al.*, 2009). We then stably integrated the *hrp*/*hrc* region directly into the genome of *P. fluorescens* Pf0-1 to ensure a robust delivery system (Silby *et al.*, 2009). We present our development of Effector to Host Analyzer (EtHAn) and validate its use for characterizing type III effectors.

## RESULTS

### Development of a stable type III effector delivery system.

We used a variety of PCR methods and recombineering to clone the *hrp/hrc* cluster from cosmid clone pLN18 into a mini-*Tn5* Tc vector (Fig. 1; de Lorenzo *et al.*, 1990; Court *et al.*, 2002; Jamir *et al.*, 2004). The genes we cloned spanned *hrpK* to *hrpH* and we will refer to this as the T3SS-encoding region. Despite evidence that HrpH may be injected directly into host cells, we purposefully included *hrpH* because its product is necessary for efficient translocation of other type III effectors (Oh *et al.*, 2007). Our design avoided

inclusion of genes encoding *shcA-hopA1* and included <300 bp of the *avrE* coding region. We further modified the mini-*Tn5* Tc by replacing the $Tet^R$ gene with a $Kan^R$ gene flanked by site-specific recombinase recognition sequences (FRT) to facilitate excision of the antibiotic marker from the genome.

We used Tn5 to mediate the direct and stable integration of the T3SS-encoding region into the genome of the soil bacterium *P. fluorescens* Pf0-1. We selected this strain for two reasons. The genome sequence of Pf0-1 has been completed and searches have failed to confidently identify a T3SS-encoding region or candidate type III effector genes (Ma *et al.*, 2003; Grant *et al.*, 2006; Silby *et al.*, 2009). Secondly, Pf0-1 is a soil bacterium not adapted for survival within a plant and is therefore expected to elicit PTI and be incapable of dampening PTI or eliciting ETI when injected directly into plants (Compeau *et al.*, 1988).

We also used the site-specific recombinase FLP to excise the $Kan^R$ gene from the genome of Pf0-1. Through these manipulations, we developed EtHAn (Effector to Host Analyzer), an unmarked, PTI-eliciting bacterium capable of delivering individual or defined sets of type III effector proteins directly into cells of plants.

***P. fluorescens* Pf0-1 and the modified EtHAn cannot grow *in planta*.**

To validate our selection of Pf0-1 as the recipient strain for the T3SS-encoding region, we first determined its growth behavior *in planta* (Fig. 2). We infiltrated 1.0 x $10^6$ cfu/ml of wild-type Pf0-1 and EtHAn into leaves of

*Arabidopsis* and enumerated their growth. The compatible pathogen *Pto*DC3000 grew extensively *in planta* and its T3SS mutant (Δ*hrcC*) failed to grow significantly. Pf0-1 and EtHAn failed to grow even to levels comparable to the Δ*hrcC* mutant. EtHAn expressing the T3SS appears to grow even less than Pf0-1, but we note that this is more likely a reflection of differences in infiltrated bacterial concentrations (at day 0, EtHAn had lower cfu than Pf0-1). We also enumerated growth of EtHAn following inoculation with higher concentrations of bacteria and had similar results to those presented (data not shown). These results validated our selection of Pf0-1 as the recipient for the T3SS-encoding region and show that the T3SS is, by itself, insufficient to confer virulence.

**EtHAn carries a functional T3SS.**

To demonstrate that EtHAn encodes a functional T3SS, we tested its ability to deliver type III effectors *in planta*. Delivery is easily detected through elicitation of ETI where perception of a type III effector protein by a corresponding R protein leads to a rapid hypersensitive response (HR). We tested whether EtHAn carrying either *avrRpm1* or *avrRpt2* would elicit an HR in *Arabidopsis* Col-0. Their protein products are perceived by corresponding R proteins, RPM1 and RPS2, leading to a visible HR approximately 6 hpi and 20 hpi, respectively (Kunkel *et al.*, 1993; Yu *et al.*, 1993; Bisgrove *et al.*, 1994).

We infiltrated 1.0 x $10^8$ cfu/ml of EtHAn carrying *avrRpm1* or *avrRpt2* into leaves of *Arabidopsis* (Fig. 3a). At 6 hpi, the majority of leaves infected

with *Pto*DC3000 or EtHAn carrying *avrRpm1*, showed visible HR phenotypes. By 20 hpi, the majority of the leaves infected with *Pto*DC3000 or EtHAn carrying *avrRpt2* had collapsed. By 28 hpi, with the exception of EtHAn carrying an empty vector, all infected leaves had collapsed. This latter observation emphasizes one additional advantage of using EtHAn as opposed to wild-type *Pto*DC3000 for studying ETI. Because Pf0-1 is not adapted for survival within a plant, it is unable to elicit any visible symptoms in *Arabidopsis*, unlike wild-type *Pto*DC3000, which causes massive tissue collapse because of disease.

We next asked whether EtHAn carrying a single type III effector gene would gain virulence. The type III effector AvrPto has global effects on perturbing host PTI because it interacts with PAMP receptors to dampen plant defense responses (Shan *et al.*, 2008; Xiang *et al.*, 2008). Transgenic plants overexpressing AvrPto support more growth of the Δ*hrcC* mutant of *Pto*DC3000 (Hauck *et al.*, 2003). These transgenics are also severely compromised in responding to PAMPs as assayed by enumerating callose deposition (Hauck *et al.*, 2003). Finally, analysis of microarray data for host transcriptional changes suggested that the changes caused by AvrPto account for the majority of changes caused by wild-type *Pto*DC3000 delivering its entire set of type III effectors (Hauck *et al.*, 2003). We therefore determined if expression of AvrPto in EtHAn would be sufficient to confer virulence to a soil bacterium (Fig. 2). EtHAn carrying *avrPto* did not have any significant increase

in growth relative to EtHAn carrying an empty vector. Thus, our growth enumeration studies not surprisingly indicated that a single type III effector protein was not sufficient to confer virulence to a soil bacterium.

**EtHAn expresses the T3SS and type III effectors to high levels.**

Our design of EtHAn relied on HrpL-controlled expression of the T3SS and its type III effector genes. Because proteins encoded by Pf0-1 are necessary to activate proteins upstream of HrpL, there is potential that EtHAn will not express type III effector genes to native levels as in *P. syringae*. We used quantitative real-time PCR (qRT-PCR) to measure the relative expression levels of a T3SS-encoding gene, *hrcV* and the type III effector gene *avrPto*. To ensure that expression levels were reflective of transcriptional regulation and not copy number of the type III effector gene, we used Tn7 to integrate a single copy of full-length *avrPto* into an intergenic region of the genome of Pf0-1 (Chang *et al.*, 2005; Peters and Craig, 2001).

As expected, *hrcV* and *avrPto* were expressed to high levels in *Pto*DC3000 seven hours after shift to *hrp*-inducing media but were not significantly expressed in the Δ*hrpL* mutant (Fig. 4a). EtHAn and EtHAn + *avrPto* expressed *hrcV* and *avrPto* in the latter strain, to higher levels than wild-type *Pto*DC3000 (Fig. 4b). Note that expression is presented as relative to *Pto*DC3000; i.e., we would have expected a normalized fold expression value of one if EtHAn expressed the HrpL-regulated genes to levels similar to *Pto*DC3000. By 24 hours after shift, expression of both genes in *Pto*DC3000

was still detectable and significant but far less than their levels at seven hours after shift to *hrp*-inducing media. In contrast, expression of both genes continued to increase in EtHAn and EtHAn + *avrPto*, respectively. We also noticed that expression of both genes was higher in cells grown in KB media than in *Pto*DC3000 (data not shown). Together, these results indicate that HrpL-regulated genes are not under the same negative control as they are in *P. syringae*. Regardless of this observation, results indicate that EtHAn expresses HrpL-regulated genes to sufficient levels as compared to *P. syringae*.

**Type III effectors delivered by EtHAn are sufficient to dampen PTI.**

Next, we determined whether EtHAn is suitable for studying the virulence functions of type III effector proteins. Several type III effectors, including AvrPto and HopM1 have demonstrable roles in significantly suppressing the deposition of callose when expressed directly in transgenic plants (Hauck *et al.*, 2003; Nomura *et al.*, 2006; Fu *et al.*, 2007; Underwood *et al.*, 2007). We therefore infiltrated leaves of *Arabidopsis* with EtHAn carrying *avrPto* or *hopM1* and enumerated the number of callose deposits (Fig. 5). We also enumerated the number of callose deposits in leaves infiltrated with *Pto*DC3000, its *ΔhrcC* mutant, and EtHAn carrying an empty vector as controls. Representative leaf pictures are presented (Fig. 5b).

As expected, the Δ*hrcC* mutant that is incapable of delivering type III effectors was unable to suppress callose deposition whereas *Pto*DC3000

significantly suppressed callose deposition. Infiltration of leaves with EtHAn without any type III effector genes resulted in even more callose deposits than the Δ*hrcC* mutant. In contrast, EtHAn carrying *avrPto* or *hopM1* significantly and reproducibly suppressed the deposition of callose as compared to EtHAn carrying an empty vector. These results indicate that EtHAn by itself elicits PTI and that type III effectors delivered by EtHAn are sufficient to dampen PTI. Thus, EtHAn is a suitable system for studying the virulence functions of type III effector proteins *in planta*.

**Delivery of type III effectors can be generalized.**

The possibility that Pf0-1 cannot deliver all *P. syringae* type III effectors is unlikely given the observation that *P. fluorescens* 55 carrying pLN18 was sufficient to deliver all tested type III effectors (Jamir *et al.*, 2004; Schechter *et al.*, 2004). Regardless, we tested approximately one half of the type III effectors from *Pto*DC3000 for delivery by EtHAn. The type III effectors were expressed as single copy genes integrated via Tn7, and from their native promoters as translational fusions to *Δ79AvrRpt2* (Guttman *et al.*, 2002; Chang *et al.*, 2005). The tested type III effectors AvrE1, AvrPto1, HopAF1, HopAB2, HopAM1-1, HopC1, HopE1, HopI1, HopP1, HopQ1-1, HopX1, HopY1, (ShcA1)-HopA1, (SchF2)-HopF2, and (SchO1)-HopO1 were all sufficient for EtHAn to elicit an HR in *Arabidopsis* encoding the corresponding *R* gene, *Rps2*. For type III effectors expressed from operons, we also included

their upstream genes (included in parentheses). We therefore concluded that EtHAn can deliver most if not all type III effectors of *P. syringae*.

**EtHAn by itself elicits a defense response in species other than *Arabidopsis*.**

Because wild-type Pf0-1 does not encode any of its own type III effector genes, we reasoned that EtHAn has potential applications for studying effectors in hosts other than *Arabidopsis*. We assayed the effects of EtHAn in leaves of tomato and *N. tabacum*. EtHAn by itself elicited a response in leaves of tomato (data not shown) and a spotty, inconsistent response in leaves of *N. tabacum* (Fig. 3b). This response was specific to EtHAn; tomato or tobacco infiltrated with *P. fluorescens* Pf0-1 had no responses (data not shown). These results suggest the T3SS-encoding locus encodes a protein that may elicit a defense response in tomato or tobacco.

Nevertheless, we asked if EtHAn could deliver type III effectors into leaves of *N. tabacum*. *Pto*DC3000 is not compatible with tobacco because it delivers the perceived type III effector HopQ1-1 and elicits ETI (Wei *et al.*, 2007). We therefore mobilized *hopQ1-1* into EtHAn and infiltrated the strain into leaves of tobacco. After ~24 hpi, *Pto*DC3000 elicited a robust HR. The compatible pathogen *P. syringae* pv *tabaci* 11528, in contrast elicited strong tissue collapse due to disease whereas its corresponding T3SS mutant (*ΔhrcV*) did not elicit a phenotype (data not shown). EtHAn carrying *hopQ1-1* elicited a strong HR similar to *Pto*DC3000 and more robust as compared to EtHAn alone.

**DISCUSSION**

Our goal is to understand the mechanisms by which a single pathogen uses its entire collection of type III effector proteins to dampen host defenses. One difficulty is that pathogens can deliver more than thirty different type III effectors, of which many share overlapping functions (Kvitko *et al.*, 2009). The redundancy within a collection of type III effectors is likely a consequence of the need to ensure robustness so that loss of single type III effectors will not compromise virulence. However, this redundancy is clearly a significant challenge that we must overcome in order to understand the contributions of each type III effector during host-association.

To address this hurdle, we developed a stable system for delivering type III effector proteins into host cells. We moved a defined fragment of ~26 kb spanning a region necessary to regulate and assemble the T3SS, into a modified mini-*Tn5* Tc vector. We used Tn5-mediated transposition to stably integrate the T3SS-encoding region directly into the genome of the soil bacterium, *P. fluorescens* Pf0-1 to make EtHAn. We show that when lacking type III effector genes, EtHAn elicits PAMP-triggered immunity and is non-pathogenic on *Arabidopsis*.

Potentially, EtHAn may be less efficient in delivering type III effector proteins into host cells than pathogenic *Pto*DC3000 as was the case for *P.*

*fluorescens* 55 carrying pLN18 (Schechter *et al.*, 2004). We did observe variability in the robustness of the HR between different type III effector-Δ79AvrRpt2 fusions (data not shown). This however, is consistent with previous results using *Pto*DC3000 to deliver fusions with Δ79AvrRpt2 or *P. fluorescens* carrying pLN18 to deliver type III effector fusions to adenylate cyclase (Schechter *et al.*, 2004; Chang *et al.*, 2005). Furthermore, our reliance on native promoters may have contributed to variable expression. Nevertheless, when carrying type III effector genes, EtHAn expressed and translocated sufficient amounts of their proteins into host cells to elicit ETI or dampen the PTI-associated deposition of callose. Therefore, EtHAn is a sufficient elicitor of PAMP-triggered immunity that can subsequently be engineered to study the roles of delivered type III effectors in perturbing host defenses.

EtHAn offers several advantages over currently used methods for studying plant-pathogen interactions. Because EtHAn is decorated with molecular patterns and can be injected into the apoplastic space of plants, it can be used directly to study host defenses without the need to introduce additional PAMPs. We selected *P. fluorescens* Pf0-1 for modification because it appears to be devoid of any endogenous type III effector genes as determined by surveying its completed genome sequence for homologous DNA sequences (Grant *et al.*, 2006). Further, Pf0-1 appears to lack most, if not all, necessary virulence factors required for growth *in planta*. Thus, any

observed phenotypes can be attributed to the delivered type III effector protein of interest. Additionally, because the T3SS-encoding region is stably integrated and the type III effector genes can be as well, EtHAn can be used to characterize type III effector proteins *in planta* over the course of days.

EtHAn has applications beyond characterizing type III effector proteins of *P. syringae*. Fusions between oomycete effectors and the amino-terminal domain of AvrRps4 or AvrRpm1 are delivered by *Pto*DC3000 directly into plant cells via its T3SS (Sohn *et al.*, 2007; Rentel *et al.*, 2008). However, in most cases the virulence functions of the oomyete effector will be masked by the functions conferred by the collection of type III effectors normally delivered by *Pto*DC3000 (Sohn *et al.*, 2007). In contrast, since EtHAn is apparently devoid of virulence factors, the functions for effectors are more apt to be observed. EtHAn has been successfully used to deliver the *Hyaloperonospora parasitica* effectors ATR1 and ATR13 fused to AvrRpm1 directly into *Arabidopsis* (Brian Staskawicz, personal communication). Furthermore, *P. fluorescens* 55 carrying pHIR11 can deliver type III effectors from *Xanthomonas* spp (Fujikawa *et al.*, 2006). Though we did not explicitly reconfirm this finding we expect EtHAn should behave similarly. EtHAn therefore has potential applications for characterizing virulence functions of other bacterial as well as fungal and oomycete effectors.

EtHAn has strong potential for studying type III effector functions in *Arabidopsis*. EtHAn by itself does not elicit any symptoms in 88 different

accessions of *Arabidopsis* (Qingli Lu, Marc Nishimura, and Jeffery Dangl, personal communication). Most notably, EtHAn does not elicit an HR in the Ws-0 accession, which is further evidence that our recombineering of the T3SS-encoding locus avoided inclusion of the type III effector gene *hopA1* (Gassmann, 2005).

In contrast however, in its current form, EtHAn may have limited use in other plant species. Our results suggest that a factor encoded by the T3SS-encoding locus elicits a weaker non-host defense response in tomato or tobacco. Alternatively, but less likely is the possibility that the phenotype elicited by EtHAn is a consequence of the *Tn*5-mediated integration event into the genome of Pf0-1. Regardless of the cause, our observations are not in agreement with published reports showing that *P. fluorescens* Pf-55 carrying pHIR11 elicits an HR in tobacco, while the same strain carrying pLN18 (disruption in *hopA1*), does not (Jamir *et al.*, 2004; Schechter *et al.*, 2004). Therefore, we speculate that EtHAn elicits a defense response in tobacco because the T3SS is stably integrated and expressed at higher levels than in pathogenic *P. syringae*.

Our results highlight the utility of recombineering for manipulating DNA fragments. This method has been used to delete genes directly from the genome of the plant pathogen, *Erwinia amylovora* but we have not had success with recombineering directly in *P. syringae* (Zhao *et al.*, 2009; Chang, unpublished). Nevertheless, we demonstrate that recombineering can be used

to alter large and otherwise challenging to manipulate fragments of DNA in *E. coli*. Other potential applications include modifying large genomic clones or plasmids such as binary vectors (Rozwadowski *et al.*, 2008).

To request EtHAn, please visit our webpage at: http://changlab.cgrb.oregonstate.edu/.

## EXPERIMENTAL PROCEDURES

### Bacterial strains, plant lines and growth conditions.

The bacterial strains used in this study are *P. syringae* pv *tomato* DC3000, its T3SS structural mutant (Δ*hrcC*; Yuan and He, 1996), its T3SS regulatory mutant, (Δ*hrpL*; Zwiesler-Vollick *et al.*, 2002), and *P. fluorescens* Pf0-1 (Silby *et al.*, 2009), *Escherichia coli* DH5α and HB101λpir. Pseudomonads were grown at 28°C in King's B (KB) liquid media with shaking or on KB agar plates. For *in vitro* induction of HrpL-regulated genes, Pseudomonads were grown overnight in KB media, washed, and resuspended at $OD_{600}$ = 0.1 in *hrp*-inducing media and grown for 7 or 24 hours (Huynh et al., 1989). *E. coli* were grown at 37°C in Luriae-Bertani (LB) liquid media with shaking or on LB agar plates. Antibiotics were used at the final concentrations of: 25 μg/ml rifampicin, 30 μg/ml kanamycin (100 μg/ml for Pf0-1), 30 μg/ml chloramphenicol, 5 μg/ml tetracycline (50 μg/ml for Pf0-1), 25 μg/ml gentamycin (100 μg/ml for Pf0-1) and 100 μg/ml ampicillin. Concentrations listed were for *P. syringae* and *E. coli* unless otherwise noted.

*Arabidopsis thaliana* accession Col-0 and *Nicotiana tabacum* were grown in a controlled-environment growth chamber (9 hrs of day at 22°C, 15 hrs of night at 20°C) for 5~6 weeks.

**Plasmid constructions.**

Plasmids used were pBBR1-MCS1, -MCS2, and -MCS5 (Kovach *et al.*, 1994; Kovach *et al.*, 1995), pRK2013 (Figurski and Helinski, 1979), mini-*Tn5* Tc (de Lorenzo *et al.*, 1990), pKD4 and pKD46 (Datsenko and Wanner, 2000), pBH474 (House *et al.*, 2004), pME3280a (Zuber *et al.*, 2003), pUX-BF13 (Bao *et al.*, 1991), and pLN18 (Jamir *et al.*, 2004). Oligonucleotides used are listed in supplementary table 1. All restriction enzymes were purchased from New England Biolabs (NEB; Ipswich, MA).

We first used recombinant and sticky-end PCR to fuse together two 0.5 kb-sized fragments flanking the T3SS-encoding region carried on pLN18 (Jamir *et al.*, 2004). A 0.5 kb fragment immediately downstream of *hrpK* was amplified in two separate reactions using Pfu and primer pairs JHC124 + JHC125 or JHC148 + JHC150 (fragments 1 and 1' of figure 1, respectively). A 0.5 kb fragment immediately upstream of *hrpH* (aka ORF1 of Conserved Effector Locus; Alfano *et al.*, 2000) was amplified in two separate reactions with primer pairs JHC126 + JHC128 or JHC151 + JHC128 (fragments 2 and 2', respectively). The products were gel-purified. Fragments 1 + 2 and 1' + 2' were mixed in approximately equal ratios, and each amplified in two separate reactions using Pfu and primer pairs JHC123 + JHC125 and JHC124 +

JHC127, or JHC148 + JHC128 and JHC149 + JHC152, respectively. Recombined products carried an inserted unique *Xba*I or *Acc*65I site, respectively. Corresponding ~1.0 kb products were mixed in approximately equal ratios, incubated at 95°C for 5 minutes, and slowly cooled to room temperature leading to a proportion of the PCR products with overhanging ends compatible to recipient vectors.

Sticky-end products derived from the 1 + 2 fusion were gel-purified and cloned into pBBR1-MCS5 cleaved with *Eco*RI + *Xho*I and subsequently subcloned into pBBR1-MCS1 cleaved with *Sac*I + *Xho*I (Kovach *et al.*, 1994; Kovach *et al.*, 1995). Sticky-end products from the 1' + 2' fusion were gel-purified and cloned into mini-*Tn5* Tc cleaved with *Not*I and treated with calf-intestinal phosphatase (CIP).

**Recombineering.**

We digested pBBR1-MCS1 carrying the 1 + 2 fusion with *Xba*I and transformed the gel-purified, linear fragment into electrocompetent cells made from arabinose-induced (10 mM) DH5$\alpha$ cells carrying pKD46 and pLN18 (de Lorenzo *et al.*, 1990; Jamir *et al.*, 2004). Transformants were selected on chloramphenicol. Successful transfer of the entire ~26 kb T3SS-encoding region was confirmed by restriction digestion and PCR with pairs of primers that amplified different fragments along the length of the T3SS-encoding region.

We next cleaved mini-*Tn5* Tc carrying the 1' + 2' fusion with *Acc*65I and transformed the gel-purified, linear fragment into electrocompetent cells made from arabinose-induced (10 mM) HB101λpir cells carrying pKD46 and pBBR1-MCS1 with the T3SS-encoding region. Transformants were selected on tetracycline. Successful transfer of the T3SS-encoding region was confirmed by restriction digestion and PCR with pairs of primers that amplified different fragments along the length of the T3SS-encoding region.

We also used recombineering to replace the $Tet^R$ gene of mini-*Tn5* Tc vector with $Kan^R$ flanked by site-specific recombinase FRT sequences. We used a two-step PCR to amplify $Kan^R$ flanked by FRT sites from pKD4 with CAT0001 and CAT0002 and subsequently with CAT0003 and CAT0004 to include 50 bp of homology to either side of the $Tet^R$ gene of mini-*Tn5* Tc (de Lorenzo *et al.*, 1990). Recombineering was done as described above. Transformants were selected on kanamycin and successful recombineering was confirmed via PCR and lack of growth on LB + Tet.

**Plasmid mobilization.**

Plasmids were mobilized into recipients as previously described (Chang *et al.*, 2005). For mobilizing *Tn*7-based vectors, cells were mixed at a ratio of 10:1:1:1 with the latter being *E. coli* carrying pUX-BF13 (Bao *et al.*, 1991). Integration of genes into the genome of *P. fluorescens* Pf0-1 was confirmed using PCR. Eviction of $Kan^R$ was mediated by pBH474, which encodes the FLP site-specific recombinase and confirmed by replica plating on KB agar

plates with and without kanamycin (House *et al.*, 2004). Finally, cells resistant to 5% sucrose were selected to identify those that lost pBH474.

**Quantitative Real Time PCR (qRT-PCR).**

Cells were collected and immediately suspended in RNAprotect (Qiagen, Valencia, CA), and stored at -80°C. RNA was extracted using RNeasy according to the instructions of the manufacturer (Qiagen, Valencia, CA), and subsequently treated with DNase I (NEB, Ipswich, MA). Qualities of each RNA preparations were assessed on 1.0 X FA, 1.2% agarose gels. We measured the quantity of RNA using a Nanodrop ND-1000 (Thermo Scientific, Wilmington, DE).

One μg of total RNA from each sample was used to synthesize single stranded cDNA according to the instructions of the manufacturer (Superscript III; Invitrogen, Carlsbad, CA). We used reverse transcriptase PCR to determine the quality of the cDNA. We used oligonucleotides 23S-T and 23S-B to amplify 23S rRNA from 1.0 ng of each sample. We also used hrpEF-T and hrpEF-B oligonucleotides in separate reactions to confirm the complete removal of DNA from each sample. These oligonucleotides span the intergenic region of separately transcribed genes *hrpE* and *hrpF* (Rahme *et al.*, 1991).

One ng of single-stranded cDNA was used in SYBR-Green qRT-PCR on a CFX96 real-time machine (BioRad, Hercules, CA). Oligonucleotides hrcV-T and hrcV-B as well as avrPto-T and avrPto-B were used to amplify portions of *hrcV* and *avrPto*, respectively. All reactions were done in triplicate

and normalized to 23S rRNA. To calculate efficiencies for each primer pair, we did qRT-PCR on 1.0 ng, 0.1 ng, and 0.01 ng of single-stranded cDNA made from *Pto*DC3000 grown in *hrp*-inducing media for 24 hrs. Efficiency-corrected $2^{-\Delta\Delta Ct}$ values were determined using the CFX Manager software (Biorad; Pfaffl, 2001).

***In planta* assay.**

Bacterial cells were grown overnight in KB with appropriate antibiotics, washed and resuspended in 10 mM $MgCl_2$. For *in planta* growth assay, we resuspended *P. syringae* and *P. fluorescens* to an $OD_{600}$ of 0.002 (1.0 x $10^6$ cfu/ml). We used 1.0 ml syringes lacking needles to infiltrate bacterial suspensions into the abaxial side of leaves of 5 ~ 6 week-old plants. We used a number 2 cork borer to core four discs for each triplicate of each treatment at 0 and 3 days post inoculation (dpi). Leaf discs were ground to homogeneity in 10 mM $MgCl_2$, serially diluted, and plated on KB with appropriate antibiotics**.** Colonies were enumerated once visible. Experiments were repeated at least three times. For HR assays, cells were infiltrated in leaves at an $OD_{600}$ of 0.2 (equivalent to 1.0 x $10^8$ cfu/ml). Phenotypes were scored starting at six hours post inoculation (hpi) and examined up until disease symptoms were visible.

**Callose Staining.**

We collected leaves 7 hpi, and cleared them of chlorophyll by heating at 65°C in 70% EtOH. Leaves were then rinsed and stained in aniline blue staining solution (150 mM $K_2HPO_4$, 0.1% aniline blue, pH 8.4) overnight. Leaves were placed in 70% glycerol containing 1.0% of the aniline blue

staining solution and mounted onto microscope slides. We viewed the leaves using a light microscope (BX40, Olympus) with a filter under ultraviolet 10X magnification. Digital images of 10 fields per leaf were taken (Magnafire, Optronics) and 15 leaves were imaged per treatment. We used a custom Perl script to enumerate callose deposits.

## ACKNOWLEDGEMENTS

We thank Mark Silby and Stuart Levy for *P. fluorescens* Pf0-1, Brenda Schroeder and Michael Kahn for pBH454, Dieter Haas for pME3280a and pUX-BF13, Sheng Yang He for Δ*hrcC*, Brian Staskawicz for Δ*hrpL*, and Alan Collmer for pLN18 and the Δ*hrcV* mutant of *Pta* 11528. We thank Jeffery Dangl, Jason Cumbie, Rebecca Pankow, and Philip Hillebrand for their assistance and critical reading of this manuscript. We gratefully acknowledge Jim Carrington for use of his light microscope. Finally, we thank Ethan S. Chang for his assistance and inspiring the naming of the type III effector delivery system. This research was supported in part by start-up funds from Oregon State University to JHC and a grant from the National Research Initiative of the USDA Cooperative State Research, Education, and Extension Service (Grant 2008-35600-18783).

## REFERENCES

**Alfano, J.R., Charkowski, A.O., Deng, W.L., Badel, J.L., Petnicki-Ocwieja, T., van Dijk, K. and Collmer, A.** (2000) The *Pseudomonas syringae Hrp*

pathogenicity island has a tripartite mosaic structure composed of a cluster of type III secretion genes bounded by exchangeable effector and conserved effector loci that contribute to parasitic fitness and pathogenicity in plants. *Proc. Natl. Acad. Sci. USA*, **97**, 4856-4861.

**Almeida, N.F., Yan, S., Lindeberg, M., Studholme, D.J., Schneider, D.J., Condon, B., Liu, H., Viana, C.J., Warren, A., Evans, C., Kemen, E., Maclean, D., Angot, A., Martin, G.B., Jones, J.D., Collmer, A., Setubal, J.C. and Vinatzer, B.A.** (2009) A draft genome sequence of *Pseudomonas syringae* pv. *tomato* T1 reveals a type III effector repertoire significantly divergent from that of *Pseudomonas syringae* pv. *tomato* DC3000. *Mol. Plant-Microbe Interact.*, **22**, 52-62.

**Ausubel, F.M.** (2005) Are innate immune signaling pathways in plants and animals conserved? *Nature immunology*, **6**, 973-979.

**Axtell, M.J. and Staskawicz, B.J.** (2003) Initiation of RPS2-specified disease resistance in *Arabidopsis* is coupled to the AvrRpt2-directed elimination of RIN4. *Cell*, **112**, 369-377.

**Bao, Y., Lies, D.P., Fu, H. and Roberts, G.P.** (1991) An improved *Tn*7-based system for the single-copy insertion of cloned genes into chromosomes of Gram-negative bacteria. *Gene*, **109**, 167-168.

**Bisgrove, S.R., Simonich, M.T., Smith, N.M., Sattler, A. and Innes, R.W.** (1994) A disease resistance gene in *Arabidopsis* with specificity for two different pathogen avirulence genes. *Plant Cell*, **6**, 927-933.

**Chang, J.H., Goel, A.K., Grant, S.R. and Dangl, J.L.** (2004) Wake of the flood: ascribing functions to the wave of type III effector proteins of phytopathogenic bacteria. *Curr Opin Microbiol*, **7**, 11-18.

**Chang, J.H., Urbach, J.M., Law, T.F., Arnold, L.W., Hu, A., Gombar, S., Grant, S.R., Ausubel, F.M. and Dangl, J.L.** (2005) A high-throughput, near-saturating screen for type III effector genes from *Pseudomonas syringae*. *Proc. Natl. Acad. Sci. USA*, **102**, 2549-2554.

**Compeau, G., Al-Achi, B.J., Platsouka, E. and Levy, S.B.** (1988) Survival of rifampin-resistant mutants of *Pseudomonas fluorescens* and *Pseudomonas putida* in soil systems. *Appl. Environ. Microbiol.*, **54**, 2432-2438.

**Court, D.L., Sawitzke, J.A. and Thomason, L.C.** (2002) Genetic engineering using homologous recombination. *Annu. Rev. Genet.*, **36**, 361-388.

**Cunnac, S., Lindeberg, M. and Collmer, A.** (2009) *Pseudomonas syringae* type III secretion system effectors: repertoires in search of functions. *Curr. Opin. Microbiol.*, **12**, 53-60.

**Datsenko, K.A. and Wanner, B.L.** (2000) One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. USA*, **97**, 6640-6645.

**de Lorenzo, V., Herrero, M., Jakubzik, U. and Timmis, K.N.** (1990) Mini-*Tn*5 transposon derivatives for insertion mutagenesis, promoter probing, and chromosomal insertion of cloned DNA in gram-negative eubacteria. *J. Bacteriol.*, **172**, 6568-6572.

**DeYoung, B.J. and Innes, R.W.** (2006) Plant NBS-LRR proteins in pathogen sensing and host defense. *Nature immunology*, **7**, 1243-1249.

**Figurski, D.H. and Helinski, D.R.** (1979) Replication of an origin-containing derivative of plasmid RK2 dependent on a plasmid function provided in trans. *Proc. Natl. Acad. Sci. USA*, **76**, 1648-1652.

**Fouts, D.E., Abramovitch, R.B., Alfano, J.R., Baldo, A.M., Buell, C.R., Cartinhour, S., Chatterjee, A.K., D'Ascenzo, M., Gwinn, M.L., Lazarowitz, S.G., Lin, N.C., Martin, G.B., Rehm, A.H., Schneider, D.J., van Dijk, K., Tang, X. and Collmer, A.** (2002) Genomewide identification of *Pseudomonas syringae* pv. *tomato* DC3000 promoters controlled by the HrpL alternative sigma factor. *Proc. Natl. Acad. Sci. USA*, **99**, 2275-2280.

**Fouts, D.E., Badel, J.L., Ramos, A.R., Rapp, R.A. and Collmer, A.** (2003) A *Pseudomonas syringae* pv. *tomato* DC3000 *Hrp* (Type III secretion) deletion mutant expressing the *Hrp* system of bean pathogen *P. syringae* pv. *syringae* 61 retains normal host specificity for tomato. *Mol. Plant-Microbe Interact.*, **16**, 43-52.

**Fu, Z.Q., Guo, M., Jeong, B.R., Tian, F., Elthon, T.E., Cerny, R.L., Staiger, D. and Alfano, J.R.** (2007) A type III effector ADP-ribosylates RNA-binding proteins and quells plant immunity. *Nature*, **447**, 284-288.

**Fujikawa, T., Ishihara, H., Leach, J.E. and Tsuyumu, S.** (2006) Suppression of defense response in plants by the avrBs3/pthA gene family of *Xanthomonas* spp. *Mol. Plant Microbe Interact.*, **19**, 342-349.

**Galan, J.E. and Wolf-Watz, H.** (2006) Protein delivery into eukaryotic cells by type III secretion machines. *Nature*, **444**, 567-573.

**Gassmann, W.** (2005) Natural variation in the *Arabidopsis* response to the avirulence gene *hopPsyA* uncouples the hypersensitive response from disease resistance. *Mol. Plant-Microbe Interact.*, **18**, 1054-1060.

**Grant, S.R., Fisher, E.J., Chang, J.H., Mole, B.M. and Dangl, J.L.** (2006) Subterfuge and Manipulation: Type III Effector Proteins of Phytopathogenic Bacteria. *Annu. Rev. Microbiol.*, **60**, 425-449.

**Greenberg, J.T. and Yao, N.** (2004) The role and regulation of programmed cell death in plant-pathogen interactions. *Cell. Microbiol.*, **6**, 201-211.

**Guttman, D.S., Vinatzer, B.A., Sarkar, S.F., Ranall, M.V., Kettler, G. and Greenberg, J.T.** (2002) A functional screen for the type III (Hrp) secretome of the plant pathogen *Pseudomonas syringae*. *Science*, **295**, 1722-1726.

**Hauck, P., Thilmony, R. and He, S.Y.** (2003) A *Pseudomonas syringae* type III effector suppresses cell wall-based extracellular defense in susceptible *Arabidopsis* plants. *Proc. Natl. Acad. Sci. USA*, **100**, 8577-8582.

**House, B.L., Mortimer, M.W. and Kahn, M.L.** (2004) New recombination methods for *Sinorhizobium meliloti* genetics. *Appl. Environ. Microbiol.*, **70**, 2806-2815.

**Huang, H.C., Schuurink, R., Denny, T.P., Atkinson, M.M., Baker, C.J., Yucel, I., Hutcheson, S.W. and Collmer, A.** (1988) Molecular cloning of a *Pseudomonas syringae* pv. *syringae* gene cluster that enables *Pseudomonas fluorescens* to elicit the hypersensitive response in tobacco plants. *J. Bacteriol.*, **170**, 4748-4756.

**Huynh, T.V., Dahlbeck, D. and Staskawicz, B.J.** (1989) Bacterial blight of soybean: regulation of a pathogen gene determining host cultivar specificity. *Science*, **245**, 1374-1377.

**Innes, R.W., Bent, A.F., Kunkel, B.N., Bisgrove, S.R. and Staskawicz, B.J.** (1993) Molecular analysis of avirulence gene *avrRpt2* and identification of a putative regulatory sequence common to all known *Pseudomonas syringae* avirulence genes. *J. Bacteriol.*, **175**, 4859-4869.

**Jamir, Y., Guo, M., Oh, H.S., Petnicki-Ocwieja, T., Chen, S., Tang, X., Dickman, M.B., Collmer, A. and Alfano, J.R.** (2004) Identification of *Pseudomonas syringae* type III effectors that can suppress programmed cell death in plants and yeast. *Plant J.*, **37**, 554-565.

**Jones, J.D. and Dangl, J.L.** (2006) The plant immune system. *Nature*, **444**, 323-329.

**Kim, Y.J., Lin, N.C. and Martin, G.B.** (2002) Two distinct *Pseudomonas* effector proteins interact with the Pto kinase and activate plant immunity. *Cell*, **109**, 589-598.

**Kovach, M.E., Elzer, P.H., Hill, D.S., Robertson, G.T., Farris, M.A., Roop, R.M., 2nd and Peterson, K.M.** (1995) Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes. *Gene*, **166**, 175-176.

**Kovach, M.E., Phillips, R.W., Elzer, P.H., Roop, R.M., 2nd and Peterson, K.M.** (1994) pBBR1MCS: a broad-host-range cloning vector. *Biotechniques*, **16**, 800-802.

**Kunkel, B.N., Bent, A.F., Dahlbeck, D., Innes, R.W. and Staskawicz, B.J.** (1993) RPS2, an *Arabidopsis* disease resistance locus specifying recognition of *Pseudomonas syringae* strains expressing the avirulence gene *avrRpt2*. *Plant Cell*, **5**, 865-875.

**Kvitko, B.H., Park, D.H., Velasquez, A.C., Wei, C.F., Russell, A.B., Martin, G.B., Schneider, D.J. and Collmer, A.** (2009) Deletions in the repertoire of *Pseudomonas syringae* pv. *tomato* DC3000 type III secretion effector genes reveal functional overlap among effectors. *PLoS pathogens*, **5**, e1000388.

**Lindgren, P.B., Peet, R.C. and Panopoulos, N.J.** (1986) Gene cluster of *Pseudomonas syringae* pv. "*phaseolicola*" controls pathogenicity of bean plants and hypersensitivity of nonhost plants. *J. Bacteriol.*, **168**, 512-522.

**Ma, Q., Zhai, Y., Schneider, J.C., Ramseier, T.M. and Saier, M.H., Jr.** (2003) Protein secretion systems of *Pseudomonas aeruginosa* and *P. fluorescens*. *Biochimica et biophysica acta*, **1611**, 223-233.

**Mackey, D., Belkhadir, Y., Alonso, J.M., Ecker, J.R. and Dangl, J.L.** (2003) *Arabidopsis* RIN4 is a target of the type III virulence effector AvrRpt2 and modulates RPS2-mediated resistance. *Cell*, **112**, 379-389.

**Mackey, D., Holt, B.F., 3rd, Wiig, A. and Dangl, J.L.** (2002) RIN4 interacts with *Pseudomonas syringae* type III effector molecules and is required for RPM1-mediated resistance in *Arabidopsis*. *Cell*, **108**, 743-754.

**Niepold, F., Anderson, D. and Mills, D.** (1985) Cloning determinants of pathogenesis from *Pseudomonas syringae* pathovar *syringae*. *Proc. Natl. Acad. Sci. USA*, **82**, 406-410.

**Nomura, K., Debroy, S., Lee, Y.H., Pumplin, N., Jones, J. and He, S.Y.** (2006) A bacterial virulence protein suppresses host innate immunity to cause plant disease. *Science*, **313**, 220-223.

**Oh, H.S., Kvitko, B.H., Morello, J.E. and Collmer, A.** (2007) *Pseudomonas syringae* lytic transglycosylases coregulated with the type III secretion system contribute to the translocation of effector proteins into plant cells. *J. Bacteriol.*, **189**, 8277-8289.

**Peters, J.E. and Craig, N.L.** (2001) *Tn*7: smarter than we thought. *Nat. Rev. Mol. Cell Biol.*, **2**, 806-814.

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

**Rahme, L.G., Mindrinos, M.N. and Panopoulos, N.J.** (1991) Genetic and transcriptional organization of the *hrp* cluster of *Pseudomonas syringae* pv. *phaseolicola*. *J. Bacteriol.*, **173**, 575-586.

**Rahme, L.G., Mindrinos, M.N. and Panopoulos, N.J.** (1992) Plant and environmental sensory signals control the expression of *hrp* genes in *Pseudomonas syringae* pv. *phaseolicola*. *J. Bacteriol.*, **174**, 3499-3507.

**Rentel, M.C., Leonelli, L., Dahlbeck, D., Zhao, B. and Staskawicz, B.J.** (2008) Recognition of the *Hyaloperonospora parasitica* effector ATR13 triggers resistance against oomycete, bacterial, and viral pathogens. *Proc. Natl. Acad. Sci. USA*, **105**, 1091-1096.

**Rozwadowski, K., Yang, W. and Kagale, S.** (2008) Homologous recombination-mediated cloning and manipulation of genomic DNA regions using Gateway and recombineering systems. *BMC biotechnology*, **8**, 88.

**Schechter, L.M., Roberts, K.A., Jamir, Y., Alfano, J.R. and Collmer, A.** (2004) *Pseudomonas syringae* type III secretion system targeting signals and novel effectors studied with a Cya translocation reporter. *J Bacteriol.*, **186**, 543-555.

**Schwessinger, B. and Zipfel, C.** (2008) News from the frontline: recent insights into PAMP-triggered immunity in plants. *Curr. Opin. Plant Biol.*, **11**, 389-395.

**Shan, L., He, P., Li, J., Heese, A., Peck, S.C., Nurnberger, T., Martin, G.B. and Sheen, J.** (2008) Bacterial effectors target the common signaling partner BAK1 to disrupt multiple MAMP receptor-signaling complexes and impede plant immunity. *Cell host & microbe*, **4**, 17-27.

**Sharan, S.K., Thomason, L.C., Kuznetsov, S.G. and Court, D.L.** (2009) Recombineering: a homologous recombination-based method of genetic engineering. *Nature protocols*, **4**, 206-223.

**Silby, M.W., Cerdeno-Tarraga, A.M., Vernikos, G.S., Giddens, S.R., Jackson, R.W., Preston, G.M., Zhang, X.X., Moon, C.D., Gehrig, S.M., Godfrey, S.A., Knight, C.G., Malone, J.G., Robinson, Z., Spiers, A.J., Harris, S., Challis, G.L., Yaxley, A.M., Harris, D., Seeger, K., Murphy, L., Rutter, S., Squares, R., Quail, M.A., Saunders, E., Mavromatis, K., Brettin, T.S., Bentley, S.D., Hothersall, J., Stephens, E., Thomas, C.M., Parkhill, J., Levy, S.B., Rainey, P.B. and Thomson, N.R.** (2009) Genomic and genetic analyses of diversity and plant interactions of *Pseudomonas fluorescens*. *Genome Biol*, **10**, R51.

**Sohn, K.H., Lei, R., Nemri, A. and Jones, J.D.** (2007) The downy mildew effector proteins ATR1 and ATR13 promote disease susceptibility in *Arabidopsis thaliana*. *Plant Cell*, **19**, 4077-4090.

**Underwood, W., Zhang, S. and He, S.Y.** (2007) The *Pseudomonas syringae* type III effector tyrosine phosphatase HopAO1 suppresses innate immunity in *Arabidopsis thaliana*. *Plant J.*, **52**, 658-672.

**Wei, C.F., Kvitko, B.H., Shimizu, R., Crabill, E., Alfano, J.R., Lin, N.C., Martin, G.B., Huang, H.C. and Collmer, A.** (2007) A *Pseudomonas syringae* pv. *tomato* DC3000 mutant lacking the type III effector HopQ1-1 is able to cause disease in the model plant *Nicotiana benthamiana*. *Plant J.*, **51**, 32-46.

**Xiang, T., Zong, N., Zou, Y., Wu, Y., Zhang, J., Xing, W., Li, Y., Tang, X., Zhu, L., Chai, J. and Zhou, J.M.** (2008) *Pseudomonas syringae* effector AvrPto blocks innate immunity by targeting receptor kinases. *Curr. Biol.*, **18**, 74-80.

**Xiao, Y., Heu, S., Yi, J., Lu, Y. and Hutcheson, S.W.** (1994) Identification of a putative alternate sigma factor and characterization of a multicomponent regulatory cascade controlling the expression of *Pseudomonas syringae* pv. *syringae* Pss61 *hrp* and *hrmA* genes. *J. Bacteriol.*, **176**, 1025-1036.

**Xiao, Y., Lu, Y., Heu, S. and Hutcheson, S.W.** (1992) Organization and environmental regulation of the *Pseudomonas syringae* pv. *syringae* 61 hrp cluster. *J. Bacteriol.*, **174**, 1734-1741.

**Yu, G.L., Katagiri, F. and Ausubel, F.M.** (1993) *Arabidopsis* mutations at the RPS2 locus result in loss of resistance to *Pseudomonas syringae* strains expressing the avirulence gene *avrRpt2*. *Mol. Plant-Microbe Interact.*, **6**, 434-443.

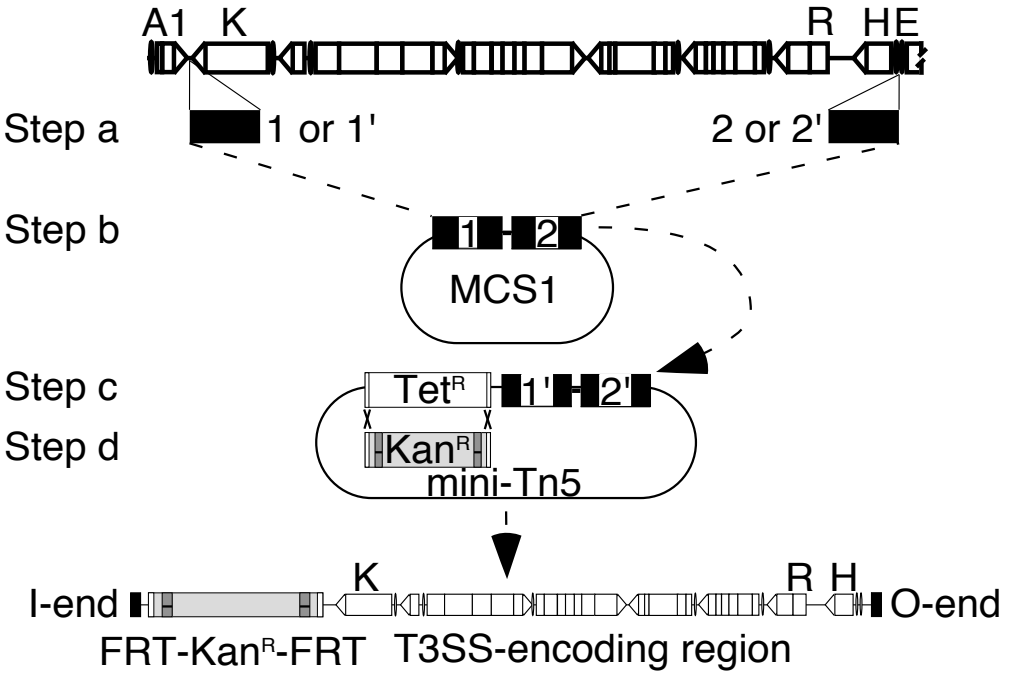
**Yuan, J. and He, S.Y.** (1996) The *Pseudomonas syringae Hrp* regulation and secretion system controls the production and secretion of multiple extracellular proteins. *J. Bacteriol*, **178**, 6399-6402.

**Zhao, Y., Sundin, G.W. and Wang, D.** (2009) Construction and analysis of pathogenicity island deletion mutants of *Erwinia amylovora*. *Canadian journal of microbiology*, **55**, 457-464.

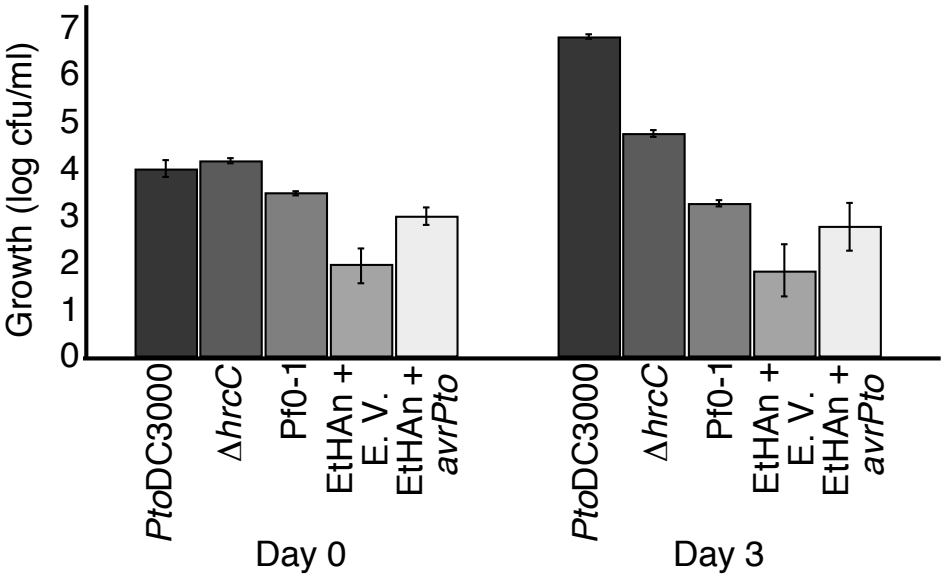
**Zuber, S., Carruthers, F., Keel, C., Mattart, A., Blumer, C., Pessi, G., Gigot-Bonnefoy, C., Schnider-Keel, U., Heeb, S., Reimmann, C. and Haas, D.** (2003) GacS sensor domains pertinent to the regulation of exoproduct formation and to the biocontrol potential of *Pseudomonas fluorescens* CHA0. *Mol. Plant-Microbe Interact.*, **16**, 634-644.

**Zwiesler-Vollick, J., Plovanich-Jones, A.E., Nomura, K., Bandyopadhyay, S., Joardar, V., Kunkel, B.N. and He, S.Y.** (2002) Identification of novel *hrp-*

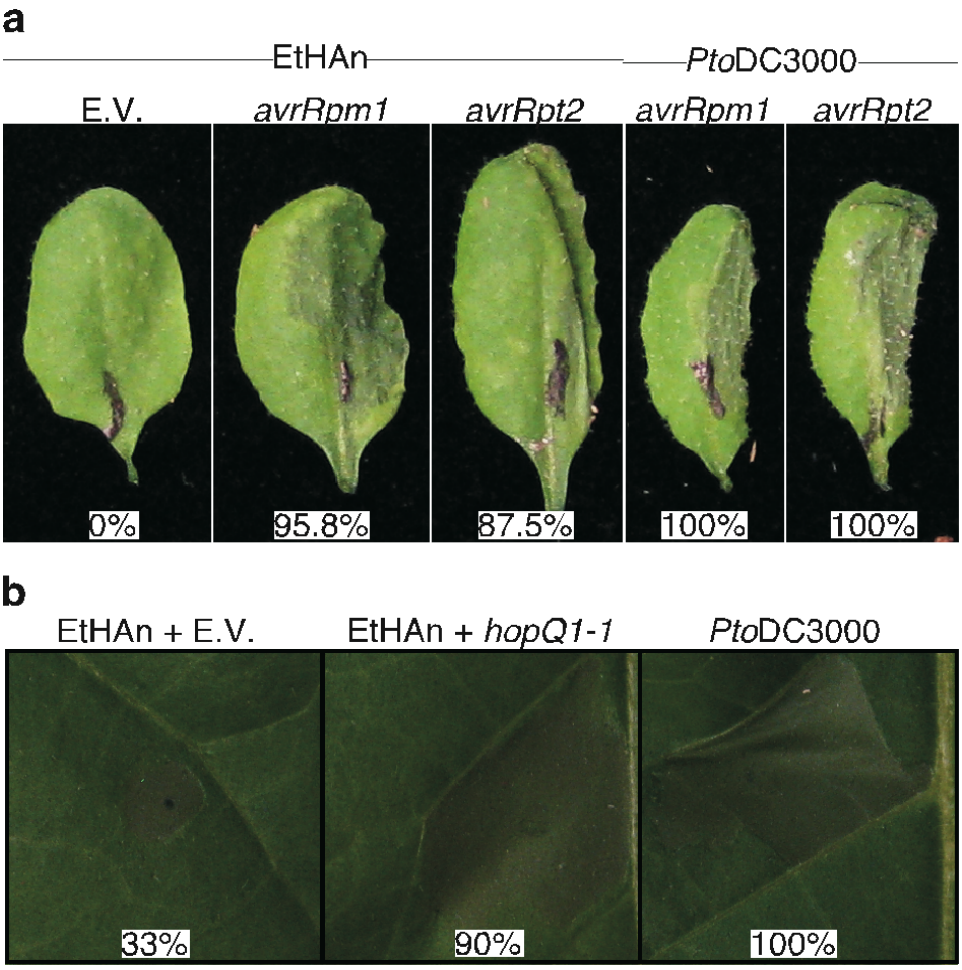
regulated genes through functional genomic analysis of the *Pseudomonas syringae* pv. *tomato* DC3000 genome. *Mol. Microbiol*, **45**, 1207-1218.



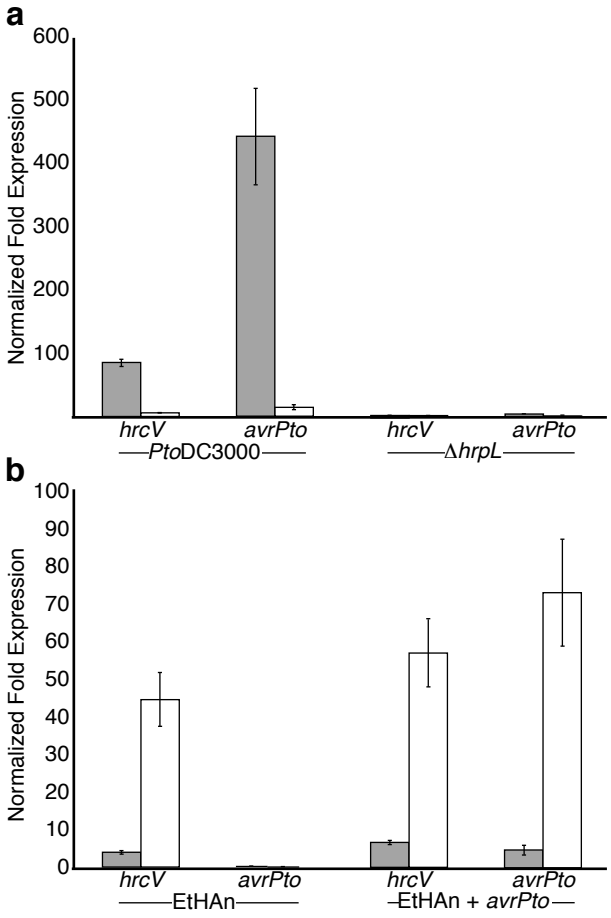

**Appendix I, Figure 1. Construction of EtHAn.**
The top bar represents the cloned insert of pLN18. Boxes represent protein-coding regions and triangles indicate their direction of transcription. Black vertical discs represent *hrp*-boxes. Some protein-coding regions are labeled for landmarks; abbreviations are: A1 = *shcA1-hopA1*, K = *hrpK*, R = *hrpR*, H = *hrpH*, and E = *avrE1*. The *avrE1* gene is truncated in pLN18 (broken line). **Step a:** Two 0.5 kb fragments (black boxes; 1, 1', 2 and 2') flanking the T3SS-encoding region were amplified and recombined. The 1 + 2 and 1' + 2' recombined products were cloned into pBBR1-MCS1 (MCS1) or mini-*Tn5* Tc, respectively. **Steps b** and **c:** Vectors containing recombined flanking DNA fragments were linearized by restriction digestion and used to sequentially capture the T3SS-encoding region by recombineering, first into pBBR1-MCS1 (step b) then to mini-*Tn5* Tc (step c). **Step d:** Recombineering was used to replace the $Tet^R$ gene with the $Kan^R$ gene flanked by FRT sites (gray bars with a horizontal dashed line). The resulting DNA fragment between the *Tn5*-repeats (I and O-ends) is shown at the bottom. The orientation of the T3SS-encoding region relative to the *Tn5* repeats is unknown and arbitrarily presented in one direction. The entire fragment was stably integrated into the genome of *P. fluorescens* Pf0-1 via *Tn5* transposition and the $Kan^R$ gene was excised using the site-specific recombinase FLP (not shown).



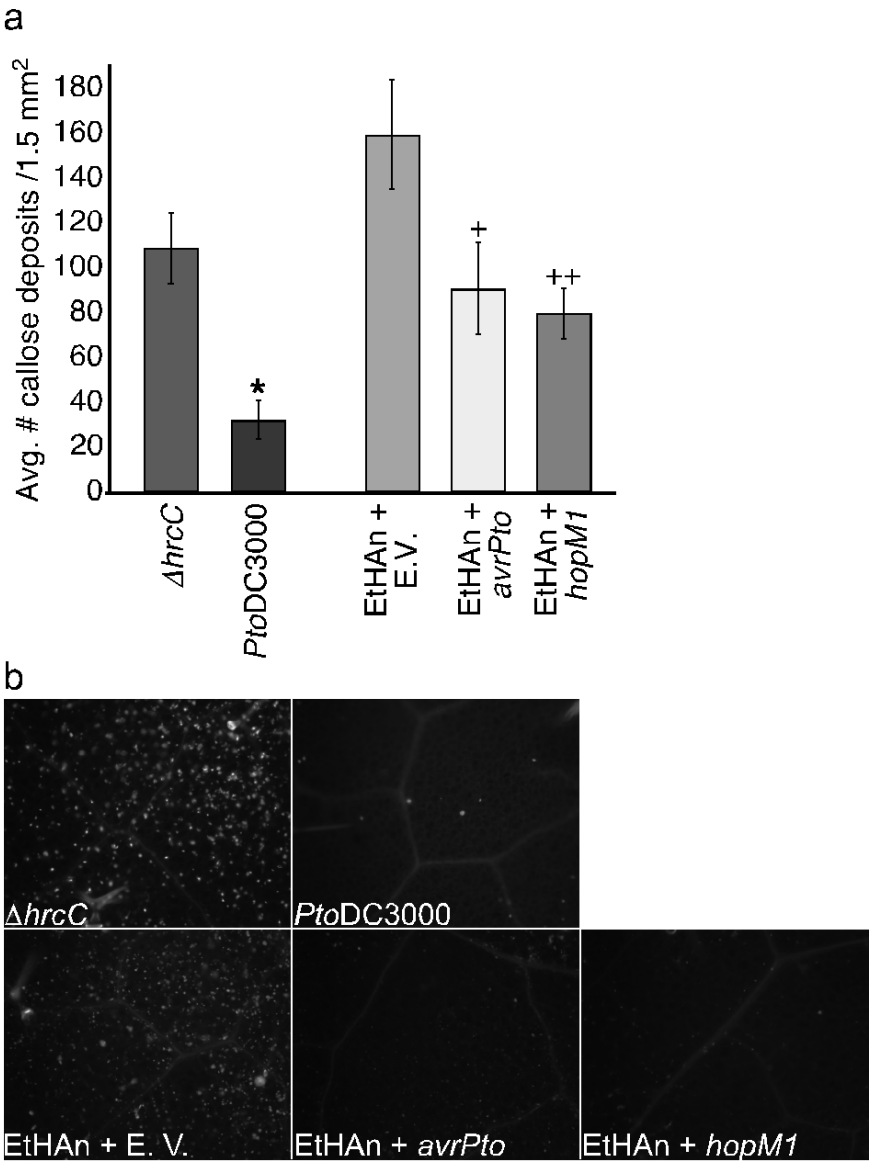

**Appendix I, Figure 2. Enumeration of bacterial growth *in planta*.**
Wild-type *Pto*DC3000, a Δ*hrcC* mutant, Pf0-1, EtHAn carrying an empty vector (E.V.), and EtHAn carrying *avrPto* were assessed for growth in *Arabidopsis* over a period of 3 dpi. Each time point was measured in triplicate; standard errors are presented. Experiments were repeated at least three times with similar results.




**Appendix I, Figure 3. EtHAN has a functional type III secretion system.**
**a)** EtHAn and *Pto*DC3000 carrying *avrRpm1* and *avrRpt2* elicited a strong HR in *Arabidopsis* Col-0. Twenty-four leaves were infiltrated for each strain and the percentages of responding leaves are presented at the bottom of each panel. Experiments were repeated at least three times with similar results. Black lines denote infiltrated leaves. **b)** EtHAn carrying full-length *hopQ1-1* elicited an HR in *N. tabacum*. EtHAn by itself elicited a spotty and inconsistent HR. Eighteen leaves were infiltrated for each strain and the percentages of responding leaves are shown. This experiment was repeated three times with similar results.




**Appendix I, Figure 4. Expression of HrpL-regulated genes in EtHAn.**
**a)** Normalized fold expression of *hrcV* and *avrPto* in *Pto*DC3000 and Δ*hrpL* were measured using the efficiency corrected $2^{-\Delta\Delta Ct}$ qRT-PCR method. Gene expression was normalized to 23S rRNA and calculated relative to corresponding genes of cells grown in KB media. Expression was assessed at 7 (gray bars) and 24 (white bars) hours after shift to *hrp*-inducing media. **b)** Normalized fold expression of *hrcV* and *avrPto* in EtHAn or EtHAn carrying *avrPto* were measured at 7 (gray bars) and 24 (white bars) hours after shift to *hrp*-inducing media. Expression was normalized to 23S rRNA and calculated relative to corresponding genes in *Pto*DC3000 grown in *hrp*-inducing media (a). All genes were measured in triplicate. Standard error is shown.




**Appendix I, Figure 5. EtHAn carrying *avrRpm1* or *hopM1* dampens the callose response.**
**a)** Average number of callose deposits per field of view (1.5 $mm^2$). Fifteen leaves were photographed per treatment with ten fields photographed per leaf. Significance of differences between averages were determined using Student's *t*-test (*Pto*DC3000 versus *hrcC* or EtHAn carrying *avrPto* or *hopM1* versus empty vector control,). $^{*,+}$ Denotes a p-value < 0.05; $^{++}$denotes a p-value < 0.01. This experiment was repeated multiple times with similar results. **b)** Representative gray-scaled images of *Arabidopsis* leaves stained with aniline blue for callose deposition.

# Appendix II: Genome sequencing and analysis of the model grass *Brachypodium distachyon*

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, Tice H, Grimwood J, McKenzie N, Huo N, Anderson OD, Febrer M, Idziak D, Hasterok R, Lindquist E, Wang M, Fox SE, Priest HD, Filichkin SA, Givan SA, Bryant DW, Chang JH, Wu H, Wu W, Hsia AP, Schnable PS, Kalyanaraman A, Barbazuk B, Michael TP, Hazen SP, Bragg JN, Laudencia-Chingcuanco D, Weng Y, McKenzie N, Haberer G, Spannagl M, Mayer K, Rattei T, Mitros T, Lee SJ, Rose JK, Mueller LA, York TL, Wicker T, Buchmann JP, Tanskanen J, Schulman AH, Gundlach H, Wright J, de Oliveira AC, Maia Lda C, Belknap W, Gu YQ, Jiang N, Lai J, Zhu L, Ma J, Sun C, Pritham E, Salse J, Murat F, Abrouk M, Bruggmann R, Messing J, Dvorak J, You FM, Luo MC, Fahlgren N, Sullivan CM, Carrington JC, Chapman EJ, May GD, Zhai J, Ganssmann M, Gurazada SG, German M, Meyers BC, Green PJ, Tyler L, Wu J, Lazo GR, Thomson J, Chen S, Scheller HV, Harholt J, Ulvskov P, Kimbrel JA, Bartley LE, Cao P, Jung KH, Sharma MK, Vega-Sanchez M, Ronald P, Dardick CD, De Bodt S, Verelst W, Inzé D, Heese M, Schnittger A, Yang X, Kalluri UC, Tuskan GA, Hua Z, Vierstra RD, Cui Y, Ouyang S, Sun Q, Liu Z, Yilmaz A, Grotewold E, Sibout R, Hematy K, Mouille G, Höfte H, Pelloux J, O'Connor D, Schnable J, Rowe S, Harmon F, Cass CL, Sedbrook JC, Byrne ME, Walsh S, Higgins J, Li P, Brutnell T, Unver T, Budak H, Belcram H, Charles M, Chalhoub B, Baxter I.

Nature.
2010. Feb 11;463(7282):763-8.
PMID: 20148030

## ABSTRACT

Three subfamilies of grasses, the Ehrhartoideae, Panicoideae and Pooideae, provide the bulk of human nutrition and are poised to become major sources of renewable energy. Here we describe the genome sequence of the wild grass Brachypodium distachyon (Brachypodium), which is, to our knowledge, the first member of the Pooideae subfamily to be sequenced. Comparison of the Brachypodium, rice and sorghum genomes shows a precise history of genome evolution across a broad diversity of the grasses, and establishes a template for analysis of the large genomes of economically important pooid grasses such as wheat. The high-quality genome sequence, coupled with ease of cultivation and transformation, small size and rapid life cycle, will help Brachypodium reach its potential as an important model system for developing new energy and food crops.


## INTRODUCTION

Three subfamilies of grasses, the Ehrhartoideae, the Panicoideae, and the Pooideae provide the bulk of human nutrition and are poised to become major sources of renewable energy. Here we describe the genome sequence of the wild grass *Brachypodium distachyon* (Brachypodium), the first member of the Pooideae subfamily to be sequenced. Comparison of the Brachypodium, rice and sorghum genomes reveals a precise history of genome evolution across a broad diversity of the grasses and establishes a template for analysis of the large genomes of economically important pooid grasses such as wheat. The high quality genome sequence, coupled with its ease of cultivation and transformation,

small size and rapid life-cycle, will help Brachypodium reach its potential as an important model system for developing new energy and food crops.

Grasses provide the bulk of human nutrition, and highly productive grasses are promising sources of sustainable energy[1]. The grass family (Poaceae) comprises over 600 genera and more than 10,000 species that dominate many ecological and agricultural systems[2,3]. To date, genomic efforts have largely focused on two economically important grass subfamilies, the Ehrhartoideae (rice) and Panicoideae (maize, sorghum, sugarcane and millets). The rice[4] and sorghum[5] genome sequences and a detailed physical map of maize[6] revealed extensive conservation of gene order[5,7] and both ancient and relatively recent polyploidization.

Most cool season cereal, forage and turf grasses belong to the subfamily Pooideae, which is also the largest grass subfamily. The genomes of many pooids are characterized by daunting size and complexity. For example, the bread wheat genome is approximately 17,000 Mb and contains three independent genomes[8]. This has prohibited genome-scale comparisons spanning the three most economically important grass subfamilies.

Brachypodium, a member of the Pooideae subfamily, is a wild annual grass endemic to the Mediterranean and Middle East[9] that has promise as a model system. This has led to development of highly efficient transformation[10,11] germplasm collections[12,13,14] genetic markers[14], a genetic linkage map[15], BAC libraries[16,17], physical maps[18](MF unpublished), mutant collections[19], and databases[20] that are facilitating the use of Brachypodium by the research

community. The genome sequence described here will allow Brachypodium to serve as a powerful functional genomics resource for the grasses. It is also a major advance in grass structural genomics, permitting, for the first time, whole genome comparisons between members of the three most economically important grass subfamilies. Full Supplementary Information available at www.nature.com/.

## RESULTS

### Genome Sequence Assembly and Annotation.

Diploid inbred line Bd21[21] was sequenced using whole genome shotgun sequencing (Table S1). The 10 largest scaffolds contained 99.6% of all sequenced nucleotides (Table S2). Comparison of these 10 scaffolds with a genetic map (Figure S1) detected two false joins and created an additional seven joins to produce five pseudomolecules that spanned 272 Mb (Table S3), within the range measured by flow cytometry[22,23]. The assembly was confirmed by cytogenetic analysis (Figure S2), and alignment with two physical maps and sequenced BACs (Supplementary Data). Over 98% of ESTs mapped to the sequence assembly, consistent with a near-complete genome (Table S4 and Figure S3). Compared to other grasses, the Brachypodium genome is very compact, with retrotransposons concentrated at the centromeres and syntenic breakpoints (Figure 1). DNA transposons and derivatives are broadly distributed and primarily associated with gene-rich regions.

We analyzed small RNA populations from inflorescence tissues with deep Illumina sequencing and mapped them onto the genome sequence (Figure 2A,

Figure S4, and Table S5). Small RNA reads were most dense in regions of high repeat density, similar to the distribution reported in Arabidopsis[24]. We identified 413 and 198 21 and 24 nt phased trans-acting (ta-) siRNA loci, respectively. Using the same algorithm, only 5 ta-siRNA were identified in Arabidopsis, and none were 24 nt phased. The biological functions of these clusters of ta-siRNA, which account for a significant number of small RNAs that map outside repeat regions, are currently not known.

A total of 25,532 protein coding gene loci was predicted in the v1.0 annotation (Supplementary Information), Table S6). This is in the same range as rice (RAP2, 28,236)[25] and sorghum (v1.4, 27,640)[5], suggesting similar gene numbers across a broad diversity of grasses. Gene models were evaluated using ~10.2 Gb of Illumina RNA-seq data (Figure S5)[26]. Overall, 92.7% of predicted coding sequences (CDS) were supported by Illumina data (Figure 2B), demonstrating the high accuracy of the Brachypodium gene predictions. These gene models are available from several databases[20].

Between 77-84% of gene families (defined according to Figure S6) are shared among the three grass subfamilies represented by Brachypodium, rice and sorghum, reflecting a relatively recent common origin (Fig 2C). Grass-specific genes include transmembrane receptor protein kinases, glycosyltransferases, peroxidases and P450 proteins (Table S7B). The Pooideae-specific gene set contains only 265 gene families (Table S7C) comprising 811 genes (1400 including singletons). Genes enriched in grasses were significantly more likely to be contained in tandem arrays than random genes, demonstrating a prominent

role for tandem gene expansion in the evolution of grass-specific genes (Figure S7 and Table S8).

To validate and improve the v1.0 gene models, we manually annotated 2,755 gene models from 72 diverse gene families (Tables S9-S11) relevant to bioenergy and food crop improvement. We identified 866 genes involved in cell wall biosynthesis and 802 transcription factors from 16 families[27]. Only 13% of the gene models required modification and very few pseudogenes were identified, demonstrating the accuracy of the v1.0 annotation. Phylogenetic trees for 62 gene families were constructed using genes from rice, Arabidopsis, sorghum and poplar. In nearly all cases, Brachypodium genes had a distribution similar to rice and sorghum, demonstrating that Brachypodium is suitably generic for grass functional genomics research (Figures S8 and S9). Analysis of the predicted secretome identified substantial differences in the distribution of cell wall metabolism genes between dicots and grasses (Tables S12 and S13, Figure S10), consistent with their different cell walls[28]. Signal peptide probability curves also suggested that start codons were accurately predicted (Figure S11).

**Genome size is maintained by balancing retroelement replication and loss.**

Exhaustive analysis of transposable elements (Supplementary Information, Table S14) showed retrotransposon sequences comprise 21.4% of the genome, compared to 26% in rice, 54% in sorghum, and over 80% in wheat[29]. Thirteen retroelement sets were younger than 20,000 years, showing a recent activation compared to rice[30] (Figure S12), and a further 53 retroelement sets were less than 0.1 MY old. A minimum of 17.4 Mbp has been lost by LTR:LTR

recombination, demonstrating that retroelement expansion is countered by removal through recombination. In contrast, retroelements persist for very long times in the closely related Triticeae[30].

DNA transposons comprise 4.77% of the Brachypodium genome, within the range found in other grass genomes[5,31]. Transcriptome data and structural analysis suggest that many non-autonomous *Mariner DTT* and *Harbinger* elements recruit transposases from other families. Two *CACTA DTC* families (M and N) carried 5 non-element genes, and the *Harbinger U* family has amplified a NBS-LRR gene family (Figures S13 and S14), adding it to the group of transposable elements implicated in gene mobility[32,33]. Centromeric regions were characterized by low gene density, characteristic repeats and retroelement clusters (Figure S15). Other repeat classes are described in Table S15. Conserved non-coding sequences are described in Figure S16.

**Whole genome sequence comparison across three diverse grass genomes**

The evolutionary relationships between Brachypodium, sorghum, rice and wheat were assessed by measuring the mean synonymous substitution rates (Ks) of orthologous gene pairs (Supplementary Information, Figure S17 and Table S16), from which divergence times of Brachypodium from wheat 29.4 (±4.9) MYA (Million Years Ago), rice 42.1 (±6.9) MYA and sorghum 50.5 (±7.5) MYA (Figure 3A) were estimated. The Ks of orthologous gene pairs in the intra-genomic Brachypodium duplications (Figure 3B) suggests duplication ~65-73 MYA ago, prior to the diversification of the grasses. This is consistent with previous evolutionary histories inferred from a small numbers of genes[3,34-36].

Paralogous relationships among Brachypodium chromosomes revealed six major inter-chromosomal duplications covering 99.7% of the genome (Figure 3B) representing ancestral whole genome duplication[37]. Using the rice and sorghum genome sequences, genetic maps of barley[38] and *Aegilops tauschii* (the D genome donor of hexaploid wheat)[39], and bin-mapped wheat ESTs[40,41], 21,045 orthologous relationships between Brachypodium / rice / sorghum / Triticeae were identified (Supplementary Information). These identified 59 blocks of collinear genes covering 99.42% of the Brachypodium genome (Figures 3C, 3D and 3E). The orthologous relationships are consistent with an evolutionary scenario that shaped five Brachypodium chromosomes from a five chromosome ancestral genome *via* a 12 chromosome intermediate involving seven major chromosome fusions[41] (Figure S18). These collinear blocks of orthologous genes provide a robust and precise sequence framework for understanding grass genome evolution and aiding the assembly of sequences from other pooid grasses. We identified 14 major syntenic disruptions between Brachypodium and rice/sorghum that can be explained by nested insertions of entire chromosomes into centromeric regions (Figures 4A and 4B)[2,39,42]. Similar nested insertions in sorghum[39], and barley (Figure 4C and 4D) were also identified. Centromeric repeats and peaks in retroelements at the junctions of chromosome insertions are footprints of these insertion events (Figure S15C and Figure 1), as is higher gene density at the former distal regions of the inserted chromosomes (Figure 1). Interestingly, the reduction in chromosome number in Brachypodium and wheat

occurred independently because none of the chromosome fusions are shared by Brachypodium and the Triticeae[39] (Figure S18).

Comparisons of evolutionary rates between Brachypodium, sorghum, rice, and *Ae. tauschii* demonstrated a substantially higher rate of genome change in *Ae. tauschii* (Table S17). This may be due to retroelement activity that increases syntenic disruptions, as proposed for chromosome 5S below[43]. Among seven relatively large gene families, four were highly syntenic and two (NBS-LRR and F-box) were almost never found in syntenic order when compared to rice and sorghum (Table S18), consistent with the rapid diversification of the NBS-LRR and F-box gene families[44].

The short arm of chromosome 5 (Bd5S) has a gene density roughly half of the rest of the genome, high density LTR retrotransposon density, the youngest intact *Gypsy* elements and the lowest solo LTR density. Thus, unlike the rest of the Brachypodium genome, Bd5S is gaining retrotransposons by replication and losing fewer by recombination. Syntenic regions of rice (Os4S) and sorghum (Sb6S) demonstrate maintenance of this high repeat content for ~60-70 MY (Figure S19)[45]. Bd5S, Os4S and Sb6S also have the lowest proportion of collinear genes (Figures 4B, S19). We propose that the chromosome ancestral to Bd5S reached a tipping point where high retrotransposon density had deleterious effects on genes.

## DISCUSSION

As the first genome sequence of a pooid grass, the Brachypodium genome aids genome analysis and gene identification in the large and complex genomes of wheat and barley, two other pooid grasses that are among the world's most important crops. The very high quality of the Brachypodium genome sequence, in combination with those from two other grass subfamilies, enabled reconstruction of chromosome evolution across a broad diversity of grasses. This analysis contributes to our understanding of grass diversification by explaining how the varying chromosome numbers found in the major grass subfamilies derive from an ancestral set of five chromosomes by nested insertions of whole chromosomes into centromeres. The relatively small genome of Brachypodium contains many active retroelement families, but recombination between these keeps genome expansion in check. The short arm of chromosome 5 deviates from the rest of the genome by exhibiting a trend toward genome expansion through increased retroelement numbers and disruption of gene order more typical of larger genomes of closely related grasses.

Grass crop improvement for sustainable fuel[46] and food[47] production requires a substantial increase in research in species such as *Miscanthus*, switchgrass, wheat and cool season forage grasses. These considerations have led to the rapid adoption of Brachypodium as an experimental system for grass research. The similarities in gene content and gene family structure between Brachypodium, rice and sorghum supports the value of Brachypodium as a functional genomics model for all grasses. The Brachypodium genome sequence

analysis reported here is therefore an important advance towards securing sustainable supplies of food, feed and fuel from new generations of grass crops.

## METHODS SUMMARY

Full Methods and any associated references are available in the Supplementary Information (www.nature.com/).

Genome sequencing and assembly. Sanger sequencing was used to generate paired-end reads from 3kb, 8kb, fosmid (35kb) and BAC (100kb) clones to generate 9.4 x coverage (Table S1). The final assembly of 83 scaffolds covers 271.9 Mb (Table S3). Sequence scaffolds were aligned to a genetic map to create pseudomolecules covering each chromosome (Figures S1 and S2).

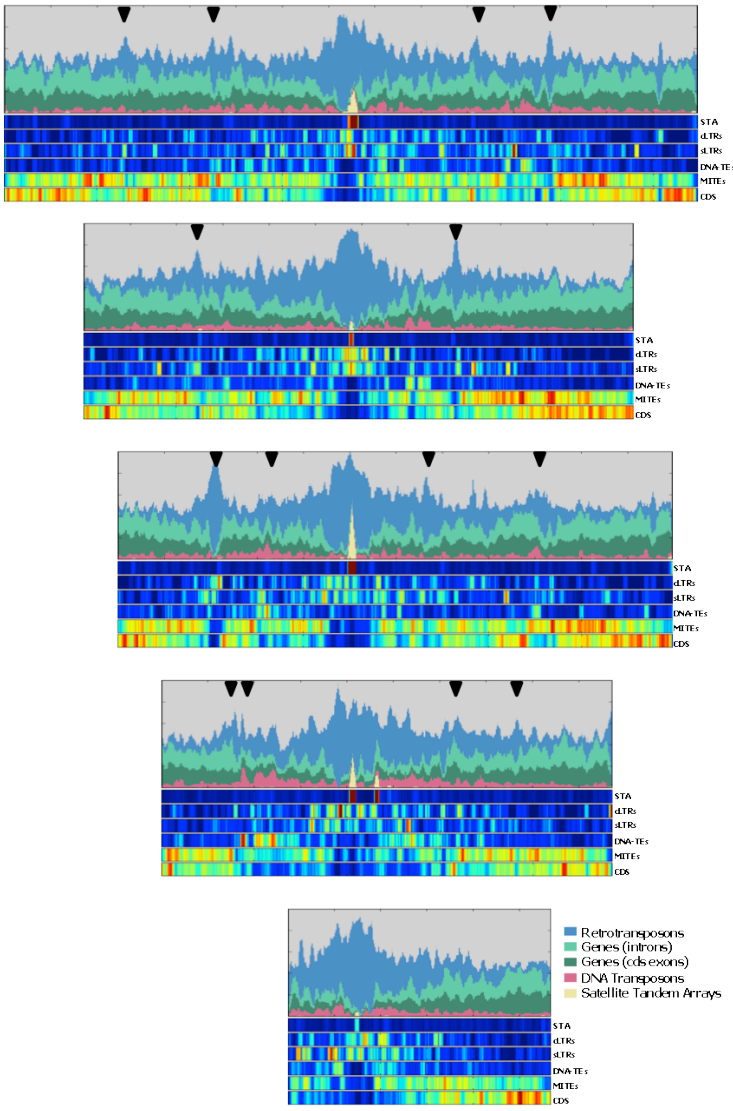
Protein coding gene annotation. Gene models were derived from weighted consensus prediction from several *ab initio* gene finders, optimal spliced alignments of ESTs and transcript assemblies, and protein homology. Illumina transcriptome sequence was aligned to predicted genome features to validate exons, splice sites and alternatively spliced transcripts.

Repeats analysis. The MIPS ANGELA pipeline was used to integrate analyses from expert groups. LTR-STRUCT and LTR-HARVEST[48] were used for *de novo* retroelement searches.

## ACKNOWLEDGEMENTS

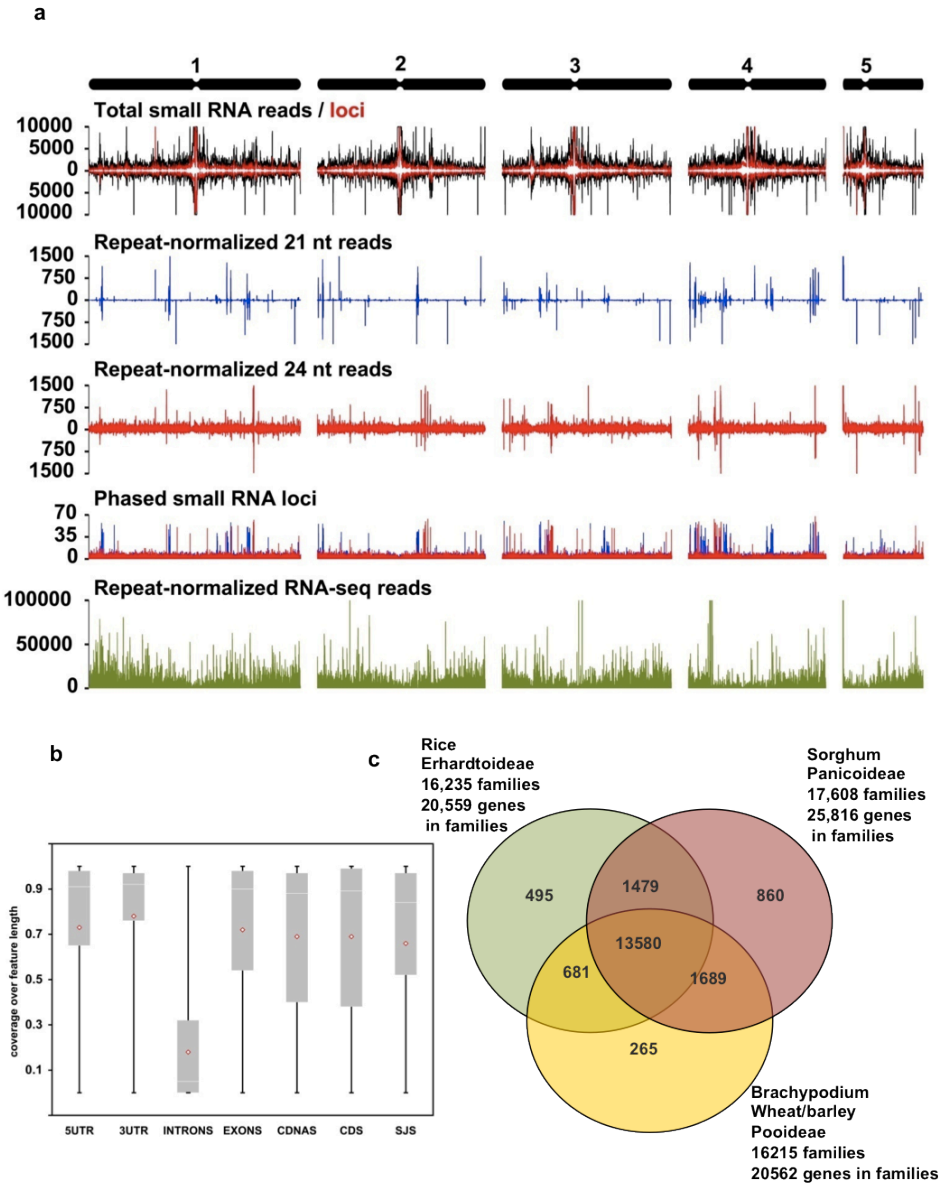
This paper is dedicated to the memory of Mike Gale, who identified the importance of conserved gene order in grass genomes. This work was mainly supported by the US Department of Energy Joint Genome Institute Community

Sequencing Program project with J.P.V, D.F.G., T.C.M. and M.W.B., a BBSRC grant to M.W.B., and EU Contract Agronomics to M.W.B. and K.F.X.M., and GABI Barlex to K.F.X.M., Illumina transcriptome sequencing was supported by a DOE Plant Feedstock Genomics for Bioenergy grant and Oregon State Agricultural Research Foundation grant to T.C.M., Small RNA research was supported by DOE Plant Feedstock Genomics for Bioenergy grants to P.J.G. and T.C.M., Annotation was supported by DOE Interagency Agreement with J.P.V., A full list of support and acknowledgements is in the Supplementary Information (www.nature.com).




**Appendix II: Figure 1. Chromosomal distribution of the main Brachypodium genome features.**

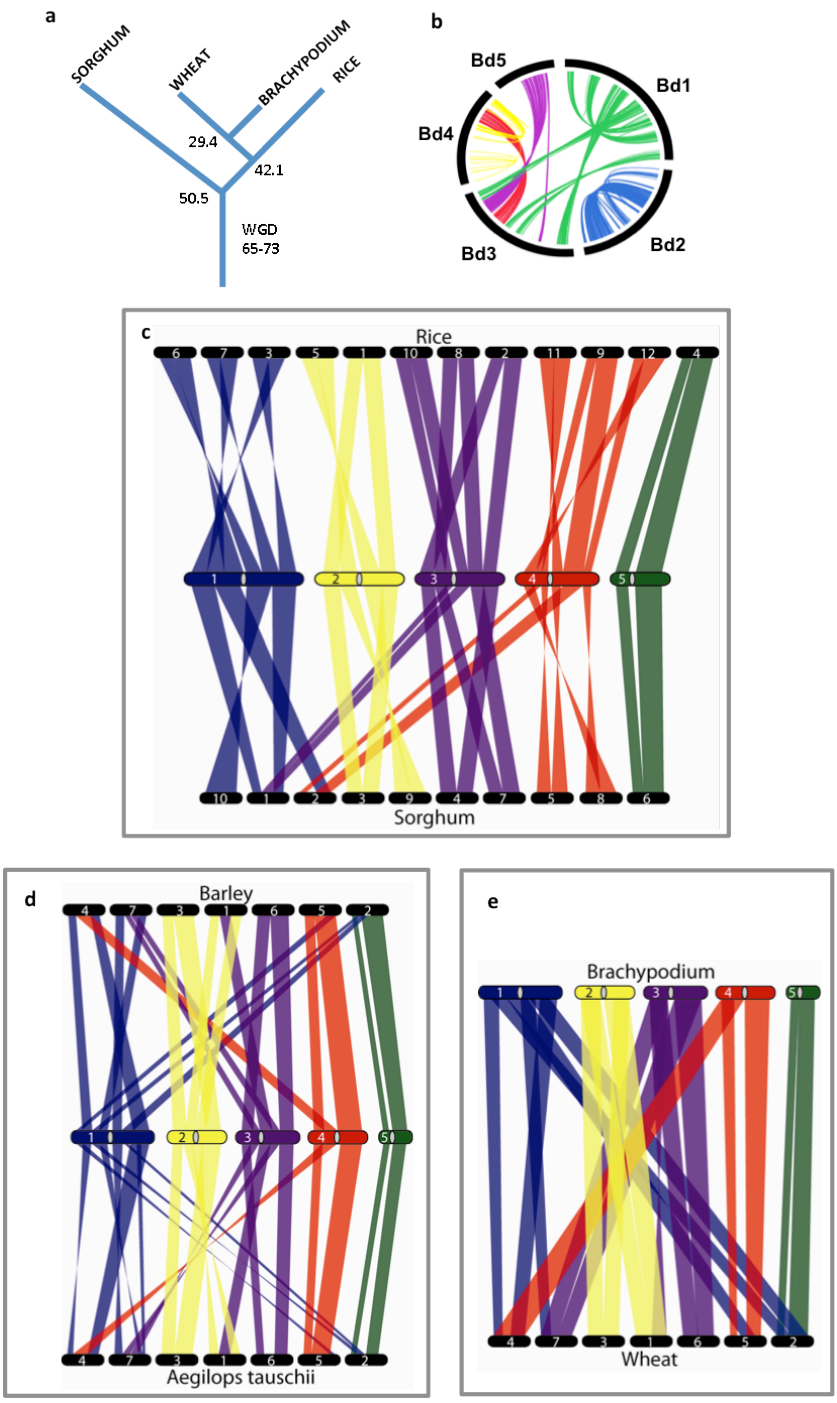
The abundance and distribution of the following genome elements are shown: complete LTR retroelements (cLTR); solo-LTRs (sLTR); potentially autonomous DNA transposons without MITEs (DNA-TEs); deletion derivatives of DNA transposons (MITES); gene exons (CDS); gene introns and satellite tandem arrays (STA). Graphs are from 0 to 100 percent bp coverage of the respective window. The heat map tracks have different scales: STA [0-55|scaled to max10] %bp; cLTRs [0-36|scaled to max 20] %bp; sLTRs [0-4] %bp; DNA-TEs [0-20] %bp; MITEs [0-22] %bp; CDS (exons) [0-22.3%] %bp. The triangles identify syntenic breakpoints.




**Appendix II: Figure 2. Transcript and gene identification and distribution among three grass subfamilies.**

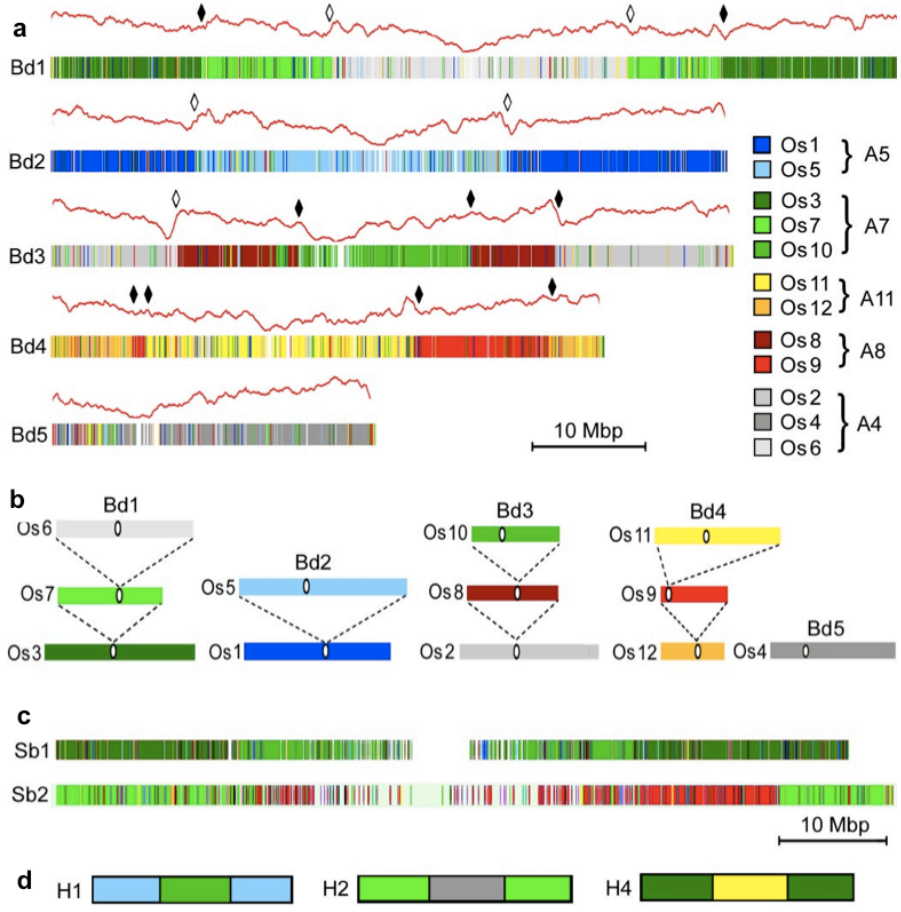
A. Genome-wide distribution of small RNA loci and transcripts in the Brachypodium genome. Brachypodium chromosomes (1-5) are shown at the top of the Figure. Total small RNA reads (black lines) and total small RNA loci (red lines) are shown on the top panel. Histograms plot 21 nt (blue) or 24 nt (red) small RNA reads normalized for repeated matches to the genome. The phased loci histograms plot the position and phase-score of 21 nt (blue) and 24 nt (red) phased small RNA loci. Repeat-normalized RNA-seq reads histograms plot the abundance of reads matching RNA transcripts (green), normalized for ambiguous matches to the genome. B. Transcript coverage over gene features. Perfect

match 32-mer Illumina reads were mapped to the Brachypodium v1.0 annotation features using HashMatch (http://mocklerlab-tools.cgrb.oregonstate.edu/). Plots of Illumina coverage were calculated as the percentage of bases along the length of the sequence feature supported by Illumina reads for the indicated gene model features. The bottom and top of the box represent the 25th and 75th quartiles, respectively. The white line is the median and the open red diamonds are the mean. C. Venn diagram showing the distribution of shared gene families between representatives of the the *Ehrhartoideae* (rice RAP2), *Panicoideae* (Sorghum v1.4) and *Pooideae* (Brachypodium v1.0, and *Triticum aestivum* and *Hordeum vulgare* TCs/EST sequences). Paralogous gene families were collapsed in these datasets.




**Appendix II: Figure 3. Brachypodium genome evolution and synteny between grass subfamilies.**
A. The distribution maxima of mean synonymous substitution rates (Ks) of Brachypodium, rice, sorghum and wheat orthologous gene pairs (Figure S13) were used to define the divergence times of these species and the age of inter-

chromosomal duplications in Brachypodium. WGD= Whole Genome Duplication. The numbers refer to the predicted divergence times measured as millions of years ago (MYA). B. Diagram showing the six major inter-chromosomal Brachypodium duplications, defined by 723 paralogous relationships, as coloured bands linking the five chromosomes. C. Identification of chromosome relationships between the Brachypodium, rice, and sorghum genomes. Orthologous relationships between the 25,532 protein-coding Brachypodium genes, 7,216 sorghum orthologs (12 syntenic blocks), and 8,533 rice orthologs (12 syntenic blocks) were defined. Sets of collinear orthologous relationships are represented by a coloured band according to each Brachypodium chromosome (blue-chr. 1; yellow-chr.2; violet-chr.3; red-chr.4; green-chr.5). The white region in each Brachypodium chromosome represents the centromeric region. D. Orthologous gene relationships between Brachypodium and barley and *Ae. tauschii* were identified using genetically-mapped ESTs. 2,516 orthologous relationships defined 12 syntenic blocks. These are shown as coloured bands. E. Orthologous gene relationships between Brachypodium and hexaploid bread wheat defined by 5,003 ESTs mapped to wheat deletion bins. Each set of orthologous relationships is represented by a band that is evenly spread across each deletion interval on the wheat chromosomes.



**Appendix II: Figure 4. A recurring pattern of nested chromosome fusions in grasses.**

A. The five Brachypodium chromosomes are coloured according to homology with rice chromosomes (Os1-Os12). Chromosomes descended from an ancestral chromosome (A4-A11) through whole genome duplication are displayed in shades of the same color. Gene density is indicated as a red line above the chromosome maps. Major discontinuities in gene density identify syntenic breakpoints, which are marked by a diamond. B. A pattern of nested insertions of whole chromosomes into centromeric regions explains the observed syntenic break points. Bd5 has not undergone chromosome fusion. C. Examples of nested chromosome insertions in sorghum (Sb) chromosomes 1 and 2. D. Examples of nested chromosome insertions in barley inferred from genetic maps. Nested insertions were not identified in other chromosomes, possibly due to the low resolution of genetic maps.

## BIBLIOGRAPHY

1. Somerville, C. The billion-ton biofuels vision. *Science* **312**, 1277 (2006).

2. Kellogg, E.A. Evolutionary history of the grasses. *Plant Physiol* **125**, 1198-205 (2001).

3. Gaut, B.S. Evolutionary dynamics of grass genomes. *New Phytologist* **154**, 15-28 (2002).

4. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793-800 (2005).

5. Paterson, A.H. et al. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-6 (2009).

6. Wei, F. et al. Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet* **3**, e123 (2007).

7. Moore, G., Devos, K.M., Wang, Z. & Gale, M.D. Cereal genome evolution. Grasses, line up and form a circle. *Curr Biol* **5**, 737-9 (1995).

8. Salamini, F., Ozkan, H., Brandolini, A., Schafer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nat Rev Genet* **3**, 429-41 (2002).

9. Draper, J. et al. Brachypodium distachyon. A new model system for functional genomics in grasses. *Plant Physiol* **127**, 1539-55 (2001).

10. Vain, P. et al. Agrobacterium-mediated transformation of the temperate grass Brachypodium distachyon (genotype Bd21) for T-DNA insertional mutagenesis. *Plant Biotechnol J* **6**, 236-45 (2008).

11. Vogel, J. & Hill, T. High-efficiency Agrobacterium-mediated transformation of Brachypodium distachyon inbred line Bd21-3. *Plant Cell Rep* **27**, 471-8 (2008).

12. Vogel, J.P., Garvin, D.F., Leong, O.M. & Hayden, D.M. Agrobacterium-mediated transformation and inbred line development in the model grass Brachypodium distachyon. *Plant Cell, Tissue and Organ Culture* **84**, 199-211 (2006).

13. Filiz, E. et al. Molecular, morphological and cytological analysis of diverse Brachypodium distachyon inbred lines. *Genome* **52,** 876-890 (2009).

14. Vogel, J.P. et al. Development of SSR markers and analysis of diversity in Turkish populations of Brachypodium distachyon. *BMC Plant Biol* **9**, 88 (2009).

15. Garvin, D.F. et al. An SSR-based genetic linkage map of the model grass Brachypodium distachyon. *Genome* **53**, 1-13 (2009).

16. Huo, N. et al. Construction and characterization of two BAC libraries from Brachypodium distachyon, a new model for grass genomics. *Genome* **49**, 1099-108 (2006).

17. Huo, N. et al. The nuclear genome of Brachypodium distachyon: analysis of BAC end sequences. *Funct Integr Genomics* **8**, 135-47 (2008).

18. Gu, Y.Q. et al. A BAC-based physical map of Brachypodium distachyon and its comparative analysis with rice and wheat. *BMC Genomics* **10**, 496 (2009).

19. http://www.brachypodium.pw.usda.gov; http://www.brachytag.org

20. http://www.brachybase.org; http://phytozome.net; http://modelcrop.org; http://mips.org; http://brachypodium.pw.usda.gov.

21. Garvin, D.F. et al. Development of Genetic and Genomic Research Resources for Brachypodium distachyon, a new model system for Grass Crop Research. *Crop Science* **48**, S69-S84 (2008).

22. Bennett, M.D. & Leitch, I.J. Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Ann Bot (Lond)* **95**, 45-90 (2005).

23. Vogel, J.P. et al. EST sequencing and phylogenetic analysis of the model grass Brachypodium distachyon. *Theor Appl Genet* **113**, 186-95 (2006).

24. Rajagopalan, R., Vaucheret, H., Trejo, J. & Bartel, D.P. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev* **20**, 3407-25 (2006).

25. Tanaka, T. et al. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res* **36**, D1028-33 (2008).

26. Fox, S., Filichkin, S. & Mockler, T. Applications of ultra high throughput sequencing. In: *Plant Systems Biology*. Humana Press **553**, 79-108 (2009)

27. Gray, J. et al. A recommendation for naming transcription factor proteins in the grasses. *Plant Physiol* **149**, 4-6 (2009).

28. Vogel, J. Unique aspects of the grass cell wall. *Curr Opin Plant Biol* **11**, 301-7 (2008).

29. Bennetzen, J.L. & Kellogg, E.A. Do Plants Have a One-Way Ticket to Genomic Obesity? *Plant Cell* **9**, 1509-1514 (1997).

30. Wicker, T. & Keller, B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res* **17**, 1072-81 (2007).

31. Wicker, T. et al. Analysis of intraspecies diversity in wheat and barley genomes identifies breakpoints of ancient haplotypes and provides insight into the structure of diploid and hexaploid triticeae gene pools. *Plant Physiol* **149**, 258-70 (2009).

32. Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. & Wessler, S.R. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**, 569-73 (2004).

33. Morgante, M. et al. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* **37**, 997-1002 (2005).

34. Grass Phylogeny Working Group. Phylogeny and subfamilial classification of the grasses (Poaceae). *Annals of the Missouri Botanical Garden* **88**, 373-457 (2001).

35. Bossolini, E., Wicker, T., Knobel, P.A. & Keller, B. Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. *Plant J* **49**, 704-17 (2007).

36. Charles, M. et al. Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of Pooideae and Ehrhartoideae, after their divergence from Panicoideae. *Mol Biol Evol* **26**, 1651-61 (2009).

37. Paterson, A.H., Bowers, J.E. & Chapman, B.A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A* **101**, 9903-8 (2004).

38. Stein, N. et al. A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. *Theor Appl Genet* **114**, 823-39 (2007).

39. Luo, M.C. et al. Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proc Natl Acad Sci U S A* **106**, 15780-5 (2009).

40. Qi, L.L. et al. A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168**, 701-12 (2004).

41. Salse, J. et al. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**, 11-24 (2008).

42. Srinivasachary, Dida, M.M., Gale, M.D. & Devos, K.M. Comparative analyses reveal high levels of conserved colinearity between the finger millet and rice genomes. *Theor Appl Genet* **115**, 489-99 (2007).

43. Vicient, C.M., Kalendar, R. & Schulman, A.H. Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon. *J Mol Evol* **61**, 275-91 (2005).

44. Meyers, B.C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R.W. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. *Plant Cell* **15**, 809-34 (2003).

45. Ma, J. & Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* **101**, 12404-10 (2004).

46. DOE. Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda (ed. U.S. Department of Energy) (2006).

47. FAO. World Agriculture: towards 2030/2050. Vol. Interim Report. Global Perspective Studies Unit, Food and Agriculture Organisation of the United Nations:Rome, Italy (2006).

48. McCarthy, E.M. & McDonald, J.F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362-7 (2003).

# Appendix III: The membrane-associated monooxygenase in the butane-oxidizing Gram-positive bacterium *Nocardioides* sp. strain CF8 is a novel member of the AMO/PMO family

Luis A. Sayavedra-Soto, Natsuko Hamamura, Chih-Wen Liu, Jeffrey A. Kimbrel, Jeff H. Chang and Daniel J. Arp

Environmental Microbiology Reports
2011, vol. 3, issue 3, pages 390-396
doi:10.1111/j.1758-2229.2010.00239.x

**SUMMARY**

The Gram-positive bacterium *Nocardioides* sp. strain CF8 uses a membrane-associated monooxygenase (pBMO) to grow on butane. The nucleotide sequences of the genes encoding this novel monooxygenase were revealed through analysis of a *de novo* assembled draft genome sequence determined by high-throughput sequencing of the whole-genome. The pBMO genes were in a similar arrangement to the genes for ammonia monooxygenase (AMO) from the ammonia-oxidizing bacteria and for particulate methane monooxygenase (pMMO) from the methane-oxidizing bacteria. The pBMO genes likely constitute an operon in the order *bmoC*, *bmoA* and *bmoB*. The nucleotide sequence was less than 50% similar to the genes for AMO and pMMO. The operon for pBMO was confirmed to be a single copy in the genome by Southern and computational analyses. In an incubation on butane the increase of transcriptional activity of the pBmoA gene was congruent with the increase of pBMO activity and suggested correspondence between gene expression and the utilization of butane. Phylogenetic comparison revealed distant but significant similarity of all three pBMO subunits to homologous members of the AMO/pMMO family and indicated that the pBMO represents a deeply branching third lineage of this group of particulate monooxygeases. No other *bmoABC*-like genes were found to cluster with pBMO lineage in phylogenetic analysis by database searches including genomic and metagenomic sequence databases. pBMO is the first example of the AMO/pMMO-like monooxygenase from Gram-positive bacteria showing similarities to proteobacterial pMMO and AMO sequences.

## INTRODUCTION

*Nocardioides* sp. strain CF8 is a Gram-positive bacterium (Hamamura and Arp, 2000) that can grow on C(2) to C(16) n-alkanes through the use of two monooxygenases, a binuclear-iron containing AlkB-type alkane monooxygenase and a particulate butane monooxygenase (pBMO). The AlkB-type alkane monooxygenase supported CF8 growth with C(6) to C(16) n-alkanes, while pBMO supported CF8 growth with C(2) to C(10) n-alkanes (Hamamura et al., 2001). The AlkB-type alkane monooxygenase belongs to the well-characterized rubredoxin-dependent alkane monooxygenase family. Members of this family are found in various organisms from oil-rich soils and typically AlkB catalyzes the degradation of medium- to long-chain n-alkanes (Hamamura et al., 2005; van Beilen and Funhoff, 2007; Wentzel et al., 2007; Schulz et al.). In contrast, the characterization of pBMO in *Nocardioides* sp. strain CF8 is not as advanced. While inhibition profiles and physiological characterizations suggested that pBMO is similar to the particulate monooxygenases (pMMO) of methane oxidizers and to the ammonia monooxygenases (AMO) of ammonia oxidizers (Hamamura et al., 1999; Hamamura and Arp, 2000), molecular evidence supporting these similarities was lacking.

The initial interest in diverse monooxygenases was that all have broad substrate specificity catalyzing the oxidation of alkanes, alkenes, aromatic hydrocarbons and ethers (Hyman and Wood, 1984; Juliette et al., 1993; Keener and Arp, 1994; Arp, 1995; Hamamura et al., 1999). The bacteria possessing monooxygenases capable of oxidizing these varied substrates have potential

applications in bioremediation (Field and Sierra-Alvarez, 2004). In fact, a search for microorganisms capable of degrading chloroform in a hydrocarbon-contaminated site led to the isolation of *Nocardioides* sp. strain CF8 (Hamamura et al., 1997; Hamamura and Arp, 2000). Furthermore the diverse monooxygenases also enable methane-, n-alkane- and ammonia-oxidizing bacteria to play important roles in the balances of the natural carbon and nitrogen biogeochemical cycles (Shively et al., 1998; Shively et al., 2001; van Beilen and Funhoff, 2005; Arp and Bottomley, 2006; van Beilen and Funhoff, 2007).

Most methane-oxidizing bacteria rely on particulate methane monooxygenase (pMMO) to grow on methane (Hakemian and Rosenzweig, 2007), and all ammonia-oxidizing bacteria rely on ammonia monooxygenase (AMO) to grow on ammonia (Arp et al., 2002). In these two distinctive groups of Gram-negative bacteria, pMMO and AMO are both membrane-associated enzymes and share many similarities in their encoding genes (*pmoCAB* and *amoCAB* respectively), likely structure ($a_3b_3g_3$ subunit composition kown for pMMO and possibly for AMO), metal content (Cu and possibly Fe) and substrate ranges (Arp et al., 2007; Hakemian and Rosenzweig, 2007). Recent ecological studies have shown the frequent occurrences of less similar genes for AMO in Archaea (Dang et al.; Molina et al.). The AMO operon in Archaea is in a different arrangement and has lower nucleotide sequence identity to the bacterial AMO operon (Walker et al.). Although there are many phylogenetic studies for *amoA* diversity in Archaea (Bernhard et al.; Urakawa et al.), the biochemical and

physiological studies of the archaeal AMO enzyme itself are limited to date (Martens-Habbena et al., 2009).

The evidence suggested that pBMO was similar to pMMO and AMO is: a) the butane oxidation activity in butane-grown *Nocardioides* sp. strain CF8 was inhibited by the Cu-selective chelator allylthiourea (Hamamura et al., 1999), a known inhibitor of the Cu containing AMO and pMMO; b) the butane oxidation activity of *Nocardioides* sp. strain CF8 was inactivated by light, which also inactivates AMO (Shears and Wood, 1985); c) the growth of *Nocardioides* sp. strain CF8 on short-chain, but not on long-chain n-alkanes, is stopped by the inhibitor of AMO activity 1-hexyne (Hamamura et al., 2001); d) in the membrane subcellular fraction of *Nocardioides* sp. strain CF8 a 30 kDa peptide was labeled with the inactivator $^{14}C_2H_2$ and the peptide showed a unique thermal aggregation phenomenon (Hamamura and Arp, 2000) in a similar way as the catalytic subunit of AMO in *N. europaea* (Hyman and Arp, 1992); and e) the $^{14}C_2H_2$ labeled peptide resided in the membrane fraction which suggested that pBMO is also a membrane-associated enzyme as AMO (Hyman and Arp, 1993).

In this work we show that pBMO in *Nocardioides* sp. strain CF8 represents a new branch in the phylogeny of monooxygenases that was previously only represented by pMMO in methane oxidizers and AMO in ammonia oxidizers (Bacteria and Archaea). The enzyme pBMO is the first example of a membrane-associated monooxygenase in a Gram-positive bacterium with similarity to pMMO and AMO.

## RESULTS AND DISCUSSION

### Nucleotide sequence of pBMO

Despite all indirect evidence for the similarity of pBMO to pMMO and AMO, and all prior physiological information (Hamamura et al., 1997; Hamamura et al., 1999; Hamamura and Arp, 2000; Hamamura et al., 2001), molecular evidence confirming the similarity of the genes encoding these three enzymes was not forthcoming. Primer sets derived from the nucleotide sequences of *amo* (NCBI accession AF058691) or *pmo* (NCBI accession L40804) failed to amplify DNA from *Nocardioides* sp. strain CF8. Furthermore, Southern hybridization blots with probes derived from the sequences of *amo* or *pmo* did not detect any DNA fragment in *Nocardioides* sp. strain CF8 either. Finally, attempts to isolate the $^{14}C_2H_2$ labeled polypeptide in *Nocardioides* sp. strain CF8 to obtain amino acid peptide sequences by mass spectroscopy were also unsuccessful.

High-throughput genome sequencing was the breakthrough needed to obtain the nucleotide sequence for the genes encoding pBMO. The genome of *Nocardioides* sp. strain CF8 was prepared for sequencing using paired-end DNA sample preparation and cluster generation kits as directed by the manufacturer (Illumina Inc. Ca, USA). An Illumina GAII sequencing instrument generated 1,876,239 pairs of reads (each 76 nucleotides in length) from fragments of about 300 base pairs (bp). The reads were assembled *de novo* with the software Velvet version 0.7.55 (Zerbino and Birney, 2008) and produced 121 contigs of size 100 bp and larger with an average length of 35,000 bases. The 121 contigs totaled 4,213,187 bp of sequence data and had an average of 32 x coverage. The draft

genome assembly included 4955 ambiguous bases (~1 ambiguous base every 850 bp).

The contigs were ordered according to length and open reading frames (ORFs) larger than 300 nucleotides were identified using Artemis software release 11.22 (Sanger Institute Cambridge, UK). The sequence of the genes encoding pBMO was obtained by scanning the draft genome sequence using Artemis Navigator and the feature "Find Amino Acid String" to find the amino acid sequence EQ**D**ASW**H**, a locus highly conserved in AmoC in ammonia oxidizers and PmoC in methane oxidizers (Fig. 1). The sequence of the amino acids in bold at this locus was reported to bind Zn in pMMO crystals and is highly conserved in methanotrophs and in ammonia oxidizers (Hakemian and Rosenzweig, 2007). The amino acids colored in Fig. 1 are the ligands to the metal cofactors deduced from the crystal structure of pMMO (Lieberman and Rosenzweig, 2005; Hakemian and Rosenzweig, 2007; Balasubramanian et al., 2010) and are conserved, with exception of B48 and B72, in pBMO. A single locus encoding this string of amino acids was found in the draft genome sequence of *Nocardioides* sp. strain CF8 and formed part of a locus with similarity to pMMO or AMO. The genes in this locus were in the same arrangement as the genes for pMMO (Semrau et al., 1995) or AMO (McTavish et al., 1993), namely in the order *bmoC*, *bmoA* and *bmoB* (Fig. 2). We refer to this genetic region as the pBMO-encoding locus.

**Analysis of the DNA sequence**

The 16S ribosomal RNA gene of the draft genome sequence had seven mismatches in comparison to its deposited partial sequence of 1460 bp from

*Nocardioides* sp. strain CF8 (Hamamura and Arp, 2000). The mismatches compelled us to examine the accuracy of the pBMO-encoding genes we discovered in the draft genome sequence. We designed primers based on the draft genome sequence and used genomic DNA from strain CF8 as a template for amplification by PCR (GoTaq green master mix; Promega). We amplified six DNA fragments that together, spanned the entire pBMO-encoding locus. Each product was sequenced using Sanger sequencing. The nucleotide sequence of the pBMO-encoding genes was 100 % identical to the locus in the draft genome sequence. The corroborated nucleotide sequence of the pBMO-encoding genes was deposited in the NCBI database with accession number HM623169.

At the pBMO locus the open reading frames (ORFs) for pBMO had classical Shine-Dalgarno sequences upstream of their annotated start codons. The intergenic region between *bmoC* and *bmoA* is 61 bp and the intergenic region between *bmoA* and *bmoB* is 4 bp; all pBMO- encoding ORFs form a single operon as demonstrated by the DNA products of the intergenic regions that were amplified by RT-PCR (not shown). No reliable predictions of promoter/functional motifs could be inferred in this DNA sequence.

NCBI/BLASTN analysis showed that the nucleotide sequence of *bmoC* had no more than 45 % identity with any other genes encoding for PmoC or AmoC. The nucleotide sequences of *bmoA* and *bmoB* had less than 43% identity to that in the genes for PmoA or AmoA, and to the genes for PmoB or AmoB respectively (Table 1). Given that the genes for pBMO are less than 50% similar to the genes of pMMO and AMO, it is not surprising that previous PCR and

Southern-based approaches failed to identify pBMO-encoding genes from *Nocardioides* sp. strain CF8 genomic DNA. Analysis of the draft genome sequence suggested *Nocardioides* sp. strain CF8 had only one copy of the pBMO-encoding locus, which was contained within a contig of 10,300 bp (three times as large as the pBMO operon). We used a probe for *bmoA* in Southern analysis of genomic DNA of CF8 digested with four restriction enzymes that do not cut the pBMO-encoding operon to confirm our observations. Results for all four restriction enzymes showed single fragments, confirming the presence of only one locus (Fig. 2).

**The expression of bmoA and pBMO activity are induced concomitantly upon exposure to butane.**

*Nocardioides* sp. strain CF8 grown with a non-alkane carbon source still have butane-oxidation activity. To demonstrate indirectly that the expression of *bmoA* is linked to pBMO enzymatic activity, butane grown cells were harvested and exposed to acetylene to inactivate pBMO. The cells were tested to assure that there was no pBMO activity, then washed and stored overnight in the form of a pellet to deplete the mRNA pool. To show the specific expression of the genes for pBMO, the cells were suspended in growth medium to the original cell density and incubated in butane to recover the lost pBMO activity (Fig. 3). The activity of pBMO was detected by the oxidation of ethylene to ethylene oxide using gas chromatography as described (Hamamura et al., 1999). The oxidation of ethylene is as a surrogate reaction catalyzed by pBMO, pMMO and AMO (Hamamura et al., 1999). Cells grown on butane, and those that recovered butane oxidation after acetylene inactivation, produced at least 14 nmol of ethylene oxide•$min^{-1}$•mg

protein$^{-1}$ (Fig. 3), in agreement with previous results (Hamamura et al., 1999). The cells upon incubation in butane showed increases in *bmoA* levels as pBMO activity increased (Fig. 3). Control reactions where no reverse transcriptase was added did not amplify products (not shown). The amplification of the cDNA for the 16S rRNA indicated that approximately equal amounts of RNA were tested by RT-PCR at 2 and 4-hour time points (Fig. 3).

**Phylogeny of pBMO**

To explore the potential phylogenetic linkage of pBMO to AMO and pMMO, phylogenetic trees of translated amino acid sequences of *bmoABC* genes were constructed with representative AMO and pMMO sequences (Fig 4). A multiple sequence alignment was performed with ClustalX (version 1.81) using default values (Thompson et al., 2002; Larkin et al., 2007) and edited manually. Phylogenetic trees were constructed on aligned sequences using evolutionary distance (Jukes-Cantor model, with neighbor-joining) and parsimony methods of PAUP*4.0 software (Sinauer Associates, Sunderland, MA). Phylogenetic comparison revealed distant but significant similarity of all three pBMO subunits to members of the AMO/pMMO family and clearly indicated that pBMO represents a deeply branching third lineage of the bacterial family (Fig 4). This pBMO sequence is the first example of the AMO/pMMO-like monooxygenase from Gram-positive bacteria and showed comparable levels of similarities to both proteobacterial pMMO and AMO sequences (e.g. the pBmoA sequence showed 37% and 39% average amino acid identity to pMMO and AMO sequences, respectively), while it is more distantly related to archaeal AMO sequences (20%

average amino acid identity). Similarities of pBMO phylogenetic tree topology to the 16S rRNA gene-based tree and the operon structures within bacterial family could expand the previously proposed hypothesis of divergent, orthologous *amo(pmo)*-homologous operon evolution (Klotz and Norton, 1998) to include *bmo* in Gram-positive bacteria. In their hypothesis, Klotz and Norton suggested that the evolution of *amo/pmo*-homologous monooxygenase genes in the last common ancestor started with speciation and diverged to acquire distinct yet rather overlapping substrate specificities.

Although numerous Gram-positive isolates have been characterized for their ability to grow on short-chain gaseous alkanes as reviewed elsewhere (Shennan, 2006), BLAST sequence search against published genome sequences (including ethylene-utilizing *Nocardioides* sp. JS614 (Mattes et al., 2005) and other hydrocarbon-degrading related Actinobacteria), and metagenomic sequences yielded no *bmoCAB*-like genes which clustered with pBMO lineage in phylogenetic analysis (data not shown). A highly diverged AMO/pMMO-related monooxygenase sequence, ERBWC_3B, from seafloor methane vent environments (Tavormina et al., 2008) also did not cluster with the pBMO lineage (Supporting Figure 1A). Recently, additional highly diverged monooxygenase-encoding gene fragments with sequence similarity to AMO and pMMO genes have been reported in a *pmoA* survey among pelagic marine environments (Tavormina et al. 2010). Phylogenetic analysis of these short sequences (~50 amino acids: Supporting Figure 1B) showed that, although pBMO is still deeply-branched and formed a separate lineage, it was more closely related (~30%

amino acid similarity) to Group O and Group W sequences (obtained from oligotrophic gyre and California seep/continental margin, respectively) than to AMO or pMMO sequences (22 % and 16% average amino acid similarity, respectively). This result could imply the potential affiliation of pBMO with these environmental clone sequences. However, further phylogenetic analysis with longer sequences and functional identification of these environmental sequences are necessary to confirm this speculation.

Our results revealed the novelty of pBMO in *Nocardioides* sp. strain CF8, and the potential for further study of this poorly understood lineage to gain insight into the evolution and diversity of the monooxygenases AMO/pMMO/pBMO family.

## ACKNOWLEDGEMENTS

We thank Mark Dasenko and Chris Sullivan of the Center for Genome Research and Biocomputing (CGRB) for sequencing and computational support. This research was supported in part by the Oregon Agricultural Experiment Station (DJA and JHC) and by the National Research Initiative Competitive Grant no. 2008-35600-18783 from the USDA's National Institute of Food and Agriculture, Microbial Functional Genomics Program to JHC. CWL funding was by the Study Abroad Program, National Science Council of Taiwan, Grant no. 98-2917-I-002-119.

## REFERENCES

Arp, D.J. (1995) Understanding the diversity of trichloroethene co-oxidations. *Current Biol.* **6**: 352-358.

Arp, D.J., and Bottomley, P.J. (2006) Nitrifiers: More than 100 years from isolation to genome sequences. *Microbe* **1**: 229-234.

Arp, D.J., Sayavedra-Soto, L.A., and Hommes, N.G. (2002) Molecular biology and biochemistry of ammonia oxidation by *Nitrosomonas europaea*. *Arch Microbiol* **178**: 250-255.

Arp, D.J., Chain, P.S., and Klotz, M.G. (2007) The impact of genome analyses on our understanding of ammonia-oxidizing bacteria. *Annu Rev Microbiol* **61**: 503-528.

Balasubramanian, R., Smith, S.M., Rawat, S., Yatsunyk, L.A., Stemmler, T.L., and Rosenzweig, A.C. (2010) Oxidation of methane by a biological dicopper centre. *Nature* **465**: 115-119.

Bernhard, A.E., Landry, Z.C., Blevins, A., de la Torre, J.R., Giblin, A.E., and Stahl, D.A. (2010) Abundance of ammonia-oxidizing Archaea and Bacteria along an estuarine salinity gradient in relation to potential nitrification rates. *Appl Environ Microbiol* **76**: 1285-1289.

Dang, H., Luan, X.W., Chen, R., Zhang, X., Guo, L., and Klotz, M.G. (2010) Diversity, abundance and distribution of *amoA*-encoding archaea in deep-sea methane seep sediments of the Okhotsk Sea. *FEMS Microbiol Ecol*.

Field, J.A., and Sierra-Alvarez, R. (2004) Biodegradability of chlorinated solvents and related chlorinated aliphatic compounds. *Reviews Environ Science & Bio/Technol.* **3**: 185:254.

Hakemian, A.S., and Rosenzweig, A.C. (2007) The biochemistry of methane oxidation. *Annu Rev Biochem* **76**: 223-241.

Hamamura, N., and Arp, D.J. (2000) Isolation and characterization of alkane-utilizing *Nocardioides* sp. strain CF8. *FEMS Microbiol Lett* **186**: 21-26.

Hamamura, N., Yeager, C., and Arp, D.J. (2001) Two distinct monooxygenases for alkane oxidation in *Nocarioides* sp. strain CF8. *Appl Environ Microbiol* **67**: 4992-4998.

Hamamura, N., Storfa, R.T., Semprini, L., and Arp, D.J. (1999) Diversity in butane monooxygenases among butane-grown bacteria. *Appl Environ Microbiol* **65**: 4586-4593.

Hamamura, N., Olson, S.H., Ward, D.M., and Inskeep, W.P. (2005) Diversity and functional analysis of bacterial communities associated with natural hydrocarbon seeps in acidic soils at Rainbow Springs, Yellowstone National Park. *Appl Environ Microbiol* **71**: 5943-5950.

Hamamura, N., Page, C., Long, T., Semprini, L., and Arp, D.J. (1997) Chloroform cometabolism by butane-grown CF8, *Pseudomonas butanovora*, and *Mycobacterium vaccae* JOB5 and methane-grown *Methylosinus trichosporium* OB3b. *Appl Environ Microbiol* **63**: 3607-3613.

Hyman, M.R., and Wood, P.M. (1984) Bromocarbon oxidations by *Nitrosomonas europaea*. In *Microbial growth on $C_1$compounds: Proceedings of the 4th International Symposium.* Crawford, R.L., and Hanson, R.S. (eds). Washington, DC: American Society for MIcrobiology, pp. 49-52.

Hyman, M.R., and Arp, D.J. (1992) $^{14}C_2H_2$- and $^{14}CO_2$-labeling studies of the *de novo* synthesis of polypeptides by *Nitrosomonas europaea* during recovery from acetylene and light inactivation of ammonia monooxygenase. *J Biol Chem* **267**: 1534-1545.

Hyman, M.R., and Arp, D.J. (1993) An electrophoretic study of the thermal-dependent and reductant-dependent aggregation of the 27 kDa component of ammonia monooxygenase from *Nitrosomonas europaea*. *Electrophoresis* **14**: 619-627.

Juliette, L.Y., Hyman, M.R., and Arp, D.J. (1993) Inhibition of Ammonia Oxidation in Nitrosomonas europaea by Sulfur Compounds: Thioethers Are Oxidized to Sulfoxides by Ammonia Monooxygenase. *Appl Environ Microbiol* **59**: 3718-3727.

Keener, W.K., and Arp, D.J. (1994) Transformations of aromatic compounds by *Nitrosomonas europaea*. *Appl Environ Microbiol* **60**: 1914-1920.

Klotz, M.G., and Norton, J.M. (1998) Multiple copies of ammonia monooxygenase (*amo*) operons have evolved under biased AT/GC mutational pressure in ammonia-oxidizing autotrophic bacteria. *FEMS Microbiol Lett* **168**: 303-311.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H. et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.

Lieberman, R.L., and Rosenzweig, A.C. (2005) Crystal structure of a membrane-bound metalloenzyme that catalyses the biological oxidation of methane. *Nature* **434**: 177-182.

Martens-Habbena, W., Berube, P.M., Urakawa, H., de la Torre, J.R., and Stahl, D.A. (2009) Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* **461**: 976-979.

Mattes, T.E., Coleman, N.V., Spain, J.C., and Gossett, J.M. (2005) Physiological and molecular genetic analyses of vinyl chloride and ethene biodegradation in *Nocardioides* sp. strain JS614. *Arch Microbiol* **183**: 95-106.

McTavish, H., Fuchs, J.A., and Hooper, A.B. (1993) Sequence of the gene coding for ammonia monooxygenase in *Nitrosomonas europaea*. *J Bacteriol* **175**: 2436-2444.

Molina, V., Belmar, L., and Ulloa, O. (2010) High diversity of ammonia-oxidizing archaea in permanent and seasonal oxygen-deficient waters of the eastern South Pacific. *Environ Microbiol* **12**:Jun 28.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989) *Molecular cloning: a laboratory manual*. Cold Springs Harbor, N.Y.: Cold Springs Harbor Laboratory Press.

Schulz, S., Pérez-de-Mora, A., Engel, M., Munch, J.C., and Schloter, M. (2010) A comparative study of most probable number (MPN)-PCR vs. real-time-PCR for the measurement of abundance and assessment of diversity of *alkB* homologous genes in soil. *J Microbiol Methods* **80**: 295-298.

Semrau, J.D., Chistoserdov, A., Lebron, J., Costello, A., Davagnino, J., Kenna, E. et al. (1995) Particulate methane monooxygenase genes in methanotrophs. *J Bacteriol* **177**: 3071-3079.

Shears, J.H., and Wood, P.M. (1985) Spectroscopic evidence for a photosensitive oxygenated state of ammonia mono-oxygenase. *Biochem. J.* **226**: 499-507.

Shennan, J.L. (2006) Utilisation of $C_2$-$C_4$ gaseous hydrocarbons and isoprene by microorganisms. *J Chel Technol Biotechnol* **81**: 237-256.

Shively, J.M., van Keulen, G., and Meijer, W.G. (1998) Something from almost nothing: carbon dioxide fixation in chemoautotrophs. *Annu Rev Microbiol* **52**: 191-230.

Shively, J.M., English, R.S., Baker, S.H., and Cannon, G.C. (2001) Carbon cycling: the prokaryotic contribution. *Curr Opin Microbiol* **4**: 301-306.

Tavormina, P.L., Ussler, W., 3rd, and Orphan, V.J. (2008) Planktonic and sediment-associated aerobic methanotrophs in two seep systems along the North American margin. *Appl Environ Microbiol* **74**: 3985-3995.

Tavormina, P.L., Ussler, W., 3rd, Joye, S.B., Harrison, B.K., and Orphan, V.J. (2010) Distributions of putative aerobic methanotrophs in diverse pelagic marine environments. *ISME J* **4**: 700-710.

Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* **Chapter 2**: Unit 2 3.

Urakawa, H., Martens-Habbena, W., and Stahl, D.A. (2010) High Abundance of Ammonia-Oxidizing Archaea in Coastal Waters Determined Using a Modified DNA Extraction Method. *Appl Environ Microbiol*. **76**:2129:2135

van Beilen, J.B., and Funhoff, E.G. (2005) Expanding the alkane oxygenase toolbox: new enzymes and applications. *Curr Opin Biotechnol* **16**: 308-314.

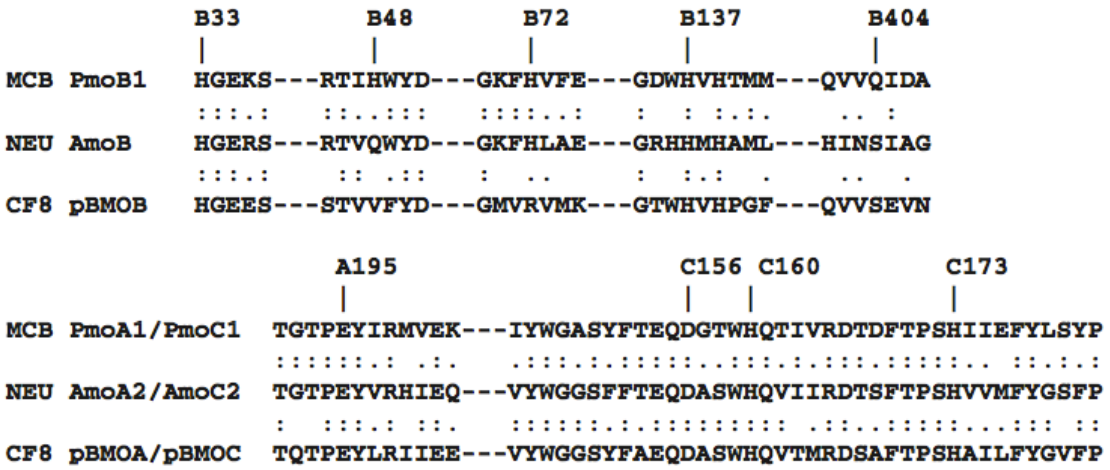
van Beilen, J.B., and Funhoff, E.G. (2007) Alkane hydroxylases involved in microbial alkane degradation. *Appl Microbiol Biotechnol* **74**: 13-21.

Walker, C.B., la Torre, J.R., Klotz, M.G., Urakawa, H., Pinel, N., Arp, D.J. et al. (2010) *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci U S A*.

Wentzel, A., Ellingsen, T.E., Kotlar, H.K., Zotchev, S.B., and Throne-Holst, M. (2007) Bacterial metabolism of long-chain n-alkanes. *Appl Microbiol Biotechnol* **76**: 1209-1221.

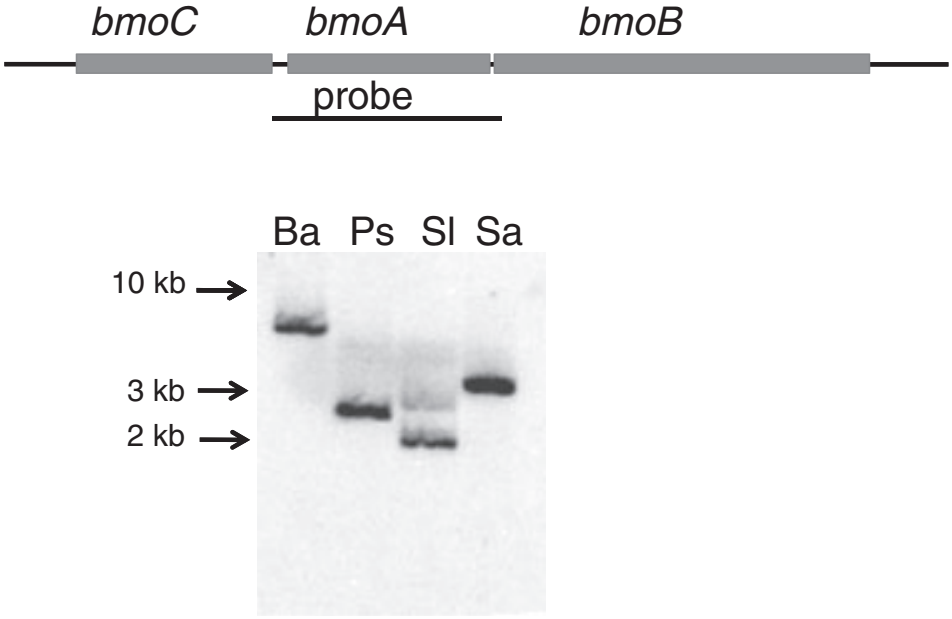
Wilson, K. (1995) Preparation of genomic DNA from bacteria. In *Current protocols in molecular biology*. Ausubel, F.M., and et al (eds). New York: John Wiley & Sons, pp. 2.4.1-2.4.5.

Zerbino, D.R., and Birney, E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res* **18**: 821-829.



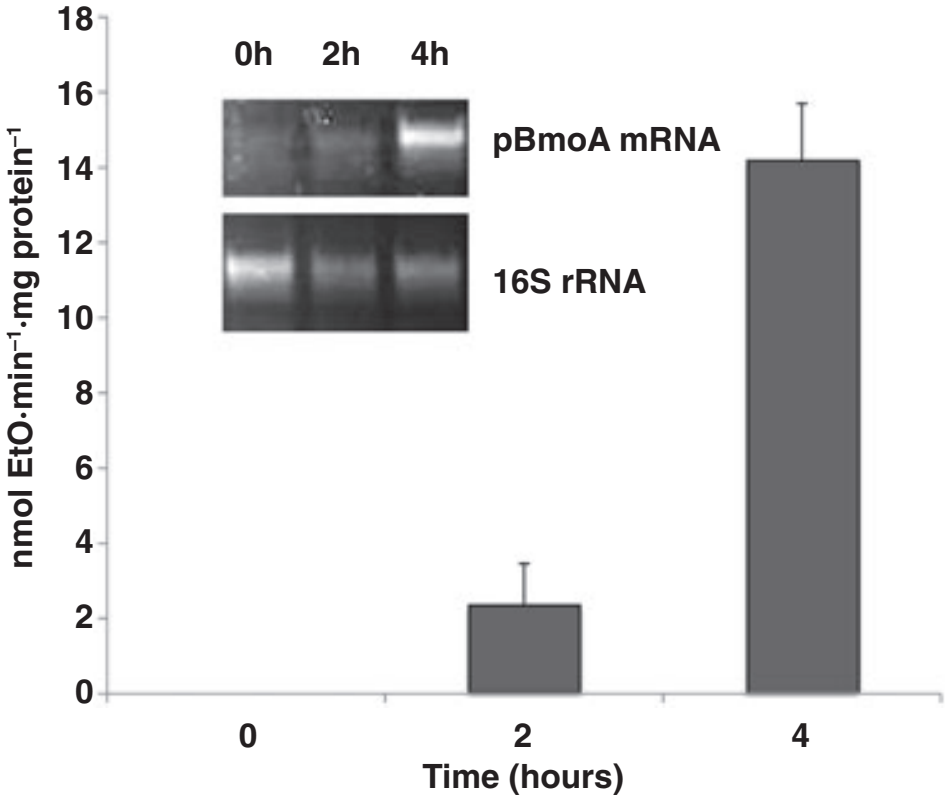

**Appendix III, Figure 1.**

Multiple sequence alignments indicating the ligands to the metal centers in the *Methylococcus capsulatus* (Bath) pMMO (top). Ligands to the putative mononuclear copper are at B48 and B72**.** Ligands to the dinuclear copper are at B33 and B137**.** Ligands to zinc are at A195, C156, C160 and C173**.** Amino acids in the ligands B48 and B72 for the putative mononuclear copper in pBMO are dissimilar. MCB: *Methylococcus capsulatus* sp. Bath; NEU: *Nitrosomonas europaea*; CF8: *Nocardioides* sp. strain CF8. Clustal W (Larkin et al., 2007) was used to align the sequences. The figure was adapted from published data (Hakemian and Rosenzweig, 2007).



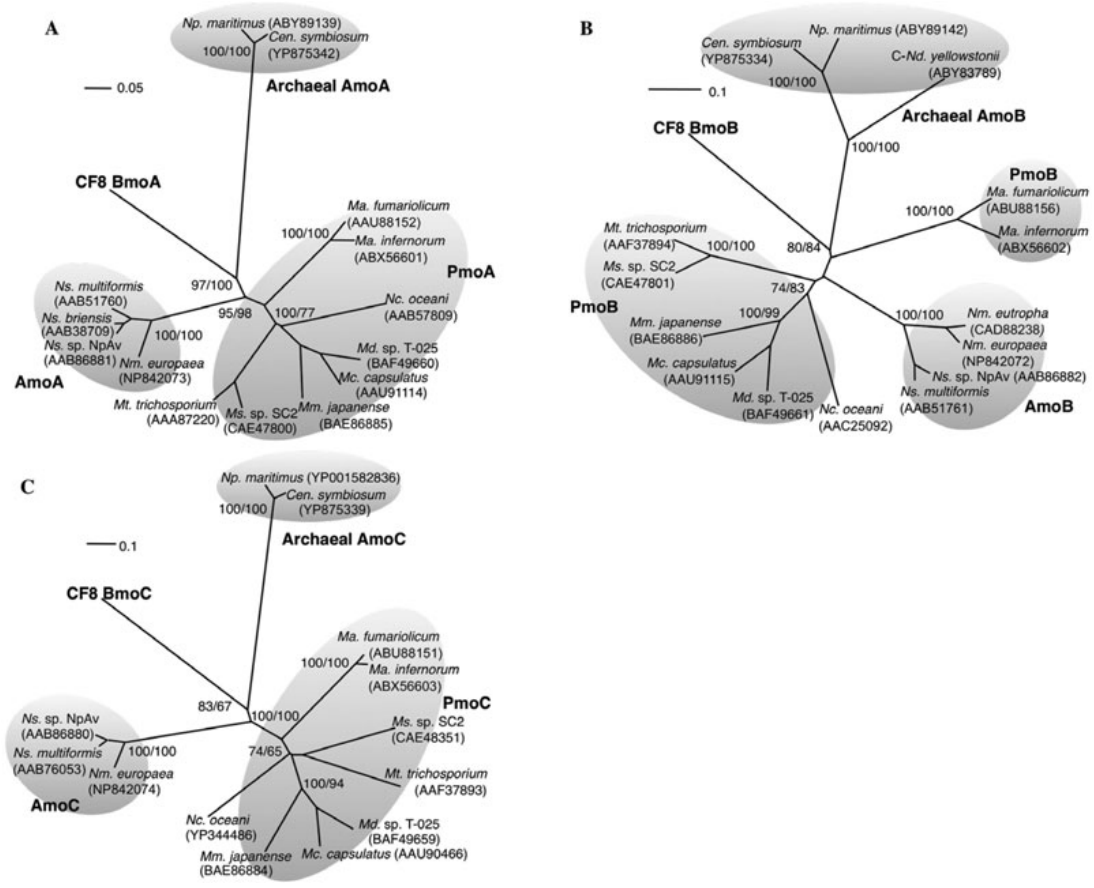

**Appendix III, Figure 2**

Physical map and phosphorimage of genomic DNA hybridized to a probe for *bmoA*. DNA was digested with *Bam*HI (Ba), *Pst*I (Ps), *Sal*I (Sl) and *Sac*I (Sa) and resolved in an agarose gel. The probe was labeled by random priming with [g$^{32}$P]-dCTP and analyzed by Southern hybridization as described (Sambrook et al., 1989). The cells were grown on butane in sealed 150 ml serum vials to which the gas was added as overpressure and extracted total nucleic acids as described (Hamamura et al., 2001). Cells were harvested at late logarithmic phase and subjected to nucleic acids extraction. To facilitate cell lysis in this Gram-positive bacterium the cells were treated with lysozyme (~ 1 mg/ ml) 10 min at 36°C. Genomic DNA was prepared by the CTAB (hexadecymethylammonium bromide) method as described (Wilson, 1995).




**Appendix III, Figure 3**
In *Nocardioides* sp. strain CF8 *bmoA* mRNA increased as the enzymatic activity of pBMO increased. In the insert are the RT-PCR products for *bmoA* and for the 16S rRNA estimating the total RNA levels in the reactions. The increase of pBMO activity was measured by the oxidation of ethylene to ethylene oxide in three biological replicates. The cells were grown in 600 ml medium with 10% butane (in the headspace) in 1 liter sealed serum bottles, harvested at early stationary phase (~0.5 $OD_{600}$) by centrifugation, and suspended in a slurry in phosphate buffer (6 ml). The slurry was placed in a 40 ml sealed serum bottle with 10 % acetylene (in the headspace) for 45 min at 30°C with shaking (100 rpm). The cells were washed and tested for total pBMO inactivation. The cells were then harvested by centrifugation and allowed to stand in a pellet at 4°C overnight to degrade the mRNA pool. The cell pellet was suspended to the original cell culture volume (~0.5 $OD_{600}$) and incubated with butane (10% vol/vol headspace). Samples were harvested by centrifugation at 0, 2 and 4 hours, suspended to 3.5 ml in growth medium without a carbon source and 0.5 ml used to test for activity. The remaining 3 ml sample was treated with lysozyme 10 min at 37 °C, at which point 1.5 ml acid phenol, 1.5 ml chloroform, 600 µl 3 M sodium acetate and 300 µl of 20 % sodium dodecyl sulfate were added and mixed thoroughly. The mix was subjected to centrifugation at 30,000 x g for 10 min and the supernatant recovered. Total RNA was purified from the supernatant in a CsCl step-gradient centrifugation. To 3 ml of supernatant 1.5 g CsCl were added and dissolved to

completion. The solution was then layered on 1.5 ml of 5.7 M CsCl in a 5 ml ultracentrifuge tube. The samples were centrifuged (150,000 x g) for 16 hours. The recovered RNA pellet was suspended in 50 µl RNAse-free water and treated with 10 units of RQ1 RNAse-free DNAse at 37 °C following the directions of the manufacturer (Promega). The RNA was further purified with a kit (RNeasy Qiagen). After all the purification steps a PCR reaction with primers for *bmoA* produced no product. The RT-PCR was carried out with an AccessQuick RT-PCR system (Promega) according to the manufacturer's directions. The primers pBMOCA-2a (5'-ccttctggtgttccaggttg-3') and pBMOCA-2b (5'-gcgaagaaggacaccaccat-3') amplified the DNA fragment of *bmoA* and the primers CF8 CF8 16S a (5'-taatggcctaccatggcttc-3') and 16S b (5'-gtattcaccgcagggttgtt-3') amplified the DNA fragment of the 16S rRNA (the inset shows the cDNA products obtained from the same RNA samples at the indicated time point).




**Appendix III, Figure 4**

Unrooted neighbor-joining (NJ) phylogenetic trees of the translated amino acid sequences encoded by (A) *bmoA*, (B) *bmoB*, and (C) *bmoC* sequences from *Nocardioides* sp. strain CF8 with known AMO and pMMO sequences. The scale bars represent changes per sequence position. Values for NJ and parsimony bootstrap (per 100 trials) are shown for major branch-points (neighbor-joining with Jukes and Cantor correction/parsimony using PAUP* 4.0). Abbreviations: Np., *Nitrosopumilus;* Cen., *Cenarchaeum;* C-Nd., Candidatus *Nitrosocaldus;* Ns., *Nitrosospira;* Nm., *Nitrosomonas;* Nc., *Nitrosococcus;* Ma., *Methylacidiphilum;*Mc., *Methylococcus;* Mm., *Methylomicrobium;* Ms., *Methylocystis;* Mt., *Methylosinus;* Md., *Methylocaldum.*

**Appendix III, Table 1. Features of the three subunits of pBMO predicted from DNA sequence analysis**

| Subunit | Molecular masses | Residues | IEP* | Amino acid similarity to pMMO or AMO (%) |
|---|---|---|---|---|
| pBmoC | 27862 | 247 | 6.68 | ≤45 |
| pBmoA | 28968 | 255 | 6.26 | ≤ 43 |
| pBmoB | 45325 | 418 | 4.55 | ≤ 31 |

*Isoelectric point.

# Appendix IV: RNA-Seq for Plant Pathogenic Bacteria

Jeffrey A. Kimbrel, Yanming Di, Jason S. Cumbie and Jeff H. Chang

Genes
2011, vol. 2 (4), pp. 689-705
doi:10.3390/genes2040689

## ABSTRACT

The throughput and single-base resolution of RNA-Sequencing (RNA-Seq) have contributed to a dramatic change in transcriptomic-based inquiries and resulted in many new insights into the complexities of bacterial transcriptomes. RNA-Seq could contribute to similar advances in our understanding of plant pathogenic bacteria but it is still a technology under development with limitations and unknowns that need to be considered. Here, we review some new developments for RNA-Seq and highlight recent findings for host-associated bacteria. We also discuss the technical and statistical challenges in the practical application of RNA-Seq for studying bacterial transcriptomes and describe some of the currently available solutions.


## INTRODUCTION: A SNEAK PEEK INTO RNA-SEQ

Genome sequences for host-associated bacteria are being generated at an extraordinary rate. Their availability has had important contributions towards deciphering the highly complex and fascinating biological interactions between symbionts and their hosts. Since the 2000s, when the first genome sequences of plant pathogens were determined, we have gained a greater appreciation into the mechanisms of virulence, such as secretion systems and repertoires of effectors, metabolic and biosynthetic capacities to adapt to different environments, biosynthesis of secondary metabolites and toxins to modulate host plants, and evolution as well as taxonomical relationships of plant pathogenic bacteria [1–9].

Genome sequences are by no means the end of the road. A genome sequence is a map with the challenge of exploration to improve and make sense

of it. As of six years ago, even *Escherichia coli*, the most heavily studied bacterium, had only 54% of its genes experimentally supported with another 32% computationally predicted [10]. No plant pathogenic bacterium is close to this level, as isolates belonging to the *Pseudomonas*, *Xanthomonas*, *Ralstonia*, and *Agrobacterium* genera have between 27%~37% of their genes annotated as "hypothetical". Adding to the challenges of studying plant pathogens is the amount of redundancy coded in their genomes and the subsequent difficulties that experimental biologists face in their efforts to map and characterize genes necessary for virulence [1].

Transcriptomic-based approaches have the potential to help rapidly address this knowledge gap. A transcriptome represents all RNA molecules, including the coding mRNAs as well as the noncoding rRNA, tRNA, sRNAs, *etc.* Investigators have mostly focused on protein coding mRNAs and, more recently, on the regulatory small RNAs, while excluding the "housekeeping" functional RNAs, such as rRNA, and tRNAs. As such, from hereafter, we use "transcriptome" to imply only mRNAs and sRNAs. The transcriptome is dynamic and is constantly changing in response to endogenous and exogenous cues. Thus, transcriptomic-based approaches typically rely on the characterization of snapshots captured from cells subjected to conditions and times of interest.

Microarrays were one of the earliest tools that offered researchers the once unique opportunity to investigate the reprogramming of a phytopathogenic bacterium's entire transcriptome. Microarrays were used to identify virulence regulons and study the physiological changes that occur in response to plant

signaling molecules or in conditions that mimic the host environment [4,11–16]. Microarrays have some constraints that cap the possible explorations into transcriptomes. Microarrays are designed according to an available genome sequence and may have limited use to only its corresponding isolate, or at best to a small number of genetically similar isolates. Additionally, microarrays are limited by the quality of the genome sequence and annotation. As a consequence, except for the genome tiling arrays, most microarrays cannot be used for gene discovery and refinement of genome annotations for improving future transcriptomic-based inquiries without subsequent redesigns.

Next generation (next gen) sequencing has pushed data generation into the logarithmic growth phase. Several next gen platforms are available that use different chemistries but offer the same advantages over traditional Sanger sequencing–dramatic increases in throughput with decreases in cost, time, and labor (reviewed in [17]). The application of next gen sequencing to transcriptomics has been coined the inaccurate term of RNA-Sequencing or RNA-Seq, which is, in practice, simply the highly parallelized sequencing of cDNA fragments. Direct sequencing of mRNA has also been demonstrated, but this approach has not yet been widely adopted [18]. As will be discussed, there are different preparation methods for RNA-Seq to yield different levels of information regarding the transcriptome.

RNA-Seq has been used for expression profiling as well as many other explorations into transcriptomes. Analysis of RNA-Seq has shown that, despite the perceived relative simplicity of bacterial genomes in comparison to their

eukaryotic hosts, bacterial transcriptomes and their regulation are nonetheless similar in complexity. Genes that escaped annotation have been uncovered using RNA-Seq, the most prominent being those of noncoding or small RNAs [19–26]. Subsequent characterization of sRNAs will contribute to a more comprehensive understanding in transcriptome regulation, as sRNAs largely function in gene regulation (reviewed in [27]). Analysis of RNA-Seq data derived from cDNA fragments prepared using enzymatic modifications to distinguish sense versus anti-sense strands or preprocessed versus processed transcripts, have helped to resolve overlapping or embedded genes as well as disputed operons, and identify transcript isoforms originating from alternative start sites [21,24–26,28,29]. In general, transcriptional initiation within upstream coding regions, anti-sense expression, and presence of alternative transcriptional start sites appear to occur with much higher prevalence than originally thought for bacterial genomes.

A distinct advantage of RNA-Seq is that cDNA fragments are directly sequenced and the reads can be *de novo* assembled to study organisms with no available reference genome sequence [30,31]. For bacteria, a more cost-effective and practical alternative is to combine analysis of RNA-Seq data with a draft genome sequence derived from next gen sequencing. This approach was successfully used to provide sufficient insights into the metabolic demands of a leech symbiont for the development of media to enable its culturing [32]. Furthermore, because of the single-base resolution and the ability to computationally predetermine and filter out ambiguous reads, RNA-Seq can also be used to study co-inhabitant or co-cultured microbes without concern for issues

such as the unknowable cross-hybridization associated with microarrays [32,33]. Thus, RNA-Seq could be used to study the potential synergistic or antagonistic interactions that occur in plant-pathogenic bacterial communities such as the case with the soft rot *Pectobacterium carotovora* [34].

On the surface, with these advantages, it almost seems absurd to not use RNA-Seq. Millions to billions of RNA-Seq reads, terabytes of data, will be available quickly and cheaply. However, to date, there has been only a single report describing the use of RNA-Seq to study the transcriptome of a plant pathogen, *Pseudomonas syringae* [25]. For many researchers, the outlook becomes bleak when faced with the task of handling and making sense of the massive amounts of data. Unlike analysis of microarrays, there are no out-of-the-box or one-size-fits-all packages for analysis of RNA-Seq for bacteria. Also, with RNA-Seq data, there may be concerns with computational hardware. Depending on the organism and scope of RNA-Seq experiment, a desktop computer is most likely insufficient.

RNA-Seq, its uses and its analytical tools, are still in their developmental stages. In the following, we briefly review options for preparing RNA from bacteria as well as some of the computational challenges associated with RNA-Seq. Many of these topics have been comprehensively reviewed [17,35–38]. We then turn our attention to the statistical challenges of analyzing RNA-Seq data, with emphasis on analysis of differential gene expression.

## TECHNIQUES FOR RNA-SEQ PREPARATIONS

One of the first tasks of RNA-Seq is to produce a transcriptome depleted of rRNAs and tRNAs. These functional RNAs typically exceed 90% of the total RNA preparation and will likely represent >99% of the RNA-Seq reads if not sufficiently addressed [33]. In eukaryotes, mRNAs are processed in part by addition of a 5' m7GpppX cap and 3' poly(A) tail, which can be exploited to enrich for mRNAs. In prokaryotes, these features are not present. Rather, newly synthesized or preprocessed RNAs have a triphosphate at the 5' end and the processed RNAs, such as rRNA and tRNAs, bear a 5' monophosphate. As a consequence, many of the available methods for transcriptomes of bacteria deplete the unwanted RNAs from preparations.

For many experiments, the tRNAs and 5 s rRNA are of little concern because they can be excluded simply based on their small sizes. However, a fraction of the sRNAs may also be lost with these approaches as some sRNAs are as small as 50 nucleotides in length [39]. Thus, if one uses a preparation method to specifically capture smaller sized RNAs, an approach to deplete tRNAs and 5 s RNAs should be considered, otherwise only a small percentage of the reads will be informative [19].

In most cases, the concern is with the 16 s and 23 s rRNAs. Three methods are commercially available that address these abundant rRNAs. Subtractive hybridization is the most popular, e.g., MicrobExpress (Ambion, Austin, TX) and Ribominus ([40]; Invitrogen, Carlsbad, CA). Subtractive hybridization is straightforward and relies on bead-associated oligonucleotides

complementary to 16 s and 23s sequences to deplete undesired rRNAs. One feature that distinguishes Ribominus from MicrobExpress is its use of locked-nucleic acids (LNAs) in the rRNA capture oligonucleotides [40]. LNAs are nucleotide analogs capable of complementary basepairing but with much higher thermal affinities allowing for the use of a higher temperature during depletion steps to increase the specificity of rRNA capture [41]. We have found that one round of MicrobExpress followed by a round of Ribominus is effective for removing a large fraction of the rRNA from RNA preparations of *P. syringae* (Figure 1A). Using qRT-PCR to assess efficiency of depletion, on average, less than 0.01% and 10% of the 16 s and 23 s rRNA, respectively, remained relative to the starting preparation (Kimbrel and Chang, unpublished). After sequencing, on average, approximately 20% of the reads aligned to the rRNA-encoding locus with 17% and 83% of those corresponding to the 16 s and 23 s rRNA, respectively. In our best case, only 12% of the total RNA-Seq reads corresponded to rRNA.

Since subtractive hybridization is a method of depletion, one must resist the temptation to use more input RNA than recommended, otherwise the transcriptome preparation may not be sufficiently devoid of rRNAs. Additionally, one needs to consult the list of compatible bacteria to determine whether the commercially available capture oligonucleotides will work for one's bacterium of interest. If inadequate, species-specific capture oligonucleotides can be designed but researchers should be aware that, due to post-transcriptional processing of precursor rRNA, the molecule is often fragmented and can exist as multiple,

separate fragments [42]. Oligonucleotides should therefore be designed to several locations along the 16 s and 23 s-encoding rRNA to sufficiently capture each of the processed forms. Processing may contribute in part to the peaks and valleys pattern of RNA-Seq read alignment to the rRNA-encoding locus (Figure 1B).

The processed rRNAs can also be preferentially degraded using a 5'-Phosphate-Dependent Exonuclease (Terminator; Epicentre, Madison, WI). This approach has important implications in downstream data analyses and can be used to characterize bacterial transcriptomes with greater precision (see below). A third and relatively new method uses enrichment by relying on “not so random” oligonucleotides during cDNA preparation to bias towards non-rRNA transcripts [43] (Ovation® Prokaryotic RNA-Seq System; NuGen, San Carlos, CA [44]). Finally, one last method is to simply sequence all cDNA fragments and computationally filter out reads corresponding to rRNA [26,33]. This method may have its appeal because there are no upfront investments of labor or cost to address rRNAs and no biases associated with the rRNA depletion methods. With the depth that can be achieved nowadays, throwing away 99.9% of the reads may still yield a substantial number of reads. Nevertheless, there is a considerable risk that if the necessary depth of sequencing is not obtained, there will be an insufficient number of informative reads for hypothesis generation or testing. Additionally, post-RNA-Seq filtering is not the most cost-effective method because the need to achieve sufficient depth of sequencing likely precludes the use of multiplex sequencing (see below).

In addition to rRNAs, we have also found that a tmRNA can sometimes be very abundant [45]. tmRNA is a bifunctional RNA that acts as both a tRNA and an mRNA in a process called trans-translation (reviewed in [46]). The high representation of tmRNA by RNA-Seq reads makes this gene a candidate worth considering for depletion prior to sequencing. Alternatively, it may be a candidate for post-RNA-Seq filtering. Its extremely high level of expression, relative to non-rRNA-encoding genes, has the potential to upset statistical testing of differential expression.

There are several methods to consider for preparing RNA for sequencing. The most straightforward method relies on sequencing randomly primed cDNAs and is sufficient for discovering genes, improving genome annotations, and assessing the transcriptome for gene expression changes. Strand-specific sequencing, in which the 3' ends of transcripts are defined using a modification to the 3' end prior to cDNA conversion, allows for a more precise interrogation of the transcriptome by distinguishing genes that are overlapping and expressed from different strands. Finally, treatment of RNA with a 5'-Phosphate-Dependent Exonuclease can be used to enrich preprocessed transcripts, which can help resolve alternative transcriptional start positions as well as overlapping and/or nested genes. Sharma *et al*., for example, developed a method they called differential RNA-Seq in which two different preparation methods were used to process fractions of RNA derived from the same sample to distinguish strand-specific 5' preprocessed transcripts [24]. Transcriptional start sites were then determined based on an enrichment of reads from the processed fractions relative

to the unprocessed fractions. Operons were also inferred in combination with bioinformatic predictions and strand-specific sequencing. This approach has provided the most detailed view into the transcriptome of a bacterium so far.

Sharma *et al*., did not fragment or size-select the RNA molecules prior to conversion to cDNA [24]. These steps are common to many RNA processing methods. Fragmentation has the potential to introduce some biases, such as sequence-specific effects on the efficiency of reverse transcription, adaptor ligation, or sequencing. Additionally, as described below, fragmentation has the potential to affect conclusions on differential expression in certain situations. However, skipping the fragmentation and size selection steps has some important considerations. First, this approach limits the sequencing platform that can be used since recommended fragment sizes for the Illumina, for example, are less than 650 bp. Furthermore, regardless of the sequencing platform, cDNAs of longer transcripts may be less represented because cDNA synthesis is done using oligonucleotides complementary to a 3' adapter sequence. Products are further amplified to enrich for products and again to amplify fragments for sequencing. Each of these steps tends to favor shorter products. However, as described below, technical biases that affect all sample preparations similarly are not expected to have major effects on conclusions regarding differential expression.

With the relatively small transcriptome sizes of plant pathogenic bacteria, one can consider using bar coding of different sample preparations and multiplex sequencing to help reduce the cost of RNA-Seq experiments. Bar coding is the

addition of nucleotide sequences that uniquely identify different sample preparations. Multiplex sequencing is simply the pooling of the bar-coded samples for more cost-effective simultaneous sequencing. A concern with this approach is the reduction in the average numbers of reads per gene and decrease in statistical power, *i.e.*, ability to identify truly differentially expressed genes. This is of greater concern with lowly expressed genes. The relation between sequencing depth and percentage of identified expressed genes for an RNA-Seq experiment of *P. syringae* is presented (Figure 2). With just ~3.5 million pre-filtered reads, 95% of the annotated, expressed protein-coding genes are represented by at least 10 RNA-Seq reads, with an average of 190 reads per gene. On an Illumina HiSeq, 3.5 million reads is easily far less than 1/10 of the number of reads expected from a single channel. Ultimately, one has to balance the tradeoff between cost and depth of sequencing. Furthermore, one needs to consider that, as more samples are pooled, there is an increasing challenge in combining approximately equal ratios of cDNA preparations to achieve approximately similar depths of sequencing for all samples. One also needs to consider the barcode sequences. We have observed that some “home-made” barcode sequences dramatically reduced the number of informative reads [45]. Commercially available multiplex sequencing kits are available and likely use rigorously tested and optimized barcodes and barcode combinations.

## COMPUTER GEEK FOR RNA-SEQ

One of the first steps of RNA-Seq data analysis is often the alignment of reads to a reference genome sequence to identify expressed genes (Figure 1A). Many short read alignment programs have been developed and the challenges these programs have in processing RNA-Seq have been comprehensively reviewed [37]. Briefly, one of the important challenges is the assignment of ambiguous reads. These are reads with sequences that can align to more than one locus in the genome and, in the case of eukaryotes, to multiple transcript isoforms. Programs that exclude ambiguous reads will cause genes or transcripts to appear depressed in expression. In contrast, programs that include ambiguous reads have the potential for incorrect assignment, which will also affect detection of gene/transcript expression. An additional concern for transcriptomes of eukaryotes is alternative splicing. A fraction of the RNA-Seq reads will not align to a genome reference sequence because their sequences span yet-to-be discovered splice junctions. Programs with computational and statistical methods to predict transcript isoform structures and assign reads to isoforms have been developed but how they perform for analysis of prokaryotic transcriptomes is unknown.

While splicing is of little concern in the analysis of bacterial transcriptomes, the density of bacterial genomes and the overlapping and nested genes do incur similar challenges in causing ambiguities in the accurate assignment of reads to genes. Based on alignments of RNA-Seq reads to a reference genome sequence of *P. syringae*, only 3% of the reads were considered ambiguous (Figure 1A).

However, this measure is based solely on genome location and does not consider reads that align to the same location encompassed by overlapping genes. Furthermore, our analyses do not take into consideration ambiguities resulting from initiation from alternate start sites. We therefore expect the percent of ambiguous RNA-Seq reads of bacteria to be higher than indicated. Fortunately, as described above, different cDNA preparations for bacteria can be used to help resolve ambiguities.

Data analysis, long-term data storage, and backup are points of concern as researchers increase the scale and scope of their RNA-Seq experiments and improvements in next gen sequencing technology yield more data with longer sequence reads. Of utmost importance is sufficient Random Access Memory (RAM) and processors. RAM acts as a very fast temporary storage space for programs that track large quantities of information. RAM is therefore critical because it directly affects the amount of data that can be analyzed per unit of time before access to the hard drive is required. Processes that rely on the latter are slower by many orders of magnitude.

Researchers may need access to large computing resources. In the absence of institutional infrastructures, cloud computing centers are cost-effective alternatives, e.g., iPlant Collaborative's Atmosphere [47]. A cloud is a computing service that provides access to processors, RAM, and disk space from multiple computers. The cloud handles the distribution of the collective resources to individual programs. The major advantage to cloud computing is their scalability in which users are able to specify the amount of RAM, disk space, and number of

processors needed when requesting for such services. Some RNA-Seq pipelines have been developed to run on a cloud [48,49]. One potential drawback is that the users must operate within the constraints of the cloud infrastructure.

## STATISTICAL ANALYSIS OF RNA-SEQ: EKE! IT'S GREEK TO ME

RNA-Seq has been used to profile gene expression changes of host-associated bacteria [20,50–53]. Comparisons to analysis of microarrays clearly highlighted the advantages in sensitivity and comprehensiveness of RNA-Seq [26]. We emphasize that, if one desires to generalize statistical conclusions from the samples to a population, one has to use independent biological replicates that are representative of the population. Some of the earlier uses of RNA-Seq relied on only technical replicates or unreplicated experiments so the conclusions only applied to the single sample from which the RNA-Seq experiments were based on.

For microarrays, one of the first steps in data analysis is normalization to correct for differences in intensities across microarrays [54]. RNA-Seq data are similar and require normalization to correct for differences in library sizes, which is the total numbers of reads for a sample. A standard approach is to use a measure of relative frequency, such as reads per million mapped reads. The use of relative frequency is not without its potential issues [55]. With a fixed library size (a sequencing run produces only so many reads for any given sample), a change in the relative frequency for some genes will be accompanied by a change in the opposite direction in the relative frequency of reads for other genes (Table 1).

This compensatory change may cause the statistical test to identify other genes as differentially expressed when in fact they are unchanged in their expression. We posit that for the large majority of cases this issue is negligible because the changes in relative frequency will be relatively small and randomly distributed through a substantial number of non-differentially expressed genes. However, problems can be envisioned for cases such as overexpression studies or in characterization of mutant genes with strong pleiotropic effects on gene expression. Methods have been proposed that effectively adjust the library sizes by some normalization factors based on the assumption that the majority of genes are not differentially expressed between different treatment groups [55,56].

Another source of variability is the different transcript lengths present within a transcriptome. Assuming comparable expression levels, genes that encode longer transcripts are expected to produce more fragments and consequently have more assigned RNA-Seq reads than those with shorter transcripts. The longer genes will therefore appear to be more abundantly expressed than comparably expressed shorter genes. Hence, one solution is to normalize per arbitrary number of bases [20,53,57]. This approach has the potential to be misleading when the length of a transcriptional unit is poorly defined, which is the case for bacterial genes belonging to polycistronic operons. Analysis of RNA-Seq derived from host-associated bacteria indicates that a significant number of genes are encoded as operons and that nearly half of the operons display a step-wise decrease in expression [28,39]. The high number of genes expressed from polycistronic operons is supported by computational

predictions in bacterial genomes [58]. As such, unless reads are equally distributed, normalization for transcript length may result in under- and overweighting of a fair number of genes unknowingly contained within an operon. The use of RNA-Seq to first resolve transcriptional units will help to overcome this concern.

After normalization of the data, the task for identifying differentially expressed genes appears simple; it is merely to apply a statistical test for comparing two treatment groups of biologically replicated samples. For analysis of microarrays, this is straightforward because the assumptions of the two-sample *t*-test are met after intensity values are log transformed. This is not the case for RNA-Seq data because the comparison is based on groups of read counts and their probability distribution cannot be approximated by a normal distribution, even after transformation. Our studies using simulated data have shown that *t*-tests are greatly underpowered and will give an unacceptably high false negative rate [59]. In other words, many truly differentially expressed genes would be missed. Thus, the tools developed for analysis of microarrays do not appear appropriate for analysis of RNA-Seq data.

The Poisson probability distribution is a natural alternative to the normal for read count data. However, the inappropriateness of the Poisson distribution for RNA-Seq data has been repeatedly demonstrated [48,56,59]. The reason is a phenomenon called overdispersion where the observed inter-library variability is substantially greater than that predicted by the Poisson model. Because of overdispersion, the variability between groups, including variability between

biological replicates, will cause a Poisson test to have an actual false discovery rate substantially greater than the nominal rate [59].

When choosing a statistics package for data analysis, the appropriateness of the method in addressing small sample size and overdispersion should therefore be considered. Several packages are available, including the updated version of Cuffdiff from the Cufflinks suite of tools, edgeR, DESeq, NBPSeq, Myrna, and LOX (http://cufflinks.cbcb.umd.edu/, [48,56,59–62]). The first four packages use the negative binomial (NB) probability distribution because the NB offers a richer model for count variability. The NB distribution can be considered as a gamma mixture of Poisson distributions. In other words, the Poisson distribution explains the technical variability and the gamma distribution explains the variability between biological replicates [63]. Another important aspect is that the NB distribution permits an exact test for two-group comparisons, which means that it does not rely on large sample size asymptotic theory. For example, the DESeq package was used to analyze an RNA-Seq experiment with only two biological replicates of host-infected *Vibrio cholerae* and identified all known key virulence factors as differentially expressed [26].

There are, however, two practical issues with the use of a NB test. The first is the pooling of information from different genes to estimate the NB “dispersion parameter”, an additional parameter for variation that circumvents the main flaw in Poisson tests. Pooling has an important benefit in providing a higher true discovery rate of differentially expressing genes, *i.e.*, substantially more power in detecting truly differentially expressed genes. For small sample sizes, the power

of the NB test would be substantially greater if the dispersion parameter were known, rather than estimated from the data because much of the information in the data used to compare the means will be sacrificed by the need to estimate the dispersion parameter. Of course, there is no way around the fact that the dispersion parameter is unknown but loss in power can be avoided if commonality in the dispersion parameter across genes can be exploited. For example, in a simple case, the dispersion parameter is the same for all genes and a single estimate can be obtained by pooling the information from all genes. Although each gene would contribute a very small bit of information about the dispersion parameter, the result of pooling from thousands of genes is an estimate that can be essentially treated as known.

In the original edgeR statistics package, the dispersion parameter was indeed assumed to be constant for all genes [61]. While this assumption may hold true for Serial Analysis of Gene Expression (SAGE) data, which was its original intended application, it does not appear to be the case for RNA-Seq data [56,59]. Henceforth, alternative methods were developed that are intermediate to assuming a constant dispersion parameter for all genes and separate dispersion parameters for each gene. The “moderated dispersion” version of the edgeR package uses an empirical Bayes approach, or inference based on the data, to shrink each gene’s dispersion estimates towards a constant value. The “trend option” of edgeR allows the genes’ dispersion estimates to vary around a nonparametric smooth curved function of the mean instead of a constant value. In the DESeq statistics package, the dispersion parameter is modeled as a

nonparametric smooth function of the mean [56]. The most recent updates to the suite of tools of the Cufflinks package include a similar approach as the DESeq method [64]. Finally, in the NBPSeq statistics package, the dispersion parameter is modeled as a simple parametric function of the mean [59].

The second issue with using the NB test is that the mathematical derivation of the exact test requires library sizes to be the same, or at least approximately equal, for all biological samples. Technically, this is a nearly impossible task as several variables beyond the control of the experimental biologist contribute to producing different numbers of reads for each sample preparation. Thus, implementation of the test requires an adjustment to read counts on a scale in which library sizes are equal. The different packages differ slightly in the methods used to adjust library sizes.

In experiments where gene expression is being compared between treatment groups, the variability due to differences in transcript lengths and other technical biases that we have not discussed, are less of an issue, since they presumably affect the same genes to the same degree across different treatment groups. The same cannot be said for other types of analyses that rely on direct or indirect comparisons of expression of a set of genes, such as network or pathway analyses, systems studies, and analysis for enriched gene ontology (GO) terms. Since tests for differential expression are usually more powerful for genes encoding longer transcripts, tests for sets of enriched and differentially expressed genes may be biased towards those that are on average longer in length [65]. To address this issue, a weighted sampling method has been proposed to

compensate for length differences [66]. We note, however, that in the original study, the problem of overdispersion was not well understood and some of the data examples that were characterized did not include biological replicates [65]. When we used NBPSeq to identify differentially induced genes from an RNA-Seq dataset comparing transcriptome changes of a host plant challenged with bacteria versus a mock inoculation, we did not observe substantial correlations between differential expression and transcript length (Figure 3) [67]. We feel that further study is needed to fully appreciate the scope and severity of this so-called “length-bias” issue.

## CONCLUSIONS: RNA-SEQ HAS YET TO PEAK

The use of RNA-Seq to investigate transcriptomes of host-associated bacteria has yielded great insights into their complexity and will do the same to help address our knowledge gap in understanding the lifestyles of plant pathogenic bacteria. Collaborative teams with plant pathologists, computer scientists, and statisticians are essential. There is a need to develop systematic and unbiased approaches for RNA-Seq to help discover genes, refine transcriptional start sites, clarify operon structures, resolve nested genes, and identify differentially expressed genes. Also necessary are new tools for integrating and visualizing large -omic datasets to help biologists formulate hypothesis. There is an urgent demand for statistical methods applicable to more complex experimental designs for RNA-Seq that involve multiple variables such as genotypes of both host and pathogen, communities of bacteria, time after

infection, *etc.* The currently available exact test based on the NB distribution, while more powerful than large sample tests, apply only to two-group comparisons and does not easily extend to the regression setting necessary for characterizing RNA-Seq experiments beyond the simple two-group comparison.

For the plant pathologists, RNA-Seq can be used in combination with ChIP-seq (Chromatin immunoprecipitation coupled with next gen sequencing) and genetic mutants to help define regulons of transcriptional regulators [68]. There will be a great gain in using RNA-Seq to study economically important, but perhaps “non-model” pathogens of food crops. RNA-Seq also has potential use in studying plant pathogens during biologically relevant interactions with their hosts [69]. Thus far, studies of bacteria associated with their hosts have relied on bacterial enrichment to help with subsequent steps of enriching for bacterial RNA [26,32,51,70]. The half-life of prokaryotic RNAs is very short, usually only a number of minutes long. In *E. coli*, for example, total mRNA is estimated to have a half-life of only 6.8 minutes [71]. Thus, the more time-consuming the bacterial purification step, the more likely that host-dependent transcriptome changes will be diminished and conclusions will be biased towards genes with more stable transcripts. To adequately capture biologically interesting transcripts, bacterial enrichment methods require an early step to stabilize RNA that does not cause excessive liberation of RNA from the host.

Another challenge is that, during certain life stages, the low densities of plant pathogenic bacteria may yield insufficient quantities of RNA for sequencing. Even at high densities in culture, there may be transcriptional heterogeneity within

a clonal, synchronized population [72]. A transcriptomic-based investigation of single cells is technically possible, as the transcriptome of a single bacterial cell, captured using laser microdissection and amplified using rolling circle amplification with φ29 DNA polymerase, can yield sufficient quantities of RNA for use in analysis of microarrays [73]. Additional studies have suggested that this method could apply to RNA-Seq, though it has not been explicitly tested.

*P. syringae* has seeded a change to RNA-Seq-based inquiries of plant pathogens [25]. This is befitting, since in addition to being an important model plant pathogen, *P. syringae* is hypothesized to seed clouds, an interesting but challenging niche for an RNA-Seq experiment [74]. Find a cloud, subscribe to a cloud, and start sequencing.

## ACKNOWLEDGMENTS

We thank Sam Fox, Carmen Wong, and Dan Schafer for critical reading of this manuscript and fruitful discussions on statistical analyses of RNA-Seq data. Work in the Chang lab is supported by the National Research Initiative Competitive Grants Program Grant no. 2008-35600-04691 and Agriculture and Food Research Initiative Competitive Grants Program Grant no. 2011-67019-30192 from the USDA National Institute of Food and Agriculture, and National Science Foundation (Grant no. IOS-1021463). JSC was supported by a Computational and Genome Biology Initiative Fellowship from OSU.

## REFERENCES


1. Schneider, D.J.; Collmer, A. Studying plant-pathogen interactions in the genomics era: Beyond molecular Koch's postulates to systems biology. *Annu. Rev. Phytopathol*. **2010**, *48*, 457–479.
2. Baltrus, D.A.; Nishimura, M.T.; Romanchuk, A.; Chang, J.H.; Mukhtar, M.S.; Cherkis, K.; Roach, J.; Grant, S.R.; Jones, C.D.; Dangl, J.L. Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 *Pseudomonas syringae* isolates. *PLoS Pathog*. **2011**, *7*, doi:10.1371/journal.ppat.1002132.
3. Gross, H.; Loper, J.E. Genomics of secondary metabolite production by *Pseudomonas* spp. *Nat. Prod. Rep*. **2009**, *26*, 1408–1446.
4. Depuydt, S.; Trenkamp, S.; Fernie, A.R.; Elftieh, S.; Renou, J.P.; Vuylsteke, M.; Holsters, M.; Vereecke, D. An integrated genomics approach to define niche establishment by *Rhodococcus fascians*. *Plant Physiol*. **2009**, *149*, 1366–1386.
5. O'Brien, H.E.; Desveaux, D.; Guttman, D.S. Next-generation genomics of *Pseudomonas syringae*. *Curr. Opin. Microbiol*. **2011**, *14*, 24–30.
6. Kimbrel, J.A.; Givan, S.A.; Temple, T.N.; Johnson, K.B.; Chang, J.H. Genome sequencing and comparative analysis of the carrot bacterial blight pathogen, *Xanthomonas hortorum* pv. *carotae* M081, for insights into pathogenicity and applications in molecular diagnostics. *Mol. Plant Pathol*. **2011**, *12*, 580–594.
7. Ryan, R.P.; Vorhölter, F.J.; Potnis, N.; Jones, J.B.; van Sluys, M.A.; Bogdanove, A.J.; Dow, J.M. Pathogenomics of *Xanthomonas*: Understanding bacterium-plant interactions. *Nat. Rev. Microbiol*. **2011**, *9*, 344–355.
8. Gelvin, S.B. *Agrobacterium* in the genomics age. *Plant Physiol*. **2009**, *150*, 1665–1676.
9. Toth, I.K.; Pritchard, L.; Birch, P.R.J. Comparative genomics reveals what makes an enterobacterial plant pathogen. *Annu. Rev. Phytopathol*. **2006**, *44*, 305–336.
10. Riley, M.; Abe, T.; Arnaud, M.B.; Berlyn, M.K.B.; Blattner, F.R.; Chaudhuri, R.R.; Glasner, J.D.; Horiuchi, T.; Keseler, I.M.; Kosuge, T.; *et al. Escherichia coli* K-12: A cooperatively developed annotation snapshot—2005. *Nucleic Acids Res*. **2006**, *34*, 1–9.
11. Guo, Y.; Figueiredo, F.; Jones, J.; Wang, N. HrpG and HrpX play global roles in coordinating different virulence traits of *Xanthomonas axonopodis* pv. *citri*. *Mol. Plant Microbe Interact*. **2011**, *24*, 649–661.
12. Ferreira, A.O.; Myers, C.R.; Gordon, J.S.; Martin, G.B.; Vencato, M.; Collmer, A.; Wehling, M.D.; Alfano, J.R.; Moreno-Hagelsieb, G.; Lamboy, W.F.; *et al.* Whole-genome expression profiling defines the HrpL regulon of *Pseudomonas syringae* pv. *tomato* DC3000, allows de novo reconstruction of the Hrp cis clement, and identifies novel coregulated genes. *Mol. Plant Microbe Interact*. **2006**, *19*, 1167–1179.
13. Lan, L.; Deng, X.; Zhou, J.; Tang, X. Genome-wide gene expression analysis of *Pseudomonas syringae* pv. *tomato* DC3000 reveals overlapping and distinct

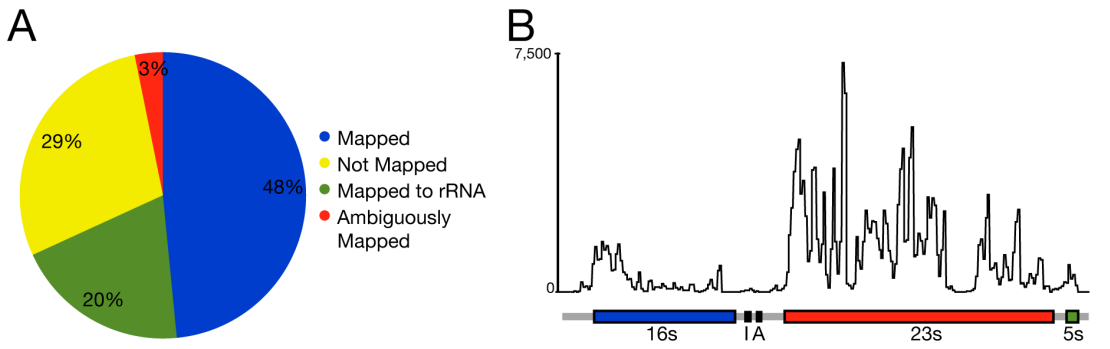
pathways regulated by *hrpL* and *hrpRS*. *Mol. Plant Microbe Interact*. **2006**, *19*, 976–987.
14. Yang, Y.; Zhao, J.; Morgan, R.L.; Ma, W.; Jiang, T. Computational prediction of type III secreted proteins from Gram-negative bacteria. *BMC Bioinforma.* **2010**, *11*, doi:10.1186/1471-2105-11-S1-S47.
15. Yuan, Z.C.; Haudecoeur, E.; Faure, D.; Kerr, K.F.; Nester, E.W. Comparative transcriptome analysis of *Agrobacterium tumefaciens* in response to plant signal salicylic acid, indole-3-acetic acid and gamma-amino butyric acid reveals signalling cross-talk and *Agrobacterium*-plant co-evolution. *Cell. Microbiol*. **2008**, *10*, 2339–2354.
16. Yuan, Z.C.; Liu, P.; Saenkham, P.; Kerr, K.; Nester, E.W. Transcriptome profiling and functional analysis of *Agrobacterium tumefaciens* reveals a general conserved response to acidic conditions (pH 5.5) and a complex acid-mediated signaling involved in *Agrobacterium*-plant interactions. *J. Bacteriol*. **2008**, *190*, 494–507.
17. MacLean, D.; Jones, J.D.G.; Studholme, D.J. Application of “next-generation” sequencing technologies to microbial genetics. *Nat. Rev. Microbiol*. **2009**, *7*, 287–296.
18. Ozsolak, F.; Platt, A.R.; Jones, D.R.; Reifenberger, J.G.; Sass, L.E.; Mcinerney, P.; Thompson, J.F.; Bowers, J.; Jarosz, M.; Milos, P.M. Direct RNA sequencing. *Nature* **2009**, *461*, 814–818.
19. Liu, J.M.; Livny, J.; Lawrence, M.S.; Kimball, M.D.; Waldor, M.K.; Camilli, A. Experimental discovery of sRNAs in *Vibrio cholerae* by direct cloning, 5S/tRNA depletion and parallel sequencing. *Nucleic Acids Res*. **2009**, *37*, doi:10.1093/nar/gkp080.
20. Oliver, H.F.; Orsi, R.H.; Ponnala, L.; Keich, U.; Wang, W.; Sun, Q.; Cartinhour, S.W.; Filiatrault, M.J.; Wiedmann, M.; Boor, K.J. Deep RNA sequencing of *L. monocytogenes* reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. *BMC Genomics*. **2009**, *10*, doi:10.1186/1471-2164-10-641.
21. Perkins, T.T.; Kingsley, R.A.; Fookes, M.C.; Gardner, P.P.; James, K.D.; Yu, L.; Assefa, S.A.; He, M.; Croucher, N.J.; Pickard, D.J.; *et al.* A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS genet*. **2009**, *5*, doi:10.1371/journal.pgen.1000569.
22. Kolev, N.G.; Franklin, J.B.; Carmi, S.; Shi, H.; Michaeli, S.; Tschudi, C. The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog*. **2010**, *6*, 1–15.
23. Schlüter, J.P.; Reinkensmeier, J.; Daschkey, S.; Evguenieva-Hackenberg, E.; Janssen, S.; Jänicke, S.; Becker, J.D.; Giegerich, R.; Becker, A. A genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium *Sinorhizobium meliloti*. *BMC Genomics* **2010**, *11*, doi:10.1186/1471-2164-11-245.
24. Sharma, C.M.; Hoffmann, S.; Darfeuille, F.; Reignier, J.; Findeiss, S.; Sittka, A.; Chabas, S.; Reiche, K.; Hackermüller, J.; Reinhardt, R.; *et al.* The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **2010**, *464*, 250–255.

25. Filiatrault, M.J.; Stodghill, P.V.; Bronstein, P.A.; Moll, S.; Lindeberg, M.; Grills, G.; Schweitzer, P.; Wang, W.; Schroth, G.P.; Luo, S.; *et al.* Transcriptome analysis of Pseudomonas syringae identifies new genes, noncoding RNAs, and antisense activity. *J. Bacteriol*. **2010**, *192*, 2359–2372.
26. Mandlik, A.; Livny, J.; Robins, W.P.; Ritchie, J.M.; Mekalanos, J.J.; Waldor, M.K. RNA-Seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression. *Cell Host Microbe* **2011**, *10*, 165–174.
27. Waters, L.S.; Storz, G. Regulatory RNAs in bacteria. *Cell* **2009**, *136*, 615–628.
28. Güell, M.; van Noort, V.; Yus, E.; Chen, W.H.; Leigh-Bell, J.; Michalodimitrakis, K.; Yamada, T.; Arumugam, M.; Doerks, T.; Kühner, S.; *et al.* Transcriptome complexity in a genome-reduced bacterium. *Science*. **2009**, *326*, 1268–1271.
29. Martin, J.; Zhu, W.; Passalacqua, K.D.; Bergman, N.; Borodovsky, M. *Bacillus anthracis* genome organization in light of whole transcriptome sequencing. *BMC Bioinforma*. **2010**, *11*, doi:10.1186/1471-2105-11-S3-S10.
30. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat biotechnol*. **2011**, *29*, 644–652.
31. Birol, I.; Jackman, S.D.; Nielsen, C.B.; Qian, J.Q.; Varhol, R.; Stazyk, G.; Morin, R.D.; Zhao, Y.; Hirst, M.; Schein, J.E.; *et al. De novo* transcriptome assembly with ABySS. *Bioinformatics* **2009**, *25*, 2872–2877.
32. Bomar, L.; Maltz, M.; Colston, S.; Graf, J. Directed culturing of microorganisms using metatranscriptomics. *mBio* **2011**, *2*, doi:10.1128/mBio.00012-11.
33. Rosenthal, A.Z.; Matson, E.G.; Eldar, A.; Leadbetter, J.R. RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J*. **2011**, *5*, 1133–1142.
34. Reiter, B.; Pfeifer, U.; Schwab, H.; Sessitsch, A. Response of endophytic bacterial communities in potato plants to infection with *Erwinia carotovora* subsp. *atroseptica*. *Appl. Environ. Microbiol*. **2002**, *68*, 2261–2268.
35. Croucher, N.J.; Thomson, N.R. Studying bacterial transcriptomes using RNA-seq. *Curr. Opin. Microbiol*. **2010**, *13*, 619–624.
36. Sorek, R.; Cossart, P. Prokaryotic transcriptomics: A new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet*. **2010**, *11*, 9–16.
37. Garber, M.; Grabherr, M.G.; Guttman, M.; Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **2011**, *8*, 469–477.
38. Oshlack, A.; Robinson, M.D.; Young, M.D. From RNA-seq reads to differential expression results. *Genome Biol*. **2010**, *11*, doi:10.1186/gb-2010-11-12-220.
39. Sharma, C.M.; Vogel, J. Experimental approaches for the discovery and characterization of regulatory small RNA. *Curr. Opin. Microbiol*. **2009**, *12*, 536–546.
40. Chen, Z.; Duan, X. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods in Mol. Biol.* **2011**, *733*, 93–103.
41. Vester, B.; Wengel, J. LNA (locked nucleic acid): High-affinity targeting of complementary RNA and DNA. *Biochemistry* **2004**, *43*, 13233–13241.

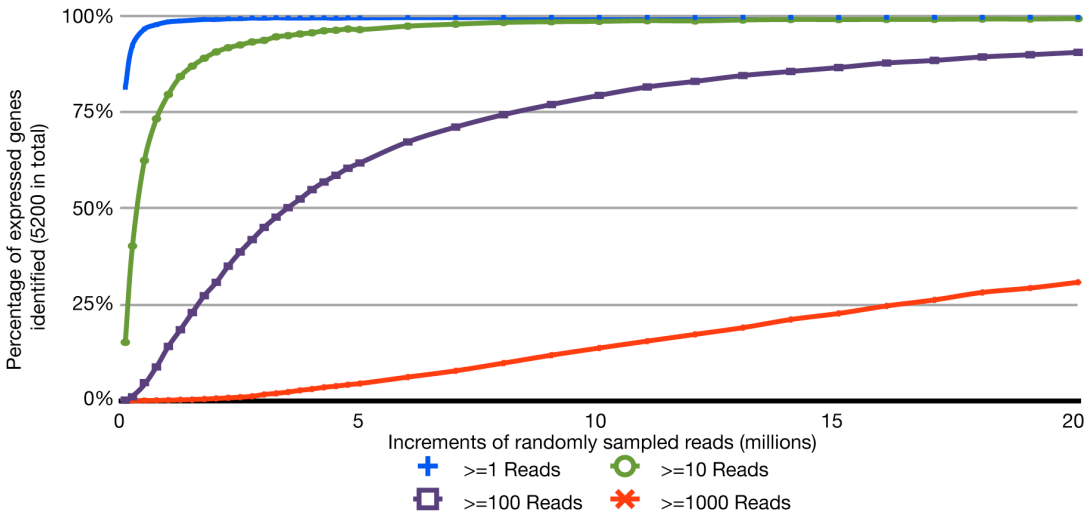
42. Evguenieva-Hackenberg, E. Bacterial ribosomal RNA in pieces. *Mol. Microbiol*. **2005**, *57*, 318–325.
43. Armour, C.D.; Castle, J.C.; Chen, R.; Babak, T.; Loerch, P.; Jackson, S.; Shah, J.K.; Dey, J.; Rohl, C.A.; Johnson, J.M.; *et al.* Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat. Methods* **2009**, *6*, 647–649.
44. Head, S.R.; Komori, H.K.; Hart, G.T.; Shimashita, J.; Schaffer, L.; Salomon, D.R.;
Ordoukhanian, P.T. Method for improved Illumina sequencing library preparation using NuGEN Ovation RNA-Seq System. *BioTechniques* **2011**, *50*, 177–180.
45. Kimbrel, J.A.; Cumbie, J.S.; Chang, J.H. Oregon State University, Corvallis, OR, USA. Unpublished work, 2011.
46. Hayes, C.S.; Keiler, K.C. Beyond ribosome rescue: tmRNA and co-translational processes. *FEBS Lett*. **2010**, *584*, 413–419.
47. iPlant Collaborative. Available online: www.iplantcollaborative.org (accessed on 13 August 2011)
48. Langmead, B.; Hansen, K.D.; Leek, J.T. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol*. **2010**, *11*, doi:10.1186/gb-2010-11-8-r83.
49. Goncalves, A.; Tikhonov, A.; Brazma, A.; Kapushesky, M. A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics* **2011**, *27*, 867–869.
50. Yoder-Himes, D.R.; Chain, P.S.G.; Zhu, Y.; Wurtzel, O.; Rubin, E.M.; Tiedje, J.M.; Sorek, R. Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 3976–3981.
51. Albrecht, M.; Sharma, C.M.; Reinhardt, R.; Vogel, J.; Rudel, T. Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. *Nucleic Acids Res*. **2010**, *38*, 868–877.
52. Camarena, L.; Bruno, V.; Euskirchen, G.; Poggio, S.; Snyder, M. Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing. *PLoS Pathog*. **2010**, *6*, doi:10.1371/journal.ppat.1000834.
53. Isabella, V.M.; Clark, V.L. Deep sequencing-based analysis of the anaerobic stimulon in *Neisseria gonorrhoeae*. *BMC Genomics* **2011**, *12*, doi:10.1186/1471-2164-12-51.
54. Quackenbush, J. Microarray data normalization and transformation. *Nat. Genet*. **2002**, *32*, 496–501.
55. Robinson, M.D.; Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. **2010**, *11*, doi:10.1186/gb-2010-11-3-r25.
56. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol*. **2010**, *11*, doi:10.1186/gb-2010-11-10-r106.
57. Mortazavi, A.; Williams, B.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*. **2008**, *5*, 621–628.

58. Mao, F.; Dam, P.; Chou, J.; Olman, V.; Xu, Y. DOOR: A database for prokaryotic operons. *Nucleic Acids Res*. **2009**, *37*, D459–D463.
59. Di, Y.; Schafer, D.W.; Cumbie, J.S.; Chang, J.H. The NBP negative binomial model for assessing differential gene expression from RNA-seq. *Stat. Appl. Genet. Mol*. **2011**, *10*, 1–28.
60. Trapnell, C.; Williams, B.A.; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M.J.; Salzberg, S.L.; Wold, B.J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*. **2010**, *28*, 511–515.
61. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140.
62. Zhang, Z.; López-Giráldez, F.; Townsend, J.P. LOX: Inferring level of eXpression from diverse methods of census sequencing. *Bioinformatics* **2010**, *26*, 1918–1919.
63. Marioni, J.C.; Mason, C.E.; Mane, S.M.; Stephens, M.; Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. **2008**, *18*, 1509–1517.
64. Cufflinks. Available online: http://cufflinks.cbcb.umd.edu (accessed on 13 August 2011)
65. Oshlack, A.; Wakefield, M.J. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct*. **2009**, *4*, doi:10.1186/1745-6150-4-14.
66. Young, M.D.; Wakefield, M.J.; Smyth, G.K.; Oshlack, A. Gene ontology analysis for RNA-seq: Accounting for selection bias. *Genome Biol*. **2010**, *11*, doi:10.1186/gb-2010-11-2-r14.
67. Cumbie, J.S.; Kimbrel, J.A.; Di, Y.; Schafer, D.W.; Wilhelm, L.J.; Fox, S.E.; Sullivan, C.M.; Curzon, A.D.; Carrington, J.C.; Mockler, T.C.; *et al.* GENE-counter: A computational pipeline for the analysis of RNA-Seq data for gene expression differences. *PLoS One* **2011**, in press.
68. Davies, B.W.; Bogard, R.W.; Mekalanos, J.J. Mapping the regulon of *Vibrio cholerae* ferric uptake regulator expands its known network of gene regulation. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 12467–12472.
69. Allen, C.; Bent, A.; Charkowski, A. Underexplored niches in research on plant pathogenic bacteria. *Plant Physiol*. **2009**, *150*, 1631–1637.
70. Poroyko, V.; White, J.R.; Wang, M.; Donovan, S.; Alverdy, J.; Liu, D.C.; Morowitz, M.J. Gut microbial gene expression in mother-fed and formula-fed piglets. *PLoS One* **2010**, *5*, doi:10.1371/journal.pone.0012459.
71. Selinger, D.W.; Saxena, R.M.; Cheung, K.J.; Church, G.M.; Rosenow, C. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res*. **2003**, *13*, 216–223.
72. Passalacqua, K.D.; Varadarajan, A.; Ondov, B.D.; Okou, D.T.; Zwick, M.E.; Bergman, N.H. Structure and complexity of a bacterial transcriptome. *J. Bacteriol*. **2009**, *191*, 3203–3211.
73. Kang, Y.; Norris, M.H.; Zarzycki-Siek, J.; Nierman, W.C.; Donachie, S.P.; Hoang, T.T. Transcript amplification from single bacterium for transcriptome analysis. *Genome Res*. **2011**, *21*, 925–935.

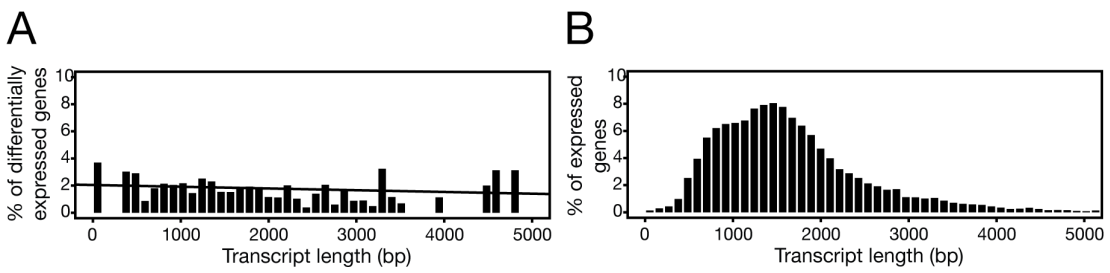
74. Morris, C.E.; Sands, D.C.; Vinatzer, B.A.; Glaux, C.; Guilbaud, C.; Buffière, A.; Yan, S.; Dominguez, H.; Thompson, B.M. The life history of the plant pathogen *Pseudomonas syringae* is linked to the water cycle. *ISME J*. **2008**, *2*, 321–334.




**Appendix IV, Figure 1. Categorization of RNA-Seq reads.**
**(A)** Alignment of 24,202,967 RNA-Seq reads to a *P. syringae* reference genome sequence. The rRNAs were depleted using Ribominus and MicrobExpress. The remaining RNA were converted to cDNA and sequenced on an Illumina IIG using single-direction 40-cycle sequencing. The first 10 and last five bases of each read were trimmed off. The 25 mers were pooled across six samples and aligned using the alignment program, CASHX version 2.3, allowing up to two mismatches. Reads were categorized based on alignment to a unique position (Mapped), the rRNA-encoding locus (Mapped to rRNA), failure to align (Not Mapped), and alignment to multiple locations in the reference genome sequence (Ambiguously Mapped). **(B)** Distribution and frequency of 25 mer RNA-Seq reads that aligned to the rRNA-encoding locus of *P. syringae* following rRNA-depletion. Reads were aligned using CASHX version 2.3.




**Appendix IV, Figure 2. Identification of expressed protein-coding genes as a function of sequencing depth.**
Increments of reads (x-axis) were randomly sampled from the set of ~24 million 25 mer reads (see Figure 1A) and aligned to a *P. syringae* reference genome features derived from the .ptt file (table of protein-coding features). The percent of expressed protein-coding genes discovered, relative to the ~5,200 identified using all 24 million 25 mers, were plotted based on a minimum of 1 (blue), 10 (green), 100 (purple) or 1,000 (red) reads (y-axis).




**Appendix IV, Figure 3. Differential expression as a function of transcript length.**

RNA-Seq data of transcriptomes from *Arabidopsis thaliana* infected with nonpathogenic bacteria or mock inoculated were analyzed using the GENE-counter pipeline configured with the NBPSeq package. **(A)** The differentially induced genes (y-axis) were binned based on equal range of transcript lengths (x-axis). A regression line is plotted. **(B)** Expressed genes from all replicates from both treatments are represented as a percentage within each bin defined based on equal range of transcript length.

**Appendix IV, Table 1. Potential effect of relative frequency on differential expression.**

| Gene name | Relative frequency | | | | | |
|---|---|---|---|---|---|---|
| | Sample 1.1 * | Sample 1.2 * | Sample 1.3 * | Sample 2.1 † | Sample 2.2 † | Sample 2.3 † |
| Gene 1 § | 11 | 13 | 14 | 55 | 52 | 57 |
| Gene 2 | 5 | 4 | 7 | 1 | 0 | 0 |
| Gene 3 | 15 | 20 | 25 | 7 | 10 | 9 |
| Gene 4 | 35 | 37 | 28 | 15 | 19 | 16 |
| Gene 5 | 34 | 26 | 26 | 22 | 19 | 18 |
| Total | 100 | 100 | 100 | 100 | 100 | 100 |

* Samples1.1-1.3 represent biological replicates from treatment group 1. † Samples2.1-2.3 represent biological replicates from treatment group 2. § Gene 1 is differentially induced in treatment group 2 relative to treatment group 1. With the fixed library size, such as an arbitrary number of 100 total reads in this example, an increase in the number of reads for gene 1 in samples 2.1–2.3 will cause compensatory decreases in the number of reads from other expressed genes 2-5 within this treatment group.

# Appendix V: GENE-counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences

Jason S. Cumbie, Jeffrey A. Kimbrel, Yanming Di, Daniel W. Schafer, Larry J. Wilhelm, Samuel E. Fox, Christopher M. Sullivan, Aron D. Curzon, James C. Carrington, Todd C. Mockler, and Jeff H. Chang

PLoS One
2011, vol 6 (10) p. e25279
PMID: 21998647

## ABSTRACT

GENE-counter is a complete Perl-based computational pipeline for analyzing RNA-Sequencing (RNA-Seq) data for differential gene expression. In addition to its use in studying transcriptomes of eukaryotic model organisms, GENE-counter is applicable for prokaryotes and non-model organisms without an available genome reference sequence. For alignments, GENE-counter is configured for CASHX, Bowtie, and BWA but an end user can use any Sequence Alignment/Map (SAM)-compliant program of preference. To analyze data for differential gene expression, GENE-counter can be run with any one of three statistics packages that are based on variations of the negative binomial distribution. The default method is a new and simple statistical test we developed based on an over-parameterized version of the negative binomial distribution. GENE-counter also includes three different methods for assessing differentially expressed features for enriched gene ontology (GO) terms. Results are transparent and data are systematically stored in a MySQL relational database to facilitate additional analyses as well as quality assessment. We used next generation sequencing to generate a small-scale RNA-Seq dataset derived from the heavily studied defense response of *Arabidopsis thaliana* and used GENE-counter to process the data. Collectively, the support from analysis of microarrays as well as the observed and substantial overlap in results from each of the three statistics packages demonstrate that GENE-counter is well suited for handling the unique characteristics of small sample sizes and high variability in gene counts.

## INTRODUCTION

The highly parallelized deep sequencing of cDNA fragments in RNA-Sequencing (RNA-Seq) is the new method of choice in transcriptomics. Its high sensitivity and single-base resolution have contributed substantially to advancing our understanding of gene expression [1]. Recent use of RNA-Seq has led to the identification of a substantial number of new transcripts and their genes, an appreciation into the abundance of a diversity of transcript isoforms as well as the diversity of alternative transcriptional start sites [2-5]. RNA-Seq has also been applied to areas of transcriptomics that in the past, were difficult to study, such as RNA editing, allele-specific expression, and study of expression changes in single cells as well as co-cultivated organisms [6-8].

RNA-Seq can be used to quantify and study genome-wide changes in gene expression. Such applications typically start with aligning RNA-Seq reads to a reference sequence to identify all expressed genome features. The numbers of reads per feature are then calculated to derive feature counts and infer expression levels. Finally, a statistical test is applied to normalized feature counts, followed by a collective assessment of significance based on an acceptable false discovery rate (FDR), to identify differentially expressed features with statistical significance [9]. From this point on, we will simply refer to features as genes.

While the use of RNA-Seq for quantifying gene expression is relatively straightforward to conceptualize, RNA-Seq experiments have considerable computational and statistical challenges. The massive quantities of short reads require ultra fast alignment programs that adequately address memory demands.

The volume of data is also of concern if the end user desires systematic storage and management, as well as integration of data into third party software for additional analyses. Importantly, the combination of a large number of comparisons and small sample sizes causes more concern than usual about the power of the statistical test.

The small sample sizes rule out the uncritical use of methods that rely on large-sample asymptotic theory. Elementary tools for the Poisson distribution will over-state differential expression because of overdispersion, the phenomenon where the count variability between biological replicates is substantially greater than that predicted from the Poisson model [10-12]. Failure to address overdispersion will cause the model to incorrectly interpret large variation between biological replicates as evidence of differential expression and provide drastically misleading conclusions [13].

The negative binomial (NB) distribution offers a more realistic model for RNA-Seq count variation and still permits an exact (non-asymptotic) test for differential gene expression [11, 14]. For each individual gene, a NB distribution uses a dispersion parameter to model the extra-Poisson variation between biological replicates. When considering all genes in an RNA-Seq experiment, statistical power of the exact NB test can be gained by sensibly combining information across genes to estimate the dispersion parameter. The constant dispersion version of the edgeR package, for example, estimates a single dispersion parameter for all genes [11, 14].

The assumption that a single parameter is constant across all genes is, however, not met for RNA-Seq data [13]. To address this, the edgeR package (version 2.0.3) includes an option for empirical Bayes estimation of the dispersion parameter for each gene, with shrinkage towards a common value as well as a 'trend' option that shrinks towards a value determined by nonparametric regression of the dispersion parameter on the mean [15]. The DESeq package, also based on the NB distribution, employs nonparametric regression to estimate the dispersion parameter as a function of the mean and treats the estimated dispersion parameters from this model as known [10]. The NBPSeq package uses a test based on a simple over-parameterized version of the NB distribution called the NBP where an additional parameter is introduced to allow the dispersion parameter to depend on the mean [13].

Some computational pipelines such as Cufflinks, Myrna, and ArrayExpressHTS have been developed for analysis of RNA-Seq data for expression changes [12, 16, 17]. Cufflinks is a pioneering pipeline that combines RNA-Seq alignment with inference of transcript isoforms directly from the RNA-Seq reads, and assessment of differential expression of the inferred transcripts [16]. Cufflinks has been updated to use a test based on the NB distribution (http://cufflinks.cbcb.umd.edu/). Myrna can use cloud computing to cost-effectively exploit large computational resources. With this pipeline, only permutation and large-sample likelihood-ratio tests were considered, which do not sufficiently address small sample sizes or the mean-variance dependence in RNA-Seq data [12, 13]. ArrayExpressHTS is an R/bioconductor-based pipeline that combines

processing, data quality assessment, a variety of alignment programs, inference of transcript isoforms, and statistical analysis with Cufflinks or MMSEQ [18]. The latter provides an estimate of expression levels but does not identify differentially expressed genes.

We describe GENE-counter, a simple pipeline with the appropriate statistical tests for studying genome-wide changes in gene expression. GENE-counter is modular and flexible to allow the end user to use different alignment programs, easily change parameters, and use different statistical tests for analysis of differential gene expression and enriched gene ontology (GO) terms. Results are transparent and systematically stored in a MySQL database, a standard format usable by most third party software. To test GENE-counter, we developed a pilot RNA-Seq dataset from *Arabidopsis thaliana* elicited for PAMP-triggered immunity (PTI). In PTI, recognition of conserved pathogen-associated molecular patterns (PAMPs) leads to a number of induced responses, including genome-wide changes in expression that can be detected 6~7 hours post inoculation (hpi) [19]. PTI is intensively studied and has a correspondingly extensive resource of publicly available microarray data that we used for comparative purposes to support our findings. RNA-Seq data were analyzed using GENE-counter and results were well supported by other statistics packages as well as analysis of microarrays. We also compared the performance of GENE-counter to Cufflinks and showed that with these data, results from the two pipelines were considerably different.

## MATERIALS AND METHODS

### Design and implementation of GENE-counter

We used a combination of Perl, MySQL, R, as well as C++ software (CASHX) to develop GENE-counter. Perl handles the decision logic for the overall pipeline flow to call different software packages for specialized needs, such as data storage and querying, statistical analysis, and fast short-read alignment, which were developed using MySQL, R, and C++, respectively. Perl is also used to handle the user-interface implementation of GENE-counter.

GENE-counter has five tools:

**Configuration tool:** this tool is used to configure GENE-counter to leverage available resources, minimize computational overhead, and reduce duplication of effort. There is potential for multiple users to connect to the same reference sequence database with one or more read databases. Similarly, an end user has the option to align the sequences from their read database to multiple installed reference sequence databases, such as different versions of the same genome sequence. All subsequent gene count and alignment data will be stored in an alignment database for each end user. This flexibility enables easy switching of read databases and/or alignment databases to test and compare results produced by GENE-counter when used with different settings such as alignment parameters.

**Processing tool:** this tool includes two modules for processing RNA-Seq reads and aligning sequences to a reference sequence, respectively. In the first module, user-defined information is recorded to describe the RNA-Seq experiment, such as treatments, replicate numbers, date, etc. The RNA-Seq

reads are processed to identify and enumerate the occurrences of each unique sequence within each replicate. Unique RNA-Seq sequences, their occurrence, and an assigned identification number populate the read database. GENE-counter can use RNA-Seq reads produced from any of the next generation sequencing (next-gen) platforms but limited information is stored if a platform other than Illumina is used.

The second module aligns all unique RNA-Seq sequences to features of a reference sequence database. Any alignment program that can output alignments in the SAM format can be used [20]. We configured GENE-counter with CASHX version 2.3, Bowtie, and BWA [21, 22]. CASHX version 2.3 is the default alignment tool. End users will need to configure other alignment programs if desired.

GENE-counter, by default, will generate gene counts using the best alignments produced with the desired alignment program settings, which are easily set by the end user. For instance, if set to allow a maximum of two mismatches, GENE-counter first relies on alignments with perfect matches, after which it will also use alignments that had one and then two mismatches that did not produce alignments with fewer mismatches. The alignments, in conjunction with their read occurrences, are used to derive gene counts for each reference sequence feature. Data are systemically stored in the alignment database.

**Assessment tool:** this tool can be used to assess the quality of the data. The assessment tool interrogates the alignment database and produces summary files that display raw count data, summary counts for types of features annotated

in the reference sequence, and intraclass correlation coefficient (ICC) values for replicates. The ICC is a descriptive statistic that can be used to quantify the degree of resemblance of quantified measurements of samples within a defined group. To derive ICC values, counts are normalized to reads per quarter million after incrementing by one to handle zeroes prior to log transformation and the 'irr' package in R is used to calculate ICC using the log transformed counts [23, 24]. There is no absolute ICC value that determines useable versus unusable replicates. Rather, the end user can inspect the values as a gauge of the quality of the replicates.

**Statistics tool:** this tool uses the NBPSeq statistics package as the default method for assessing the normalized gene counts to produce a list of differentially expressed genes [13]. GENE-counter is also configured for the edgeR and DESeq statistics packages [10, 15]. Normalization was implemented using the built-in normalization methods of each statistics package. For NBPSeq, the function nbp.test() is called with the appropriate counts and parameters, and normalization occurs automatically followed by differential expression analysis. For edgeR, the 'estimateTagwiseDispersion()' function was used, with the 'trend' parameter set to true and using the matrix counts produced by the 'estimateCommonDisp()' function, to read in the matrix of read counts and normalize counts as well as estimate the dispersion parameters [15]. The 'exactTest()' function was used to calculate p-values for each gene. For DESeq, the 'newCountDataSet()' function was used to generate a cds object from the matrix of read counts and a subsequent call to the 'estimateVarianceFunctions()'

was used to generate the variance estimates [10]. The ‘nbinomTest()’ function was called to generate the p-values for differential expression.

The conclusion about evidence for differentially expressed genes is subsequently based on an ordering of p-values and a cutoff for statistical significance to adhere to acceptable false discovery rates [9]. The ‘qvalue’ package in R was used to generate q-values using the p-values generated by the respective statistics packages.

**GORich tool:** the list of differentially expressed genes can be analyzed for enriched gene ontology (GO) terms using any one of three tests available: the parent-child-inheritance, term-for-term, and GOperm analysis methods [25-27].

**Data storage:** GENE-counter records reference sequence definitions, RNA-Seq read sequence alignments, and derived gene count data, in a MySQL relational database.

Details in installing and using GENE-counter are provided in the user’s manuals.

## Improvements to CASHX

A number of changes were made to CASHX version 1.3 [28]. We implemented a simple hashing algorithm that eliminated empty containers corresponding to preamble sequences absent from reference sequences. We further compressed the database to only store corresponding reference sequence coordinates for each of the indexed k-mers. We also changed the order in which information was stored within each container. The reference sequence

coordinates for each k-mer within a preamble container are now sorted based on the sequence of the 16 nucleotides following the preamble, allowing for sorting of 64 bit integers (2 bits for each nucleotide). Implementation of a simple binary search algorithm dramatically reduced the search time within a preamble container by an order of magnitude. Finally, we implemented a mirrored search logic to index reads to their corresponding container(s), similar to the method employed by Bowtie [21]. Two equal-length fragments derived from each query read are used to seed alignments of the read. CASHX uses the integer converted from the seed fragments and increments their integers through all possible mismatch combinations.

Mapping programs were benchmarked in a single thread on a CentOS 5.1 8 Intel Xeon X5355 x86 64-bit processor with 2.66 GHz and 32 GB RAM. For Bowtie and SOAP2, version 0.12.3 and 2.20, respectively were used [21, 29].

### Developing the *Arabidopsis thaliana* reference database

We developed a comprehensive reference database using the genome and transcript annotations in the TAIR9 genome release (www.arabidopsis.org/). The Generic Feature Format (GFF3) file was used to populate a MySQL database with information such as genes, their classifications (e.g. coding, transposable elements, pseudogene, etc.), transcript classifications (mRNA, miRNA, tRNA, rRNA, etc), coordinates, gene features, and the corresponding gene isoforms. Also included were over 18,000 sequences corresponding to splice junction sequences [3].

Information on how GENE-counter can be used to derive count data from either a list of gene features in a reference genome, or transcript features in a reference transcriptome can be found in GENE-counter's user's manual.

## RNA preparation and sequencing

Bacteria were grown in King's B media and infiltrated into plants as previously described [30]. Briefly, we used a syringe lacking a needle to infiltrate the abaxial side of leaves of six-week old Arabidopsis plants. Plants were infected with either the Δ*hrcC* mutant of *Pseudomonas syringae* pv. *tomato* DC3000 (*Pto*DC3000) or mock inoculated with 10 mM $MgCl_2$ 7 hpi. Each treatment was done as biological triplicates with each pair of replicates done at separate times and derived from independently grown plants and bacteria. Total RNA was extracted from leaves at 7 hpi, enriched for mRNA using Poly(A)Purist (Ambion Inc., Austin, TX) and processed for RNA-Seq as described [31]. The replicates were sequenced one per channel using the 36-cycle sequencing kit on an Illumina. Sequencing was done by the Center for Genome Research and Biocomputing core facility at Oregon State University (CGRB; OSU).

## Pre-processing and aligning RNA-Seq reads

Prior to processing, the first six and last five nucleotides from each RNA-Seq read were trimmed. Reads were then aligned allowing up to two mismatches in the alignment as specified in the global configuration file found in GENE-counter; this setting can be changed by the end-user. Only RNA-Seq reads that aligned to features of a single gene locus were considered, which we referred to

as unambiguous and useable reads. In cases where a read sequence aligns to a single gene locus but to multiple gene isoforms, GENE-counter assigned the reads equally to each of the mapped isoforms. Furthermore, to be considered for differential expression, genome features were required to have assignments in all replicates of at least one of the treatments. Settings can be easily modified at the command line when running the statistics tool of GENE-counter.

GENE-counter was benchmarked in a single thread on a CentOS 5.1 8 Intel Xeon X5355 x86 64-bit processor with 2.66 GHz and 32 GB RAM.

**Derivation of MA plot**

M was calculated as the difference between the $\log_2$ average of GENE-counter normalized values for all replicates in Δ*hrcC* and $MgCl_2$ ($\log_2$(Δ*hrcC*) - $\log_2(MgCl_2)$). A was calculated as the average of all $\log_2$ transformed GENE-counter normalized counts (1/2 * (($\log_2$(Δ*hrcC*) + $\log_2(MgCl_2)$)). All normalized counts had 1 added to them prior to log transformation to avoid problems with zeroes.

**Comparing results from GENE-counter with different statistics packages**

Gene expression was calculated by natural log transformation of the average number of raw gene counts for all genes. The percentage of genes was plotted per expression quantile. The plot was generated using the 'plot' function in R [24]. All genes were also ranked according to the p-value assigned by the respective statistics package and used to create a scatter plot of all genes found

significant in pair wise comparisons. Linear regression lines were plotted using the 'lm' function in R [24].

### Analysis of NBPSeq normalization

The findDGE.pl script of GENE-counter was run 1000 times to examine the effects of random thinning used by NBPSeq to normalize gene counts. For each iteration, a random seed was supplied to the '-s' option of the findDGE.pl script to randomize the thinning process. The percentage of times each gene from the original NBPSeq set of 308 induced genes was determined and plotted against their original q-values. The q-value bins were categorized in quantile increments of 0.005.

### Analysis of microarrays

The mRNA labeling, hybridization, and scanning of Affymetrix ATH1 microarrays were done by the CGRB core facility at OSU. Microarrays were normalized using RMA [32]. Significance was determined based on the overlap of genes common to each of four methods: BRAT (corrected p-value ≤ 0.3) [27], LIMMA (p-value ≤ 0.1) [33, 34], PaGE (confidence level ≥ 0.85) [35], and SAM (q-value % ≤ 10%) [36].

To compare against results from analysis of RNA-Seq, a $\log_2$ scatter plot was produced. For the RNA-Seq data, the fold-change values were calculated using the GENE-counter normalized values (Δ*hrcC* versus $MgCl_2$). For the Affymetrix ATH1 data the raw fluorescence values were used to calculate the normalized fold-change values using the Robust Multi-array Analysis

normalization method [37]. The $\log_2$ values were calculated for both ratios, and the RNA-Seq data (y-axis) was plotted against the Affymetrix ATH1 data (x-axis). Estimated regression lines and Pearson's correlation coefficient were calculated using the 'lm()' function and the 'cor()' functions in the R programming language, respectively [24].

**Cufflinks**

The same set of unambiguous and usable reads from each replicate used by GENE-counter, were also used for analysis by Cufflinks. Reads were mapped to the genome reference sequence using either Tophat version 1.1.2. with the flags '--library-type fr-unstranded -m 2' or Bowtie with the flags '-v 2 -f -a --best --strata –S' to most closely match alignment parameters used in running GENE-counter (allowing for up to two mismatches and choosing the best alignments). Bowtie alignments were converted to BAM and sorted for use with Cufflinks using SAMtools version 0.1.6 [20]. Cufflinks version 1.0.2 was run using default parameters on each replicate file. Each replicate 'transcripts.gtf' file created by Cufflinks was then merged with the Arabidopsis annotation using Cuffmerge with the final merged annotation file being used in Cuffdiff as the reference genome annotation. Cuffdiff version 1.0.2 was run to most closely emulate the way GENE-counter data was used by throwing the flags '--emit-count-tables -c 1 --FDR 0.05' with the '-b' flag being supplied the Arabidopsis reference genome sequence in order to use bias correction.

## RESULTS AND DISCUSSION

We developed GENE-counter as a modular pipeline with five tools for processing, aligning, analyzing, and storing RNA-Seq data (Fig. 1; see material and methods). Perl is used to handle the user-interface of GENE-counter, which makes its use relatively easy by only requiring the end user to be familiar with simple commands at the command line.

GENE-counter stores all processed data in a standard relational database and each of its tools therefore use the standard structure query language (SQL) to retrieve data. Thus, in order to run GENE-counter, it requires configured read, alignment, and reference sequence databases. The first two databases will be populated while running GENE-counter to contain the RNA-Seq reads and alignment information, respectively. The reference sequence database should be populated with reference sequences as well as annotation information prior to running GENE-counter. The three databases will be interrogated by each of the tools of GENE-counter to manage and analyze the data.

### Processing tool: alignment programs

The modularity of GENE-counter gives end users a preference in configuring any SAM compliant alignment program. The default configured option is an improved version of the CASHX alignment program [28]. The improved CASHX, version 2.3, is SAM compliant, and like its predecessor, uses a 2 bit-per-base binary format to compress both the RNA-Seq reads and reference sequence database to exhaustively find all possible alignments that meet user-specified criteria [20]. The improvements to CASHX allowed for mismatch alignments and

dramatically increase alignment speed to reduce the time for aligning sequences by almost 20X and memory demands by 1.5X without compromising accuracy (Table 1).

We benchmarked the CASHX ver. 2.3 alignment program against Bowtie and SOAP2 that, like several alignment programs, use the Burrows Wheeler Transformation compressed index to reduce computational weight and increase speed [21, 29] (Table 1). Using simulated data, in which we knew the exact alignments, CASHX and Bowtie were identical in accuracy but slower than SOAP2 in regards to speed. CASHX was marginally faster than Bowtie when mismatches were allowed and showed a greater advantage in alignment time as the size of the dataset increased (data not shown). In contrast, CASHX had a fairly substantial memory demand relative to the other two tested alignment programs. Though, as the number of reads increased, memory demands by SOAP2 exceeded that of CASHX (data not shown).

The memory demands are potentially limiting or end users may simply be less familiar with CASHX. To address these possibilities, we configured GENE-counter for two other alignment programs, Bowtie and Burrows-Wheeler Alignment tool (BWA) [21]. Other options to control memory demands include running fewer instances of alignment programs or using the built-in throttling mechanism to specify the number of sequences processed at a time. We did not exhaustively benchmark BWA or any other alignment programs in the same manner as presented in table 1. We therefore recommend end users to test their alignment program of preference prior to use with GENE-counter. Nonetheless,

when the accuracy of alignment by BWA was examined using reads from a pilot RNA-Seq experiment (see below), results suggested that BWA was similar to CASHX and Bowtie. We did observe differences in how each of the three programs aligned reads with ambiguous bases and used best alignments (data not shown). The default of CASHX is to exclude reads with ambiguous bases and use only the best alignment.

**Benchmarking GENE-counter**

We processed 522 million RNA-Seq reads of 40 nt in length to demonstrate extremes in running parameters of GENE-counter (S. A. Filichkin and T. C. Mockler, unpublished). In one, we maximized speed at the expense of memory by using eight instances of CASHX in the absence of throttle control. The entire process took GENE-counter ~29 hours and memory demands peaked at 17 GB to analyze the greater than half billion RNA-Seq reads (Fig. 1). Similar running parameters using BWA took ~30 hours and memory peaked at 5 GB [22]. In another setting, we emphasized memory demands over speed by using only one instance of Bowtie and maximum throttling to limit memory usage [21]. GENE-counter took ~52 hours but memory demands peaked at only ~1 GB. In both cases, up to two mismatches were allowed and all steps, from populating the read database with raw RNA-Seq reads to assessing data for enriched GO terms, were measured. These examples demonstrate the range in versatility and scalability of GENE-counter to flex to the size of the RNA-Seq experiment and operate within

the limits of an end-user's computer hardware. Running times will vary depending on hardware.

Storing and interrogating information in databases adds a considerable amount of analysis time by GENE-counter. Although this could be considered a disadvantage, it is offset by the substantial timesaving that will be gained in downstream analyses. Most production level desktop and web-based software platforms have application program interfaces (APIs) that interact with MySQL. These data can therefore be easily queried using third party programs. For example, alignment data processed by GENE-counter can be easily pulled into the generic Genome Browser (GBrowse), a robust web-based platform for visualizing genomes, gene features, and expression data [38]. The systematic storage of data contributes to the modularity of GENE-counter and gives each of the tools a high degree of independence, which allowed for the easier path in configuring different alignment programs and statistics packages. It also gives software developers the ability to leverage the comprehensive data querying language of MySQL to quickly extend the utility of GENE-counter to accelerate development of additional analytical methods and distribution tools. If time is of concern, end users can use a preferred alignment program to derive gene counts independent of GENE-counter and provide counts directly to the statistics tool. However, alignment data will not be stored.

**Analysis of a pilot RNA-Seq dataset**

To examine the efficacy of the entire GENE-counter pipeline, particularly the analysis of differential gene expression, we developed a small-scale RNA-Seq dataset using the intensively studied defense response of Arabidopsis (E-GEOD-25818; http://www.ebi.ac.uk/arrayexpress/). We chose this response because of the availability of microarray data that we could use to support results. We isolated, prepared and sequenced cDNA preparations derived from biological triplicates from Arabidopsis infected with either a Δ*hrcC* strain of *Pto*DC3000 or mock inoculated with 10 mM $MgCl_2$ 7 hpi. The Δ*hrcC* strain has a mutation that affects the assembly of the type III secretion system (T3SS). The T3SS is an apparatus required to inject type III effector proteins, which collectively dampen host defenses, directly into plant cells [39, 40]. Without the T3SS, strains are nonpathogenic and elicit PTI.

GENE-counter took ~3.0 hours when eight instances of CASHX were run in parallel, to process and analyze the ~54 million 25 nt-long reads. For the alignments, we allowed up to two mismatches. On average, ~63% of the reads from the Δ*hrcC*-challenged and mock-inoculated Arabidopsis RNA-Seq experiment aligned to the reference sequence database. We further required GENE-counter to only consider reads that aligned to a single annotated feature of an expressed gene, such as 5' and 3' UTRs, exons, splice junctions, and retained introns. Approximately 50% of the total reads met this additional criterion and were termed unambiguous and usable. Thus, based on the replicate with the fewest number of unambiguous and usable reads and our requirement for a

feature to be aligned with reads in all replicates of at least one treatment, 20,045 of the 33,518 genes annotated for Arabidopsis were considered expressed. Intraclass correlation coefficient (ICC) values for the Δ*hrcC* and mock treatments were both considered acceptable with values of 0.8 and 0.88, respectively [23]. The ICC is a quantitative statistic for assessing the degree of similarity of values within a group.

## Statistics tool

The trend version of edgeR, as well as the DESeq and NBPSeq statistics packages use different ways to model the NB dispersion parameter as a function of the mean [10, 15]. The three are similar in the exact test they use and each method provides the same power benefit associated with combining information across genes [13]. We demonstrated through systematic simulation studies that in terms of statistical power and control of false discoveries, the three methods performed similarly to each other and substantially better than alternative test procedures such as *t*-test, a test based on Poisson model, and the constant or moderated dispersion versions of edgeR [13]. We therefore configured GENE-counter with each of the three statistics packages. Since Perl handles the user-interface, end users are not required to use the R statistics programming language.

The NBPSeq package was implemented as the default method and represents the first known practical use of the NBP distribution. The NBP model has the advantage of relative transparency and model simplicity. The NBP does

not require the input of any user-defined parameters. In contrast, tuning parameters are employed by the trend version of edgeR and DESeq to control smoothing of mean-variance and mean-dispersion curves [10, 15]. How to find the best tuning parameters is still a topic of research. Additionally, while these two other methods provide more flexibility, they also run the risk of overfitting and are prone to the impact of potential unstable variance estimation in the extreme range of expression levels, or 'boundary effects' [13].

With a FDR ≤ 5%, GENE-counter running NBPSeq, returned a list of 308 differentially induced and 79 repressed genes in Δ*hrcC*-infected plants relative to mock-inoculated plants (Fig. 2A; Table S1; from hereafter referred to as the 'original NBPSeq set'). GENE-counter running the trend version of edgeR and DESeq identified 308 and 251 induced genes, respectively (Fig. 2B). Of these, 88% and 94% of the genes, respectively, were also in the original NBPSeq set. We plotted the genes identified from the three methods on an expression scale to examine the effects of gene expression levels on detection of differential expression (Fig. 2C). In general, the three methods captured broad and very similar distributions of gene expression levels. A fair proportion of genes unique to edgeR and DESeq were concentrated in the middle of the expression scale, giving a pronounced sharp peak where results from NBPSeq showed more of a plateau. The genes uniquely identified were found distributed throughout the expression scale.

We also compared the p-value rankings for the induced genes identified from each statistical package (Fig. 2D). Again, in general, there were good

correlations in rankings between all pair wise comparisons. For the genes uniquely identified by one method but not the other, the unique genes were still nevertheless highly ranked, typically within the top ~2.5% or 500 of the ~20,000 ranked genes. Our results confirmed our previous findings that all three statistics packages were comparable and therefore suitable options in GENE-counter [13].

In order to use an exact NB test, which does not rely on large-sample asymptotics for assessing differential gene expression, the three statistics packages need to normalize the counts. In other words, the total numbers of reads must be approximately equal in all replicates. The edgeR method uses quantile adjustment, DESeq adjusts the counts by scaling and NBPSeq adjusts gene counts by random thinning [10, 15]. Normalization is suggested to potentially affect the sensitivity of RNA-Seq analysis [41]. With the data tested here, similar results were produced from GENE-counter when run with each of the three different statistics packages, including their corresponding methods for normalization. This observation suggested that the different normalization methods did not have large effects on the results (Fig. 2).

The adjusting of gene counts by random thinning will yield slightly different normalized counts by separate analyses. This method, however, does not have substantial consequences to the overall conclusions on differential gene expression. As evidence, we analyzed results from running GENE-counter 1000 times with NBPSeq and randomly thinned gene counts (Fig. 3). As expected, the trend in consistency of differential expression correlated strongly with increasing significance of q-values. Of the original NBPSeq set of 308 differentially induced

genes, 87% were identified as differentially induced in ≥ 90% of the samples (Fig. 3). Thus, in general, the great majority of genes were consistently identified and thinning will not have substantial impacts on conclusions. There are however, some instances where random thinning could be viewed as undesirable, e.g., one replicate is severely under-sequenced relative to all others. We would encourage an end user to re-sequence the replicate. Nevertheless, an alternative option would be to use one of the other configured statistics packages of GENE-counter.

**Analysis of enriched GO terms**

A careful inspection of descriptions of the original NBPSeq set of differentially induced genes found that 36% of the annotated genes functions were in plant defense or were identified based on differential expression in response to pathogens, wounding, and/or stresses. Another 15% were annotated as being involved in signal perception, transduction, secretion or modification of the plant cell wall. We also analyzed the induced genes using the parent-child-inheritance method available in the GORich tool of GENE-counter and found 124 enriched GO terms (Table S2) [26]. We compared these to enriched GO terms of genes identified from publicly available microarray studies of plant defense [42-49]. A total of 88 enriched GO terms associated with the differentially induced genes were found associated with at least one other microarray study; 62 were found in at least three of the studies. We concluded that the original NBPSeq set of differentially induced genes was similar to those previously found using analysis of microarrays.

**Comparisons with analysis of microarrays**

We used analysis of microarrays as an alternative technical method to globally assess differential induction and provide independent support for the original NBPSeq set of induced genes. We hybridized the same mRNA samples to Affymetrix ATH1 microarrays and identified 366 induced genes (Table S3; GSE25818; http://www.ncbi.nlm.nih.gov/geo/). For comparisons between RNA-Seq- and microarray-based expression studies, we limited the analysis to only genes that were detectable by both methods. As a result, 254 (82%) and 364 (99%) of the genes identified using GENE-counter or analysis of microarrays, respectively, could be compared.

The $\log_2$-fold change of expression for the induced genes identified from the two methods was well correlated (Fig. 4A). As previously noted, stronger correlations were noted for genes with higher levels of expression [50]. Importantly, analysis of microarrays gave strong support for the genes found by GENE-counter and measurable using microarrays, 174 of 254 or 68% of the induced genes, were common to both expression platforms (Fig. 4B). Additionally, of 22 randomly selected induced genes, 20 were confirmed as differentially induced using qRT-PCR (≥2-fold relative expression; data not shown). We also compared results from an independent microarray study most similar to ours, infection of Arabidopsis with a Δ*hrpA* T3SS mutant of *Pto*DC3000 at 6 hpi [46]. We used the same methods to reanalyze these data and arrived at 414 differentially induced genes, which when compared, supported 58% and 57% of

the differentially induced genes identified using GENE-counter and analysis of our microarrays, respectively. Between the two microarray studies, 78% of the differentially induced genes identified using GENE-counter, and measurable by both methods, were supported. Collectively, our analyses suggested the majority of the genes identified using GENE-counter are *bona fide* differentially induced genes.

**Comparison to Cufflinks**

We compared the performance of GENE-counter to Cufflinks version 1.0.2. For alignments, Cufflinks uses Bowtie with a genome reference sequence and TopHat with an optional transcriptome reference annotation to identify splice junctions and guide inference of transcript isoforms, respectively [16]. In contrast, with GENE-counter, an end user can specify genome, transcriptome, or both reference sequences for alignments. A total of 27,968,144 reads were found to be unambiguous and usable based on alignments by GENE-counter. Cufflinks, when given this set of reads, aligned 26,873,027 to the genome and 735,520 to splice junctions. This compared favorably to GENE-counter, which aligned 26,976,496 to the genome and 991,648 to the transcriptome reference sequence. There were some rare and notable differences but they are not expected to be of much consequence; for example, 16,784 reads used by TopHat to infer splice junctions were aligned to the genome reference sequence by CASHX. As expected with the similarities in alignments, there were high correlations in mean gene expression levels for both treatments (Fig. S1).

Despite the congruence of results up to this step of the two pipelines, only ~24% of the 260 differentially induced and significant genes identified by Cufflinks overlapped with the original NBPSeq set of 308 genes (Table S4). Only ~10% of the genes unique to Cufflinks were identified in a minimum of at least one microarray study, with the majority of those found in only one [42-49]. In contrast, ~86% of genes unique to GENE-counter were often identified across several microarray studies (data not shown). Additional attempts that included increasing the 'minimum alignment count' of Cufflinks to filter out low expressing genes, using all reads in Cufflinks, using Bowtie for alignments to skip isoform predictions by Cufflinks, and comparing results to GENE-counter using an exon only reference database for alignments, resulted in no substantial increases in overlap of gene lists (data not shown). Therefore, our comparisons show that, with the settings, databases, and data used, the final outputs of GENE-counter and Cufflinks were dissimilar with no more than 30% overlap.

Results from independent statistics packages and expression platforms were largely in agreement with results from GENE-counter but the same cannot be said for Cufflinks. The different strategies for measuring isoform versus gene expression could partially explain the discrepancy in results. A study suggested that Cufflinks (ver. 1.0.0), but not methods like GENE-counter, could reliably identify differentially expressed genes when simulated total gene counts were held constant and expression was switched *in silico* from all isoforms in one group to exclusively a single isoform in another group [51]. This is, however, a unique

and extreme case and unlikely generalizable to all genes that differed in the comparisons.

The pilot RNA-Seq dataset could also have contributed to the observed differences as statistical analysis of RNA-Seq data has suggested that technical variability can be substantial and is further exacerbated with lower depth of sequencing [52]. We have used GENE-counter to analyze other RNA-Seq datasets and in these few cases, greater depth of sequencing did not appear to improve results. Particularly informative were two independent rRNA-depleted RNA-Seq experiments of *in vitro* grown bacteria. The depth of sequencing amply exceeded the depth achieved with the Arabidopsis dataset and furthermore, analyses were not complicated by the presence of alternatively spliced isoforms. Nevertheless, in one experiment the overlap in differentially expressed genes identified using GENE-counter and Cufflinks was still less than 30% (J. Dangl, and C. Jones; personal communication). In the other, the number of genes identified using Cufflinks was slightly more than 20% the number found using GENE-counter (J. Kimbrel and J. Chang, unpublished).

There are differences in the statistical methods used by the two pipelines. Uncertainties in read assignments are addressed by Cufflinks using maximum likelihood estimates. This approach has the potential to impact conclusions on differential gene expression [51]. Secondly, Cufflinks uses a different statistical test than GENE-counter, but this is very likely minor. It is also unclear to us whether Cufflinks uses an important statistical power saving feature that is used by all three statistics packages configured in GENE-counter. We are reluctant in

speculating whether these explain the differences in results as Cufflinks experienced substantial and multiple recent changes. We encourage end users to consider and test both pipelines to identify the method most suitable for their purposes.

One important consideration is that GENE-counter does not infer transcript isoforms or directly examine their differential expression. This, however, does not preclude the use of GENE-counter for studying differential expression of transcript isoforms. End users can select genome, transcriptome, or both types of reference databases for alignments. The transcriptome databases for many model organisms are continuously updated to include newly discovered transcript isoforms and when combined with the rapid advances in next-gen technology, may contribute to more accurate alignments of RNA-Seq reads to resolve transcript isoforms and homologous genes. Many software programs for *de novo* assembly of transcripts as well as empirical identification of splice junctions and inference of splice variants from RNA-Seq reads are available [16, 53-55]. These programs could be used to first develop a transcript isoform database with empirically supported sequences. This database could then be used by GENE-counter to identify differentially expressed transcript isoforms.

In summary, GENE-counter is a pipeline for analyzing RNA-Seq data for differential gene expression. Its strengths include ease of use, modularity, appropriateness of statistical tests, and systematic storage of data. Additionally, GENE-counter is well suited for studying gene expression changes of prokaryotes as well as non-model organisms with only a transcriptome reference sequence

first inferred directly from the RNA-Seq data using other software programs. GENE-counter and its user's manuals can be downloaded from our website at: http://changlab.cgrb.oregonstate.edu/. GENE-counter is also available for download from sourceforge.net.

**ACKNOWLEDGEMENTS**

We thank Mark Dasenko and Anne-Marie Girard in the Center for Genome Research and Biocomputing for sequencing and microarray support, Rebecca Pankow, Allison Creason, Philip Hillebrand, Gleb Bazilevsky for their assistance as well as Dr. Corbin Jones and Dr. Scott Givan for their fruitful discussions.

## REFERENCES

1. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10(1):57-63.

2. Toung JM, Morley M, Li M, Cheung VG (2011) RNA-sequence analysis of human B-cells. Genome Res. 21(6):991-8.

3. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, *et al.* (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res. 20(1):45-58.

4. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, *et al.* (2011) The developmental transcriptome of *Drosophila melanogaster.* Nature. 471(7339):473-9.

5. Salzberg SL (2010) Recent advances in RNA sequence analysis. F1000 Biol Rep. 2:64.

6. Rosenberg BR, Hamilton CE, Mwangi MM, Dewell S, Papavasiliou FN (2011) Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. Nat Struct Mol Biol. 18(2):230-6.

7. Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, *et al.* (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 21(7):1160-7.

8. Rosenthal AZ, Matson EG, Eldar A, Leadbetter JR (2011) RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. Isme J. 5(7):1133-42.

9. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 100(16):9440-5.

10. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol. 11(10):R106.

11. Robinson MD, Smyth GK (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 23(21):2881-7.

12. Langmead B, Hansen KD, Leek JT (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol. 11(8):R83.

13. Di Y, Schafer DW, Cumbie JS, Chang JH (2011) The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. Stat Appl Genet Mol Biol. 10(1) Article 24.

14. Robinson MD, Smyth GK (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics. 9(2):321-32.

15. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26(1):139-40.

16. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 28(5):511-5.

17. Goncalves A, Tikhonov A, Brazma A, Kapushesky M (2011) A pipeline for RNA-seq data processing and quality assessment. Bioinformatics. 27(6):867-9.

18. Turro E, Su SY, Goncalves A, Coin LJ, Richardson S, *et al.* (2010) Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. Genome Biol. 12(2):R13.

19. Dodds PN, Rathjen JP (2010) Plant immunity: towards an integrated view of plant-pathogen interactions. Nat Rev Genet. 11(8):539-48.

20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.* (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16):2078-9.

21. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10(3):R25.

22. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25(14):1754-60.

23. McGraw KO, Wong SP (1996) Forming Inferences About Some Intraclass Correlation Coefficients. Psychological Methods. 1(1):30-46.

24. R Development Core Team, *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. 2010: Vienna, Austria.

25. Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics. 24(14):1650-1.

26. Grossmann S, Bauer S, Robinson PN, Vingron M (2007) Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. Bioinformatics. 23(22):3024-31.

27. Pandelova I, Betts MF, Manning VA, Wilhelm LJ, Mockler TC*, et al.* (2009) Analysis of transcriptome changes induced by Ptr ToxA in wheat provides insights into the mechanisms of plant susceptibility. Mol Plant. 2(5):1067-83.

28. Fahlgren N, Sullivan CM, Kasschau KD, Chapman EJ, Cumbie JS*, et al.* (2009) Computational and analytical framework for small RNA profiling by high-throughput sequencing. RNA. 15(5):992-1002.

29. Li R, Yu C, Li Y, Lam TW, Yiu SM*, et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 25(15):1966-7.

30. Thomas WJ, Thireault CA, Kimbrel JA, Chang JH (2009) Recombineering and stable integration of the *Pseudomonas syringae* pv. *syringae* 61 *hrp/hrc* cluster into the genome of the soil bacterium *Pseudomonas fluorescens* Pf0-1. Plant J. 60(5):919-28.

31. Fox S, Filichkin S, Mockler TC (2009) Applications of ultra-high-throughput sequencing. Methods Mol Biol. 553:79-108.

32. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 19(2):185-93.

33. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 3:Article3.

34. Wettenhall JM, Simpson KM, Satterley K, Smyth GK (2006) affylmGUI: a graphical user interface for linear modeling of single channel microarray data. Bioinformatics. 22(7):897-9.

35. Grant GR, Liu J, Stoeckert CJ, Jr. (2005) A practical false discovery rate approach to identifying patterns of differential expression in microarray data. Bioinformatics. 21(11):2684-90.

36. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 98(9):5116-21.

37. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B*, et al.* (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 31(4):e15.

38. Stein LD, Mungall C, Shu S, Caudy M, Mangone M*, et al.* (2002) The generic genome browser: a building block for a model organism system database. Genome Res. 12(10):1599-610.

39. Deng WL, Preston G, Collmer A, Chang CJ, Huang HC (1998) Characterization of the *hrpC* and *hrpRS* operons of *Pseudomonas syringae* pathovars *syringae*, *tomato*, and *glycinea* and analysis of the ability of *hrpF*, *hrpG*, *hrcC*, *hrpT*, and *hrpV* mutants to elicit the hypersensitive response and disease in plants. J Bacteriol. 180(17):4523-31.

40. Roine E, Wei W, Yuan J, Nurmiaho-Lassila EL, Kalkkinen N, *et al.* (1997) Hrp pilus: an *hrp*-dependent bacterial surface appendage produced by *Pseudomonas syringae* pv. *tomato* DC3000. Proc Natl Acad Sci U S A. 94(7):3459-64.

41. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 11:94.

42. Denoux C, Galletti R, Mammarella N, Gopalan S, Werck D, *et al.* (2008) Activation of defense response pathways by OGs and Flg22 elicitors in Arabidopsis seedlings. Mol Plant. 1(3):423-45.

43. Glazebrook J, Chen W, Estes B, Chang HS, Nawrath C, *et al.* (2003) Topology of the network integrating salicylate and jasmonate signal transduction derived from global expression phenotyping. Plant J. 34(2):217-28.

44. Mahalingam R, Gomez-Buitrago A, Eckardt N, Shah N, Guevara-Garcia A, *et al.* (2003) Characterizing the stress/defense transcriptome of Arabidopsis. Genome Biol. 4(3):R20.

45. Navarro L, Zipfel C, Rowland O, Keller I, Robatzek S, *et al.* (2004) The transcriptional innate immune response to flg22. Interplay and overlap with Avr gene-dependent defense responses and bacterial pathogenesis. Plant Physiol. 135(2):1113-28.

46. Thilmony R, Underwood W, He SY (2006) Genome-wide transcriptional analysis of the *Arabidopsis thaliana* interaction with the plant pathogen *Pseudomonas syringae* pv. *tomato* DC3000 and the human pathogen *Escherichia coli* O157:H7. Plant J. 46(1):34-53.

47. Truman W, de Zabala MT, Grant M (2006) Type III effectors orchestrate a complex interplay between transcriptional networks to modify basal defence responses during pathogenesis and resistance. Plant J. 46(1):14-33.

48. Tsuda K, Sato M, Glazebrook J, Cohen JD, Katagiri F (2008) Interplay between MAMP-triggered and SA-mediated defense responses. Plant J. 53(5):763-75.

49. Wang L, Mitra RM, Hasselmann KD, Sato M, Lenarz-Wyatt L, *et al.* (2008) The genetic network controlling the Arabidopsis transcriptional response to

*Pseudomonas syringae* pv. *maculicola*: roles of major regulators and the phytotoxin coronatine. Mol Plant Microbe Interact. 21(11):1408-20.

50. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 18(9):1509-17.

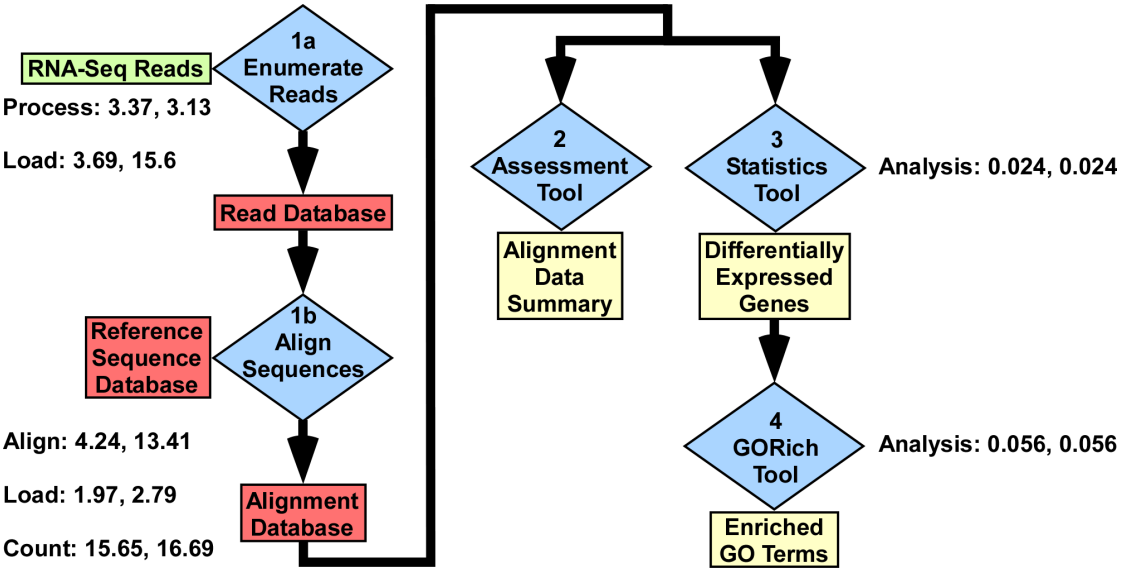
51. Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 8(6):469-77.

52. McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL*, et al.* (2011) RNA-seq: technical variability and sampling. BMC Genomics. 12(1):293.

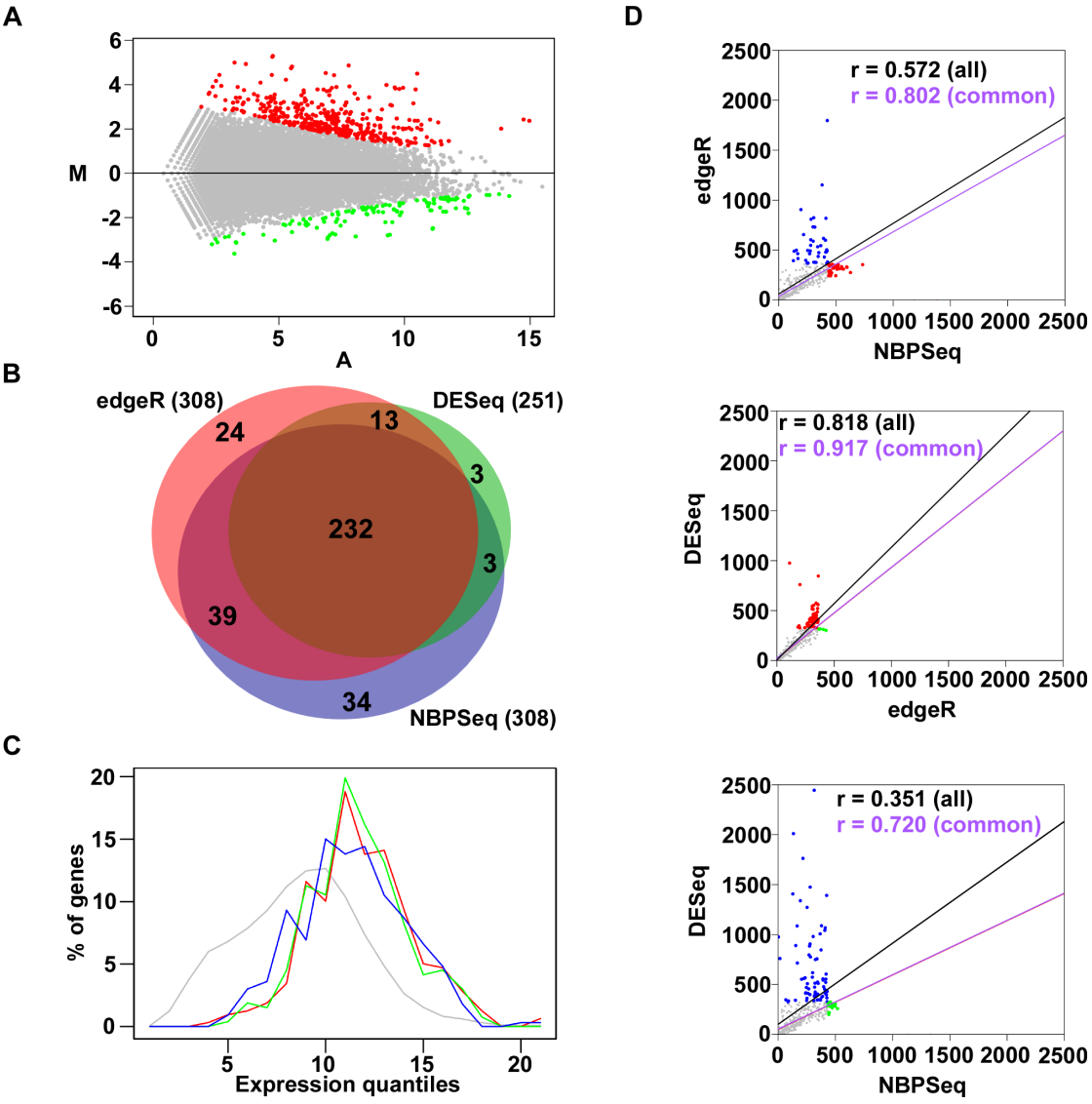
53. Bryant DW, Jr., Shen R, Priest HD, Wong WK, Mockler TC (2010) Supersplat--spliced RNA-seq alignment. Bioinformatics. 26(12):1500-5.

54. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y*, et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 38(18):e178.

55. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644-52.



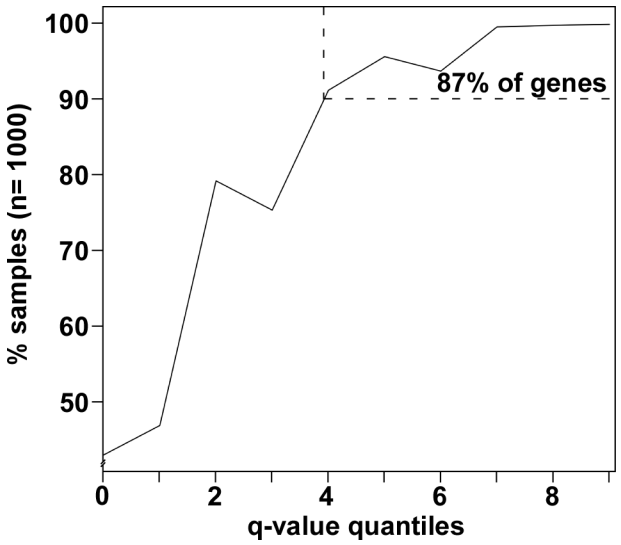

**Appendix V, Figure 1. Entity-relationship diagram for four tools of GENE-counter.**
Each tool is numbered indicating the order in which data is typically processed: **1a and 1b)** the two modules of the processing tool, **2)** the assessment tool, **3)** the statistics tool, and **4)** the GORich tool. The processing tool uses a directory of FASTA files for each replicate as an input (RNA-Seq reads) to tabulate a list of unique read sequences and enumerate the occurrence of each read sequence within each FASTA file. Data are stored in a read database. The processing tool uses a SAM compliant alignment program to align and assign read sequences to features stored in a user-developed reference sequence database. Alignment information and associated count data are stored in the alignment database. Results can be analyzed by the assessment tool to produce an alignment summary, which includes a summary report of replicates and intraclass correlation coefficient (ICC) values. For statistical analysis, the statistics tool can use the NBPSeq, trend version of edgeR, or DESeq statistics package to assess the normalized gene count data. Results are produced as a list of differentially expressed genes, their associated gene counts, normalized gene counts, p- and q-values. The GORich tool can be used to identify enriched Gene Ontology (GO) terms in a list of differentially expressed genes. Three different methods are provided. The amount of time (hours) for steps to analyze over half a billion RNA-Seq reads is shown (GENE-counter running eight instances of CASHX with no throttle control and one instance of Bowtie with maximum throttle control (separated by a comma).



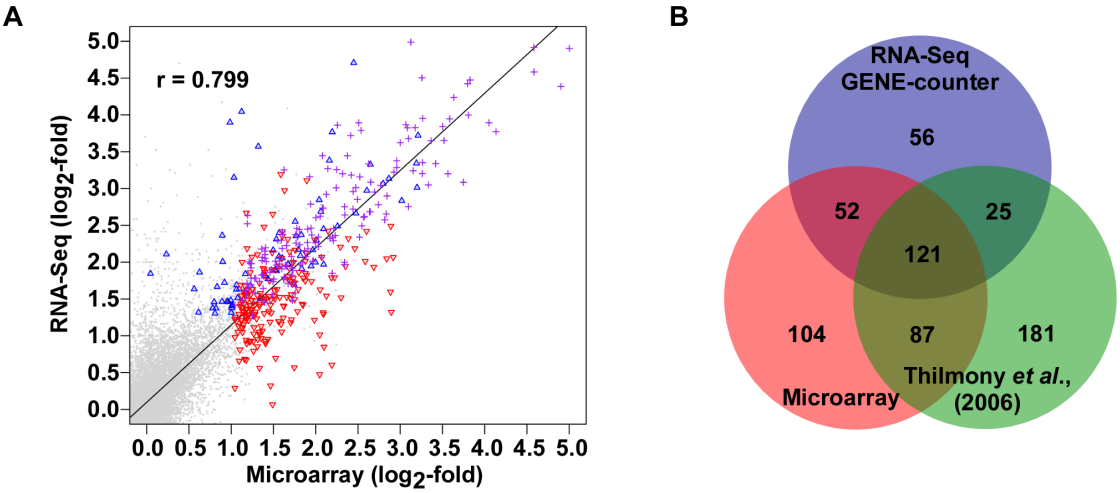

**Appendix V, Figure 2. Analysis of RNA-Seq data for genes differentially expressed in Arabidopsis infected with Δ*hrcC* relative to mock inoculation 7 hpi.**

**(A)** The differentially expressed genes identified between Δ*hrcC*- and mock-treated Arabidopsis. Results are plotted using an MA-based method. Differentially expressed genes were identified using GENE-counter with the NBPSeq statistics package. Induced and repressed genes are highlighted in red and green, respectively (FDR ≤ 5%). **(B)** Area-proportional Venn diagram comparing the differentially induced genes identified using GENE-counter running NBPSeq, the trend version of edgeR, or DESeq. Read counts were normalized using the methods provided in each statistical package prior to analysis (FDR ≤ 5%). **(C)** Distribution of gene expression levels. Percentages of total genes (y-axis) were categorized per expression quantile, increasing from left to right (x-axis; natural

log transformation of average number of normalized aligned reads per gene): gray; all genes; blue, red, and green; differentially induced as identified using GENE-counter running edgeR, DESeq, or NBPSeq, respectively. **(D)** Pair-wise comparisons of p-value rankings for genes identified as significant. Genes were color-coded gray if identified by both statistical packages, blue, red, or green, if uniquely identified by GENE-counter running NBPSeq, edgeR, or DESeq, respectively. Regression lines are plotted based on all genes (black) or only those common to both statistical packages (red). Pearson's r values are shown and colored accordingly.




**Appendix V, Figure 3. Analysis of NBPSeq normalization on differential expression.**
The percent of iterations a gene from the original set was identified as differentially induced (y-axis; n = 1000) was plotted as a function of q-value (x-axis; q-values determined for the original set of differentially induced genes categorized in quantile increments of 0.005 from least significant (q-value = 0.05) on the left to most significant (q-value = 0) on the right). For each iteration, different random number seeds were used to randomly thin gene counts. The percentage of genes found in ≥ 90% of the samples is indicated.




**Appendix V, Figure 4. Comparison of analysis of RNA-Seq with analysis of microarrays.**

**(A)** Comparison of estimated $\log_2$-fold changes from analysis of microarrays (x-axis) and RNA-Seq using GENE-counter running NBPSeq (y-axis). Only induced genes measurable by both platforms are presented. Differentially induced genes are colored to highlight genes uniquely identified using microarrays (open red down triangles) or RNA-Seq (open blue up triangles) and found common between the two methods (purple crosses). **(B)** Three-way Venn comparing differentially expressed genes identified from GENE-counter's assessment of RNA-Seq data and analysis of microarrays. Only genes measurable using both methods were included in the comparison.

**SUPPORTING INFORMATION**

**http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0025279#s4**

**Table S1: Differentially expressed genes identified using GENE-counter running NBPSeq**

**Table S2: Enriched GO terms for the original set of differentially induced genes**

**Table S3: Differentially induced genes from analysis of microarrays**

**Table S4: Differentially expressed genes identified using Cufflinks**

**Fig. S1: Comparison of the log of the mean gene expression values determined by GENE-counter and Cufflinks.**

**Appendix V, Table 1. Benchmarking CASHX ver. 2.3**

| Mapping program* | Clock time (min) | Peak memory usage (Mb) | Alignments identified | Missed alignments (% of the ~8.8 million expected found)† | Unsupported alignments‡ |
|---|---|---|---|---|---|
| **0 mismatches§** | | | | | |
| CASHX ver. 2.3 | 3.70 | 2.32 | 8,815,743 | 0 (100%) | 0 |
| CASHX ver. 1.3 | 73.23 | 3.48 | 8,815,743 | 0 (100%) | 0 |
| SOAP2 | 1.71 | 0.79 | 8,815,745 | 2 (<100%) | 4 |
| Bowtie | 3.22 | 0.13 | 8,815,743 | 0 (100%) | 0 |
| **2 mismatches§** | | | | | |
| CASHX ver. 2.3 | 16.32 | 2.32 | 9,138,971 | 0 (100%) | 0 |
| SOAP2 | 8.85 | 0.81 | 9,094,436 | 44,576 (100%) | 41 |
| Bowtie | 20.38 | 0.19 | 9,138,971 | 0 (100%) | 0 |

*CASHX ver. 1.3 does not allow for mismatches and was not benchmarked for all tests [21, 28, 29]. §We derived a simulated RNA-Seq dataset from 8,815,743 regions of the Arabidopsis genome that were unique in sequence and lacked any Ns for use in benchmarking CASHX ver. 2.3. †For no mismatches, values are based on expected unique alignments; for two mismatches, values are based on the number of alignments confirmed by at least two software programs. ‡Number of alignments that were not confirmed by at least one of the other tested software programs.