AN ABSTRACT OF THE DISSERTATION OF

Jason S. Cumbie for the degree of Doctor of Philosophy in Molecular and Cellular Biology presented on April 22, 2013.
Title: Alternative Splicing in the Obligate Biotrophic Oomycete Pathogen, *Pseudoperonospora cubensis*

Abstract approved:

_____

Jeff H. Chang

Oomycetes are some of the most devastating pathogens, causing upwards of billions of dollars of damage each year to plants. They also diminish ecological diversity and health through the destruction of trees and shrubs. The genome sequence of *Pseudoperonospora cubensis*, an obligate plant pathogen and causative agent of downy mildew in cucurbits, was originally generated as a first step for discovering candidate virulence genes. Through these efforts, the novel discovery was made that a multidrug transport encoding gene was alternatively spliced, giving rise to a truncated protein that, unlike the full length form, exhibits characteristics consistent with *in planta* virulence functions. Alternative splicing can generate different combinations of gene sequences, thereby increasing transcriptome and proteome complexity to influence gene regulation and phenotypic plasticity. Because of the limited number of studies, the impact of alternative splicing on virulence and development of oomycetes is unknown. To address this knowledge gap, we used RNA-Seq to deeply sequence *Ps. cubensis* transcriptomes to assess the impact of alternative splicing during its infection of the host *Cucumis sativus* (cucumber). In addition a number of computational and statistical tools will be described

that were developed to help improve the draft genome and faciliate the characterization of alternative splicing. We demonstrate that alternative splicing influences at least 26% of the *Ps. cubensis* genome with potential effects on gene function, thus highlighting its importance in pathogenesis. This work represents the first step towards understanding the role of alternative splicing in an obligate oomycete pathogen and lays the groundwork for further dissecting the role of alternative splicing in pathogenesis.

Alternative Splicing in the Obligate Biotrophic Oomycete Pathogen, *Pseudoperonospora cubensis*

by

Jason S. Cumbie

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented April 22, 2013

Commencement June 2013

Doctor of Philosophy dissertation of Jason S. Cumbie presented on April 22, 2013.

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.

_____

Jason S. Cumbie, Author

ACKNOWLEDGEMENTS

I would like to state my sincere gratitude for all those whose hard work and encouragement has made this possible. I would like to thank my advisor Dr. Jeff Chang for his tireless efforts at encouraging me, challenging me to always dig deeper, and pressing me to persist to the greatest of my ability. Without his constant aid, mentoring, and guidance I would never have been able to make it this far in my studies or accomplish as much as I have. I would also like to extend a warm thanks to Elizabeth Savory and Alyssa Burkhardt, whose work made this thesis possible. There continued insight and discussions have made this entire process an incredibly wonderful experience. Last, but definitely not least, I would like to thank my wife and love of my life Rachel. Her love, support, and endurance has allowed me to continue my education and made this entire process bearable. Without her presence in my life, my prospects and joy would be greatly diminished in every possible way.

CONTRIBUTION OF AUTHORS

Chapter 1: JSC wrote the manuscript.

Chapter 2: JSC developed the computational pipeline. JAK prepared and sequenced the RNA-Seq libraries. LJW did the microarray bioinformatica analysis. JSC, JAK, YD, DS, and JHC wrote the manuscript.

Chapter 3: This work is a collaborative effort with Dr. Brad Day's lab at Michigan State University. MSU: EAS prepared the RNA for sequencing; AB prepared cDNA and did the RT-PCR experiments to validate newly identified genes. OSU: JSC, EAB, and EAS did the bionformatic analysis of the genome; JSC, AB, EAS, BD, and JHC wrote the manuscript.

Chapter 4: This work is a collaborative effort with Dr. Brad Day's lab at Michigan State University. MSU: EAS prepared the RNA for sequencing. AB prepared cDNA and did the RT-PCR experiments to validate newly identified genes. OSU: JSC, EAB, and EAS did the bionformatic analysis of the genome. JSC, AB, EAS, and JHC wrote the manuscript.

Chapter 5: JSC wrote the manuscript.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

Figure Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF TABLES

LIST OF APPENDICES

Appendix Page

LIST OF APPENDIX FIGURE

Figure Page

Introduction: Plant-microbe Interactions and Next Generation Sequencing

Jason S. Cumbie

INTRODUCTION

The plant immune system is a barrier that pathogens must overcome. It consists of two layers: PAMP Triggered Immunity (PTI) and Effector Triggered Immunity (ETI). PTI recognizes conserved pathogen associated molecular patterns (PAMPs), such as flagellin, that are shared by many pathogens. Recognition of pathogens in ETI is via direct interactions with effector proteins, or the perception of pathogen effector activity on a host target protein. Effector proteins are used by pathogens to overcome or suppress host immunity.

Much of what has been learned about PTI and ETI has been gained through the characterization of model plant pathogen systems (Dangl and Jones 2001; Jones and Dangl 2006). While pathogens may differ markedly in their biology and life styles, they all must evade or suppress PTI. A common strategy used by pathogens is the deployment of secreted effector molecules. For instance, many Gram-negative bacterial pathogens employ a secretion system known as the Type III Secretion System (T3SS) that directly injects upwards of 30+ effector molecules in the host cytoplasm to suppress host immunity (Feng and Zhou, 2012). Similarly, oomycetes secrete upwards of hundreds of effector molecules via the general secretory pathway into the apoplast, with some effectors translocating into the host cytoplasm through a poorly understood process to suppress plant immunity (Schornack et al., 2009).

Recent advances in sequencing have contributed to cost-effective and deep as well as high resolution investigations of plant and microbial transcriptomes (Kunjeti et

al., 2012). This application of technology is especially helpful for gaining insights into genetically intractable organisms, such as obligate plant pathogens, since many techniques, e.g. mutagenesis screens, used in these systems would require comprehensive sequencing to quickly identify potential target genes. In the following sections, I will provide an overview of the plant immune system, discuss important model systems that lay the foundation for this work, and briefly highlight the advances made possible in this area of study through the use of high throughput sequencing.

PLANT IMMUNITY

Plant basal immunity involves the recognition of conserved microbial/pathogen associated molecular patterns (MAMPs/PAMPs), such as flagellin, via surface-associated proteins known as pattern recognition receptors (PRRs) (Zipfel and Felix 2005). The perception of these conserved molecular patterns by PRRs in turn initiates a cascade of signaling events that induce a number of general responses that collectively contribute to immunity. Responses include production of anti-microbial compounds, bursts of reactive oxygen species (ROS), activation of mitogen-activated protein kinases (MAPK), and deposition of callose (Nicaise et al., 2009; Monaghan and Zipfell, 2012). The most immediate of these responses is the influx of ions, such as Ca^{2+}, which occurs within minutes (0.5 – 2 min) of PAMP perception. Calcium-binding to such proteins as calmodulin and calcium-dependent protein kinases further transmits the perception of PAMPs (Reddy and Reddy, 2004). ROS are reduced oxygen forms that are primarily formed from the activity of membrane-localized NADPH oxidases and act as signaling

molecules (Torres et al., 2006). MAPKs are important protein kinases involved in a number of developmental processes and are critical signaling molecules in the initiation of defense-related gene expression, such as up-regulation of the WRKY family of transcription factors (Eulgem and Somssich, 2007). The accumulation of the plant β-1,3-glucan polymer, referred to as callose, at the cell wall and plasma membrane interface is one of the most well recognized outputs of PTI (Luna et al., 2011). It primarily acts to both thicken the cell wall, and to increase the barrier to microbial infection while providing a matrix for the deposition of anti-microbial compounds.

Many host-associated microbes deploy effector proteins to counter PTI. For example, HopA1, a *Pseudomonas syringae* effector, dephosphorylates MAPK3 and MAPK6 to directly interfere with MAPK signaling during PTI (Zhang et al., 2007). RxLR effectors are modular proteins used by oomycetes to overcome plant defense. They consist of an N-terminal signaling domain, an RxLR motif, and a variable C-terminal domain. While the general function of RxLRs is poorly understood, their role in pathogenesis has been demonstrated. In the oomycete *Hyaloperonospora parasitica*, the effector protein ATR13 has been shown to suppress callose deposition induced by *P. syringae* infection suggesting a general role in suppressing basal plant immunity (Sohn et al., 2007). Similarly, Avr1b, an RxLR effector of *P. sojae*, has been shown to increase pathogen virulence when overexpressed (Sohn et al., 2007).

As a counter-adaptation to pathogen effectors, plants employ a second layer of plant immunity referred to as Effector Triggered Immunity (ETI). This layer of defense

either encodes a resistance or R-gene that recognizes an effector gene or the activities of specific effector genes (Jones and Dangl 2006; Hou et al., 2011). The induction of ETI is mediated by resistance or R genes that recognize their cognate effector or effector activity, in which a localized programmed cell death response, referred to as the hypersensitive response (HR), is a possible outcome (Shao 2003; Jones and Dangl, 2006). R-genes encode a modular protein that includes a nucleotide binding domain (NB) and a leucine rich repeat sequence (LRR), and are more generally referred to as NB-LRR genes (Takken and Goverse, 2012).

MODEL PLANT-PATHOGEN SYSTEMS

A crucial advancement in the research of plant immunity and microbial pathogenesis has been the development of model systems for investigation. Two important key players in this development are the model plant system *Arabidopsis thaliana*, and the model plant pathogen *Pseudomonas syringae* pv *tomato* DC3000 which causes bacterial speck on its host. Important outputs of both PTI and ETI, such as callose deposition and HR respectively, have been directly interrogated in *A. thaliana* using screens to elucidate the activities of various cognate effector proteins (Chang et al. 2005; Sohn et al., 2007). Due to the rapid growth and easy maintenance of *A. thaliana* and the genetically tractable nature of *P. syringae*, basic scientific experiments could easily be performed to study the individual genes and gene targets in host-microbe interactions (Jones and Dangl 2006; Mansfield 2009). This is in stark contrast to other more complex systems, such as the obligate plant pathogen *Ps. cubensis*, which cannot be grown outside

of its host, making it more difficult to grow and manage compared to *A. thaliana* and *P. syringae*.

In addition to the ideal functional characteristics of these model systems, major sequencing projects were undertaken to greatly enhance the utility of both *A. thaliana* and *P. syrinage*. In 2000, the genome sequence of *A. thaliana* was published allowing an unprecedented view into the genome wide characteristics of a plant genome. For instance, studying the sequence of *A. thalian*a predicted the existence of over 100 NB-LRR genes, generating a large number of new targets for testing plant immunity (Arabidopsis Genome Initiative, 2000). Not much after this project, the genome of the model plant pathogen *Pseudomonas syringae* would be published (Buell et al., 2003). The sequencing of these respective genomes would allow an expansive genome-wide look into plant immunity and families of effector genes in pathogens that would provide the necessary framework for their bioinformatic analysis (Buell et al. 2003; Feil et al. 2005; Joardar et al. 2005; Jones and Dangl 2006).

Bioinformatic screens were used to identify many new effector genes. This was accomplished by using similarities in the promoters of effector genes to aid scans of the genome. For instance, the Hrp-box promoter in the HrpL regulon, a class of genes crucial in the regulation of the T3SS, was used to discover a number of genes with a similar promoter to identify genes co-regulated with the expression of the T3SS (Deng, 1998; Chang et al., 2005). This would allow functional screens to be carried out and

greatly increase the number of potential effector genes to investigate in pathogenesis (Chang et al., 2005; Lindeberg et al., 2006).

MODEL OBLIGATE PLANT-PATHOGEN SYSTEMS

Through the establishment and use of model plant systems we have greatly enhanced our understanding of plant immunity and pathogenesis. The establishment of simpler model systems provided many new bioinformatic tools developed for these systems to be used to study the genomes and biology of less tractable model systems such as obligate plant pathogens. With the increase in genome sequence availability, and the advancements in sequencing technologies, the development of model obligate plant pathogen systems has been greatly expanded. Oomycetes are one important group containing model obligate plant pathogens that has been gaining further attention recently due to their impact on numerous agriculturally important crops.

Oomycetes are the most destructive pathogens of plants, impacting food security and natural settings worldwide. *Pseudoperonospora cubensis* is the causative agent of cucurbit downy mildew, the most economically important foliar disease of cucurbits (Savory et al., 2011). Fruits such as cucumber, melon, watermelon, squash, and pumpkin are consumed as staple foods throughout the world, and in the U.S. these crops are valued at nearly $1.6 billion annually (Savory et al., 2011). For decades, cucumber cultivars have been bred to be resistant to the obligate oomycete pathogen, *Ps. cubensis*. However, recent epidemics suggest that the widely incorporated cucumber resistance locus, *dm-1* (Vliet and Meysing, 1977), is no longer sufficient in providing durable resistance. A

number of factors contribute to make *Ps. cubensis* an especially devastating pathogen. *Ps. cubensis* has a polycyclic life cycle, propagating and infecting multiple times throughout a growing season (Fig. 1.1). During each cycle, its sporangia have the potential to migrate long distances and infect a broad range of hosts that include more than 50 species of cucurbits grown in more than 70 countries (Lebeda et al., 2003; Savory et al., 2011). Coupled with its ability to develop resistance to fungicides, *Ps. cubensis* is a particular threat to the long-term viability of cucumber production (Blum et al., 2011; Quesada-Ocampo et al., 2012).

Six *Ps. cubensis* pathotypes have been described and all are compatible on cucumber and several melon cultivars (Cohen et al., 2003). Interestingly, the pathotypes display varying degrees of compatibility and disease severity on watermelon, squash, and/or pumpkin, suggesting that the genetic basis of host-range and virulence are linked (Savory et al., 2011). Additionally, developmentally matched *Ps. cubensis* are morphologically different on each of the host plants, suggesting a host-dependent effect (Granke and Hausbeck, 2011). Environment is another component driving downy mildew incidence and severity (Kanetis et al., 2010; Neufeld and Ojiambo, 2012; Schornack et al. 2009). For example, in Michigan, it is not uncommon to see cucumber severely diseased by *Ps. cubensis*, while watermelon displays moderate infection, and yet still, squash and pumpkin in adjacent plots show no signs of the disease. However, in the southeastern U.S., these four species may be simultaneously infected, suggesting an environmental component to host specificity. Indeed, environment has been shown to be a primary factor that affects downy mildew incidence and disease severity (Kanetis et al.,

2010; Neufeld and Ojiambo, 2012; Schornack et al. 2009). Leaf wetness and ambient temperature, in particular, and the effects of their interaction on germination and disease severity, have been quantified in controlled laboratory experiments using a single isolate of *Ps. cubensis* (Schornack et al., 2009).

NEXT GENERATION SEQUENCING AND RNA-SEQ

Next generation sequencing technologies have revolutionized biology. The development of technologies such as the Illumina Sequencer and Roche/454 pyrosequencers have made possible the parallel sequencing of many millions of short (50-400 nt) DNA molecules. These technologies exponentially increased the number of genome sequences available from a broad range of species, e.g., the Japanese quail, pear, woodland strawberry, and numerous microbial genomes, including *Ps. cubensis* (Civelek et al., 2013; Wu et al., 2013; Shulaev et al., 2011; Buell 2003; Tian et al., 2011; Savory et al., 2012).

Transcriptome wide studies have also been greatly enhanced through the development of RNA-Seq, a process that uses next generation technologies to sequence cDNA libraries. For instance, RNA-Seq allows the use of *de novo* assembly methods to construct transcriptome sequences without requiring the existence of a genome sequence. The utilization of *de novo* assembly methods in gene expression studies is especially helpful when studying organisms whose genomes may be otherwise difficult to sequence and/or assemble (Fan et al., 2013; Shanku et al., 2013). This in turn greatly increased the number of species interrogated for their use of RNA splicing, and alternatively spliced RNA transcripts, RNA-editing, and sequence dependent analyses in general that would

not be possible otherwise (Gunaratne et al., 2012; Wang et al., 2012; Trapnell et al., 2011; Bryant et al., 2012). However, software tools, appropriate statistical methods, and in particular, the development of user-friendly packages have not kept pace with advances in technology. Most statistical tools developed require sufficient training in statistics and computer science. For instance, many of the statistical packages such as edgeR, DESeq, and NBPSeq are R statistical packages written using the R programming language that requires direct use of R scripts to leverage the utility of these tools (Robinson and Smyth, 2007; Anders and Huber, 2010; Di et al., 2011).

While sequencing costs have been greatly reduced, current technologies allow for few replicates since one sequencing run (allowing for 7 separate samples) can cost thousands of dollars in addition to the large amount of data produced by each sampling, making it difficult to manage. This is offset by the ultra-deep sequencing of samples allowing for incredible depth and coverage of sample DNA sequences. This approach has by necessity required novel statistical tools, sophisticated data analyses, and the careful examination of experimental design. However, until these tools have greatly matured, the full use of this new technology will be beyond many biologists.

Hereafter, I describe in chapter II, GENE-counter, a user-friendly and flexible software package that was developed for processing and analyzing RNA-Sequencing datasets for differential gene expression. In chapter III, new tools were developed that, in combination with GENE-counter, were used to improve the draft genome sequence of an agriculturally important pathogen, *Ps. cubensis*. Results will be presented that highlights the importance of assessing data for quality and how the RNA-Seq data can be used to

help improve a draft genome sequence. In chapter IV, I characterize the RNA-Seq datasets for splicing and alternative splicing as a first step towards understanding how transcriptome plasticity could contribute to the success of *Ps. cubensis* as a broad-host range, and environmentally pliable pathogen of plants. Finally, in chapter V, I discuss the impact and future directions of my work.

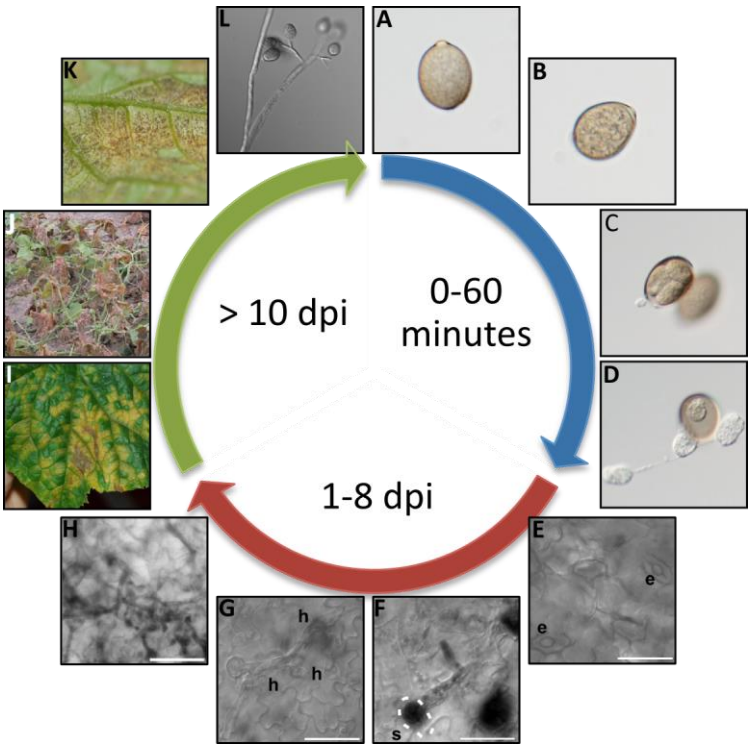


Figure 1.1. The *Ps. cubensis* life cycle.

0-60 minutes (Zoosporogenesis): Dormant sporangia **(A)** begin to differentiate upon contact with moisture **(B)** into 2-11 zoospores which emerge from the distal end **(C)**. Zoospores **(D)** are biflagellate and swim preferentially to stomata. 1-8 days post inoculation (dpi) **(**Host infection): Zoospores enter through stomata by 1 dpi, encysted zoospores are indicated with an "e" **(E)** and grow through intracellular spaces 2 dpi **(F)**. By 4 dpi **(G)** haustoria (labeled with "h") have formed. Colonization of the mesophyll continues through 8 dpi **(H)**. > 10 dpi (Disease symptoms): Yellow angular lesions with necrotic centers are indicative of heavy infestation **(I)**. Fields of plants can be completely defoliated within 14 dpi **(J)**. Sporulation occurs on the lower leaf surface **(K)** when sporangiophores **(L)** emerge through stomata and bear grey sporangia. The inner circle indicates time in minutes or days post inoculation (dpi). Scale bars: E-G, 25 μm; H, 50 μm.

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

Published in:
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 (Wang et al., 2009). 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 (Toung et al., 2011; Filichkin et al., 2011; Graveley et al., 2011; Salzberg, 2010). 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 (Rosenberg et al., 2011; Islam et al., 2011; Rosenthal et al., 2011).

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 (Storey Tibshirani, 2003). 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 (Anders and Huber, 20010; Robinson and Smyth, 2007; Langmead et al., 2010). 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 (Di et al., 2011).

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 (Robinson and Smyth, 2007; Robinson and Smyth, 2008). 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 (Robinson and Smyth, 2007; Robinson and Smyth, 2008).

The assumption that a single parameter is constant across all genes is, however, not met for RNA-Seq data (Di et al., 2011). 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 (Robinson et al., 2010). 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 (Anders and Huber, 2010). 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 (Di et al., 2011).

Some computational pipelines such as Cufflinks, Myrna, and ArrayExpressHTS have been developed for analysis of RNA-Seq data for expression changes (Langmead et al., 2010; Trapnell et al., 2010; Goncalves et al., 2011). 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 (Trapnell et al., 2010). 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 (Langmead et al.,

2010; Di et al., 2011). 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 (Turro et al., 2010). 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) (Dodds and Rathjen , 2010). 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 (Handsaker et al., 2009). We configured GENE-counter with CASHX version 2.3, Bowtie, and BWA (Langmead et al., 2009; Durbin, 2009). 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 (McGraw and Wong, 1996; R Development Core Team, 2010). 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 (Di et al., 2011). GENE-counter is also configured for the edgeR and DESeq statistics packages (Anders and Huber, 2010; Robinson et al., 2010). 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 (Robinson et al., 2010). 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 (Anders and Huber, 2010). 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 (Storey and Tibshirani, 2003). 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 (Bauer et al., 2008; Grossmann et al., 2007; Pandelova et al., 2009).

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 (Fahlgren et al., 2009). 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 (Langmead et al., 2009). 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 (Langmead et al., 2009; Li et al., 2009).

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 (Filichkin et al., 2011).

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 (Thomas et al., 2009). 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 (Fox et al., 2009). 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 (R Development Core Team, 2010). 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 (R Development Core Team, 2010).

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 (Bolstad et al., 2003). Significance was determined based on the overlap of genes common to each of four methods: BRAT (corrected p-value ≤ 0.3) (Pandelova et al., 2009), LIMMA (p-value ≤ 0.1) (Smyth, 2004; Wettenhall et al., 2006), PaGE (confidence level ≥ 0.85) (Grant et al., 2005), and SAM (q-value % $\leq 10\%$) (Tusher et al., 2001).

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 (Irizarry et al., 2003). 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 (R Development Core Team, 2010).

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 (Handsaker et al., 2009). 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. 2.1; 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 (Fahlgren et al., 2009). 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 (Handsaker et al., 2009). 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 2.1).

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 (Langmead et al., 2009; Li et al., 2009) (Table 2.1). 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) (Langmead et al., 2009). 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 2.1. 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. 2.1). Similar running parameters using BWA took ~30 hours and memory peaked at 5 GB (Durbin, 2009). In another setting, we emphasized memory demands over speed by using only one instance of Bowtie and maximum throttling to limit memory usage (Langmead et al., 2009). 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 (Stein et al., 2002). 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 (Deng et al., 1998; Roine et al., 1997). 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 (McGraw and Wong, 1996). 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 (Anders and Huber, 2010; Robinson et al., 2010). The three are similar in the exact test they use and each method provides the same power benefit associated with combining information across genes (Di et al., 2011). 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 (Di et al., 2011). 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 (Anders and Huber, 2010; Robinson et al., 2010). 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' (Di et al., 2011).

With a FDR $\leq$ 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. 2.2A; Table S2.1; 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. 2.2B). 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. 2.2C). 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. 2.2D). 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 (Di et al., 2011).

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 (Anders and Huber, 2010; Robinson et al., 2010). Normalization is suggested to potentially affect the sensitivity of RNA-Seq analysis (Bullard et al., 2010). 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. 2.2).

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. 2.3). 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. 2.3). 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 S2.2) (Grossmann et al., 2007). We compared these to enriched GO terms of genes identified from publicly available

microarray studies of plant defense (Denoux et al., 2008; Glazebrook et al., 2003; Mahalingam et al., 2003; Navarro et al., 2004; Thilmony et al., 2006; Truman et al., 2006; Tsuda et al., 2008; Wang et al., 2008). 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 S2.3; 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. 2.4A). As previously noted, stronger correlations were noted for genes with higher levels of expression (Marioni et al., 2008). 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. 2.4B). 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 (Thilmony et al., 2006). 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 (Trapnell et al., 2010). 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. S2.1).

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 S2.4). 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 (Denoux et al., 2008; Glazebrook et al., 2003; Mahalingam et al., 2003; Navarro et al., 2004; Thilmony et al., 2006; Truman et al., 2006; Tsuda et al., 2008; Wang et al., 2008). 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 (Garber et al., 2011). 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 (McIntyre et al., 2011). 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 (Garber et al., 2011). 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 (Trapnell et al., 2010; Bryant et al., 2010; Wang et al., 2010; Grabherr et al., 2011). 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.

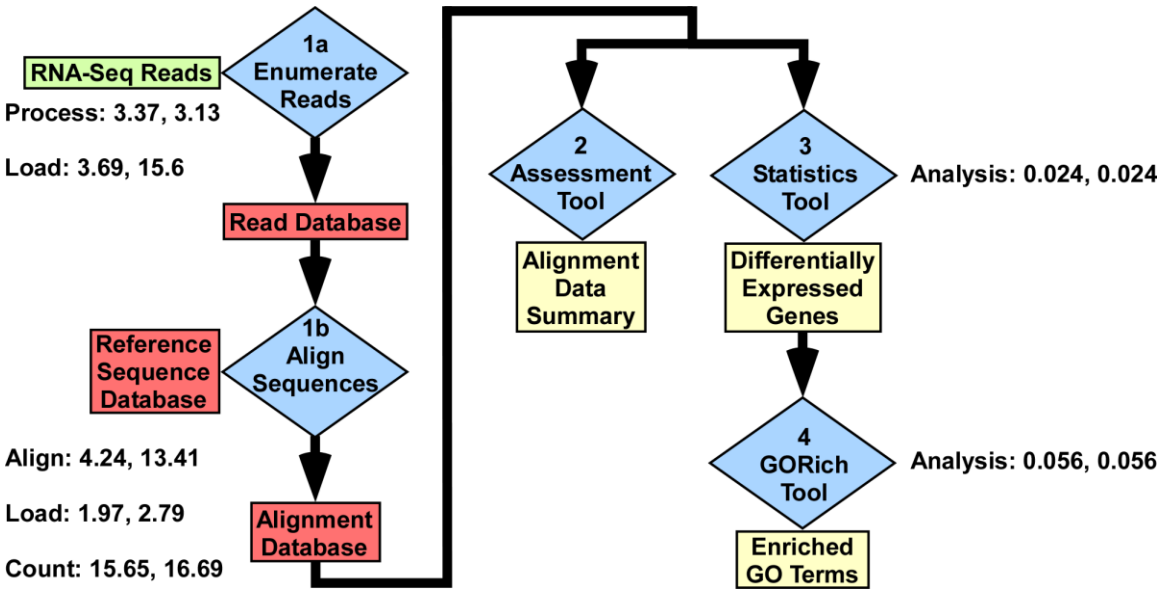


Figure 2.1. Entity-relationship diagram for four tools of GENE-counter.

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).

Figure 2.2. 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.

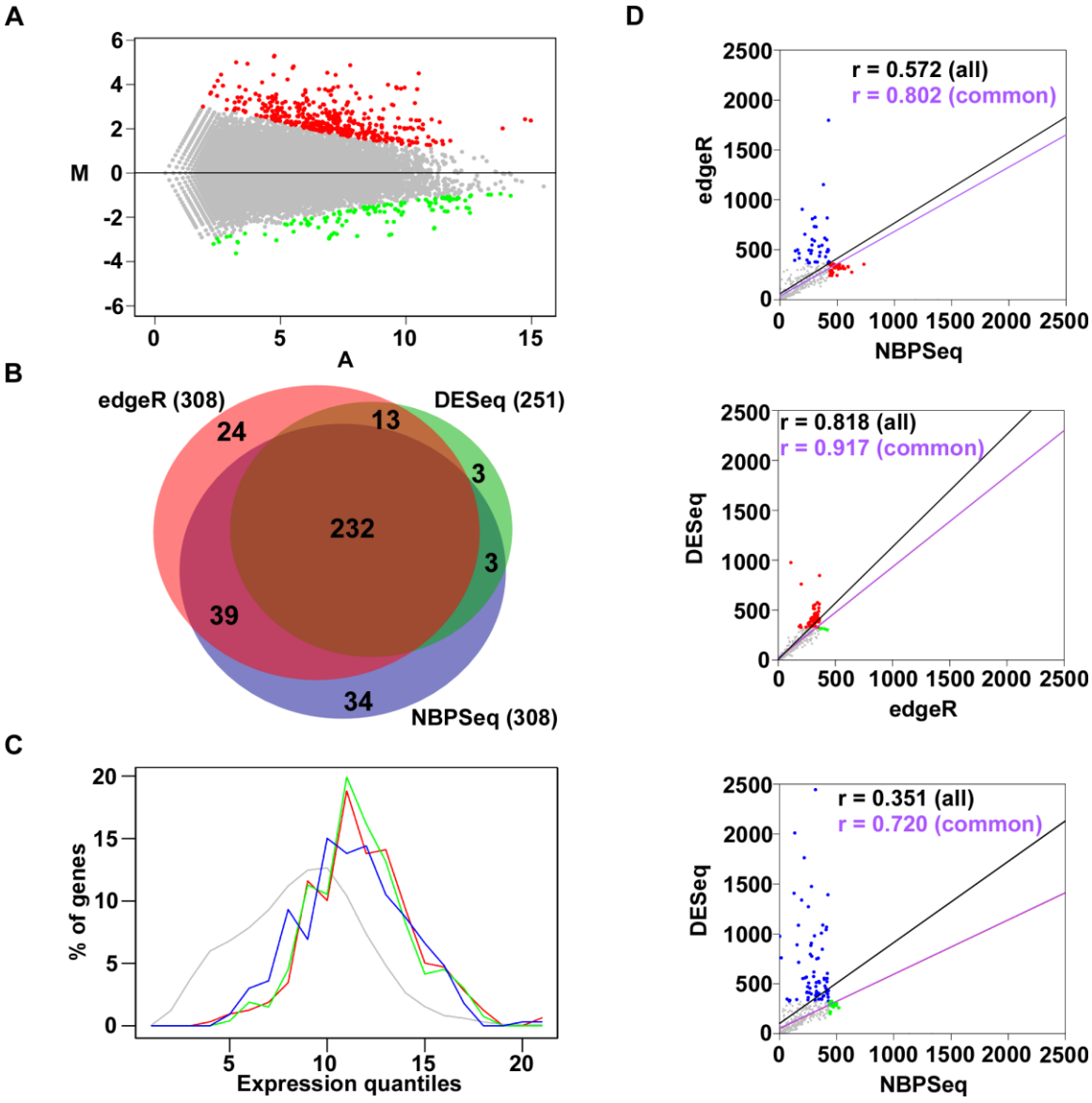


Figure 2.2. Analysis of RNA-Seq data for genes differentially expressed in Arabidopsis infected with Δ*hrcC* relative to mock inoculation 7 hpi.

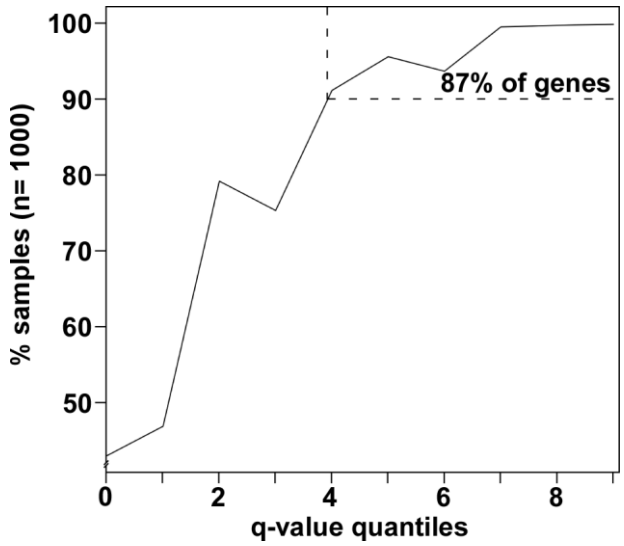


Figure 2.3. Analysis of NBPSeq normalization on differential expression.

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.

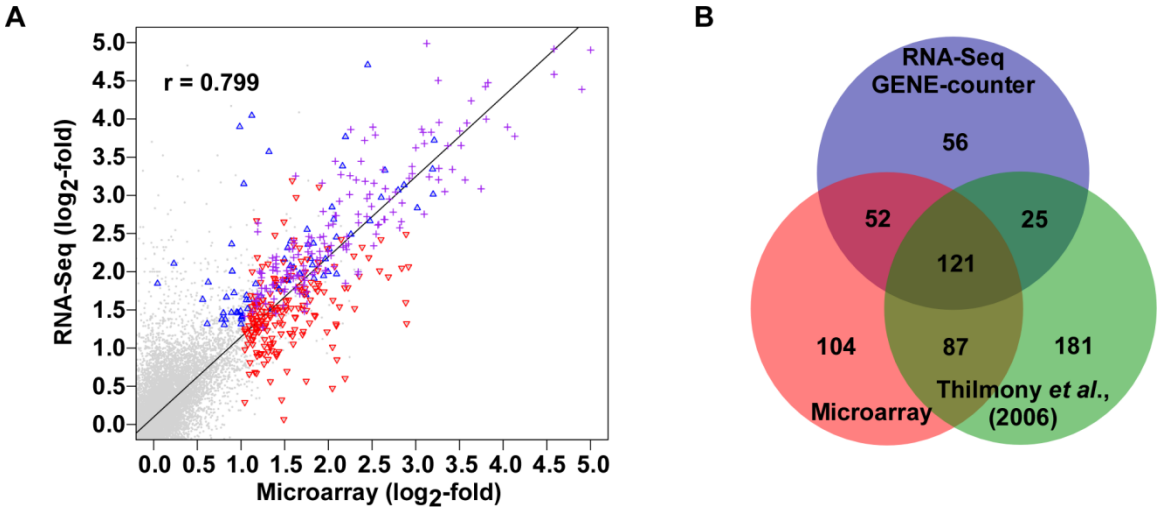


Figure 2.4. Comparison of analysis of RNA-Seq with analysis of microarrays.

(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.

Table 2.1. 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 (Langmead et al., 2009; Fahlgren et al., 2009' Li et al., 2009). [§]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.

Supporting Information

All supplementary information can be found online at http://www.plosone.org.

Table S2.1: Differentially expressed genes identified using GENE-counter running NBPSeq

Table S2.2: Enriched GO terms for the original set of differentially induced genes

Table S2.3: Differentially induced genes from analysis of microarrays

Table S2.4: Differentially expressed genes identified using Cufflinks

Fig. S2.1: Comparison of the log of the mean gene expression values determined by GENE-counter and Cufflinks.

Analysis of Transcriptome Sequences to Improve the Draft Genome Sequence of *Pseudoperonospora cubensis*

Jason S. Cumbie, Alyssa Burkhardt, Elizabeth A. Savory, E. Alex Buchanan, Jeff H. Chang

In preparation for submission to The New Phytologist

INTRODUCTION

Oomycetes are some of the most devastating plant pathogens. These pathogens aggressively attack many different species of plants, with some of the most well-known examples being *Phytophthora infestans*, the cause of late blight in potatoes, as well as *Phytophthora ramorum* which causes sudden oak death (Tyler et al., 2006; Vleeshouwers et al., 2011; Grünwald et al., 2012). Their ubiquity and, in some cases, recent emergence as central players of plant disease in both ecologically and economically important plant species has highlighted the importance of understanding their mechanisms of pathogenecity to better assess and manage these plant pathogens (Savory et al., 2011; Grünwald et al., 2012). *P. infestans* is estimated to cause upwards of $6.7 billion dollars due to its destruction of potato crops whereas *Ps. cubensis* is estimated to cause upwards of $1.6 billion of damage to economically important food staples such as cucumber, melon, watermelon, squash, and pumpkin, (Haas et al., 2009; Savory et al., 2011). The reemergence of sudden oak death over the years has wreaked havoc on numerous ecosystems due to the destruction of trees and shrubs, dramatically impacting the composition and stability of ecosystems (Grünwald et al., 2012). One of the most notable historical examples is the Irish Potato famine caused by *P. infestans* which resulted in millions of deaths and immigration of millions more (Henderson, 2006).

Much of our current knowledge of oomycete biology has been informed by studies utilizing the published genome sequence of model oomycetes *Phytophthora sojae* and *Phytophthora ramorum* (Tyler et al., 2006), and *Phytophthora infestans* (Haas et al., 2009). Through their respective analyses it has been discovered that each encodes

hundreds of effector proteins that are under positive selection, underscoring the rapid evolution and expansion of the repertoire of pathogenesis related genes (Schornack et al., 2009). Another interesting discovery was the potentially wide range (~1100 – 1400) of putatively secreted proteins found within these genomes highlighting the incredible diversity of secreted proteins used by oomycetes to interact with their respective environments (Tyler et al., 2006; Haas et al., 2009).

Oomycetes in general are challenging to study because many are recalcitrant to genetic modifications with some being obligate biotrophic pathogens that require live host tissue for survival (Schornack et al., 2009). Recent advances in sequencing, comparative genomics and analysis of transcriptomes have contributed extensively in advancing our understanding the adaptation and virulence of oomycetes to plant hosts (Huitema et al., 2003; Choi et al., 2012; Savory et al., 2012; Roy et al., 2013).

Ps. cubensis is the causative agent of cucurbit downy mildew, the most economically important foliar disease of cucurbits (Savory et al., 2011). While resistant plant cultivars were primarily used for pathogen control, the emergence of resistant strains has increased our reliance on fungicides to control the spread of disease (Savory et al., 2011). In addition to resistant cultivars, many aspects of *Ps. cubensis* make it especially difficult to control, such as a polycylcic lifecycle allowing for multiple infections throughout its lifecycle, as well as its ability to infect a large number of different hosts under a number of different environmental conditions, particularly leaf moisture and temperature (Lebeda et al., 2003; Neufeld et al., 2011; Arauz et al., 2010; Savory et al., 2011).

Transcriptome sequencing studies have been greatly enhanced through the development of RNA-Seq, a process that uses next generation technologies to sequence cDNA libraries. Using RNA-Seq, it was shown that distinct transcriptional changes are highly correlated with different stages of infection in oomycetes (Whisson et al., 2007; Savory et al., 2012). In *Ps. cubensis,* study of the transcriptome changes using RNA-Seq revealed distinct patterns in the upregulation of a number of effector genes. These changes were highly correlated with upwards of 79 orthologous effector genes in *P. infestans* shown to display a sharp upward peak during the biotrophic phase of infection (Haas et al., 2009; Savory et al., 2012).

The draft genome of *Ps. cubensis* was originally obtained to help identify RxLR effectors (Tian et al., 2011; Savory et al., 2012). Effectors are virulence proteins that function in host cells to counter host immunity and promote pathogen growth and reproduction. RxLR effectors are modular proteins that consist of an N-terminal signal peptide followed by a conserved RxLR domain, and a variable C-terminal effector domain (Morgan and Kamoun, 2007). The signal peptide allows for the secretion of the effector into the host apoplast via its recognition by the general secretory pathway (Morgan and Kamoun, 2007). The RxLR motif has been implicated in the translocation of the effector into the host cytoplasm, where it is hypothesized that the highly variable C-terminal domain acts to suppress plant immunity. It has been shown that the signal peptide and RxLR motifs together are sufficient to translocate an effector molecule into the host cytoplasm (Whisson et al., 2007), however the specific virulence functions of RxLR proteins remains largely unknown (Morgan and Kamoun, 2007). This observation

led to the establishment of an *ab initio* genome-wide screen, which facilitated the identification of a large number of candidate RxLR effectors (Win et al., 2007). RxLR effectors can induce cell death (Bhattacharyya et al., 2005; Bittner-Eddy et al., 2000; Huang et al., 2005) implicating their role in pathogenesis, and indeed the overexpression of the RxLR effector Avr1b-1 increased pathogen virulence on a compatible host, consistent with this hypothesis (Morgan and Kamoun, 2007).

The draft assembly of *Ps. cubensis* contained 38,778 contigs with an N50 contig size of 3.7 kb representing 67.9 Mb (Tian et al., 2011; Savory et al., 2012). The *Ps. cubensis* genome is estimated to be relatively compact (~65 Mb) and dense, encoding an unusually high number of ~23,500 genes relative to other sequenced oomycetes (Savory et al., 2012; Jiang and Tyler, 2012). Protein coding genes in the draft assembly were annotated using a combination of *ab initio* gene predictions, protein evidence, and transcript evidence from other sequenced oomycete genomes (Savory et al., 2012). The size of the *Ps. cubensis* genome is similar to other oomycetes with more streamlined genomes (Tyler et al., 2006); however, the number of genes exceeds predictions of other comparable oomycete genome annotations (Tyler et al., 2006; Haas et al., 2009; Baxter et al., 2010; Lévesque et al., 2010).

In this study, we used deeply re-sequenced RNA-Seq libraries from *C. sativus* infected with *Ps. cubensis* to improve its draft genome sequence as a first and necessary step for the work described in chapter IV of this thesis. Microscopy analysis also highlighted the high potential of contaminating sequence that could further reduce the quality of the genome sequence annotation. Additionally the number of gene models is

equal to the number of genes in the draft genome necessitating an update to the annotation to study alternative splicing. Considering the annotation quality of the draft genome sequence we felt it necessary to improve the predicted gene models using empirically supported analysis to validate, update, and add new gene models to the current genome annotation.

MATERIAL AND METHODS

Ps. cubensis growth, sample collection, and sequencing

Inoculation, growth, RNA isolation and library preparations were previously described (Savory et al., 2012). The libraries from biological replicates of cucumber leaves infected with *Ps. cubensis* for 2, 3, 4, and 8 days post inoculation (DPI) were re-sequenced using the 100mer paired-end sequencing kit on two channels, in two separate sequencing runs, on an Illumina HiSeq at the Michigan State University Research Technology Support Facility (RTSF). Library preparations were also done by the RTSF.

Alignment of *Ps. cubensis* RNA-Seq reads

Bowtie (Langmead et al., 2009) version 0.12.7 was used to align reads with up to 2 mismatches to the *C. sativus.* (Huang et al., 2009) (genome accession ACHR00000000 with contig accessions [ACHR01000001-ACHR01059995]) and *Ps. cubensis* (Savory et al., 2012) (genome accession AHJF00000000 contig accessions [AHJF01000001-AHJF01035539]) reference genome sequences. Sets of reads from the five channels that included the previously sequenced library from sporangia, were independently analyzed for quality by plotting mismatches as a factor of read position using custom Perl and R scripts. A version of SuperSplat (Bryant et al., 2010d) modified to output SAM

formatted data was used for gapped alignments, with parameters set for a gap size of 20 - 4,000 nucleotides (nt) and a minimum of at least 15 perfectly-matched nucleotides on one side of the gap. All alignments were stored in BAM format using samtools (Li et al., 2009) and converted to and from human readable alignments on the fly within the respective analysis scripts/programs. Paired-end RNA-Seq reads were aligned independently but assessed as pairs, using an inhouse C++ pipeline (Matchmaker), to identify those in which one or both reads aligned uniquely to the *Ps. cubensis* reference sequence and in a manner consistent with expectations for paired-end RNA-Seq reads. The latter included the alignment of one read per strand and an iterative process of searching pairs that aligned -60 to 200 nt apart, where a negative integer indicates overlapping alignments. These parameters were determined based on an expected fragment size of 120 – 400bp which would include the spacer sequence sizes.

Mismatch Distribution Analysis

Custom Perl and R scripts were used to generate and plot the distribution of mismatches along the length of the read within each channel. Only reads with unique ungapped alignments to the *Ps. cubensis* genome were used. For each alignment, the position in which a given mismatch occurred was enumerated and totaled for all reads at all positions. A final plot was generated that displayed the number of mismatches at each position as a percent of all mismatches found. The 5’ and 3’ ends were trimmed to the first base position having < ~2-3% of mismatches.

Spacer Sequence Length Distribution Analysis

Fragment size distribution estimates were generated using all fragments that had

both reads aligned without a gap to a unique locus in the *Ps. cubensis* reference genome and appropriately oriented for paired-end reads. For each of these unique alignment pairs, the nt position of the start of the downstream read alignment was subtracted from the nt position of the end of the upstream read, with a negative integer indicating an overlap. A file containing all of these distances was then plotted using R's 'density()' plot function for the respective channel. Fragments were binned according to the channels in which they were sequenced.

Unpaired Alignment Analysis

Paired-end alignment quality was assessed by plotting the read distribution of unpaired alignments within each channel across both sequencing runs. For all read pairs in which only one of the two reads aligned to the *Ps. cubensis* genome, the percent that aligned using the first read (R1) was compared against the percent that aligned using the second read (R2). The paired and unpaired reads that uniquely aligned to the *Ps. cubensis* reference genome were used to derive expression estimates and compared to test for similarities in genome expression. Estimates of expression were determined by calculating the number of unique alignments that aligned to *Ps. cubensis* genes using either paired or unpaired alignments. These counts were then normalized to Fragments Per Kilobase Per Million (FPKM) and log-transformed. The expression estimates derived from paired and unpaired reads for each gene were then plotted against each other. Linear regression models were estimated using the 'lm()' function and Pearson's R coefficients were calculated using the 'cor()' function in the R scripting language and plotted accordingly.

Annotation Updates and Gene Discovery

Annotation updates and gene discoveries were generated using three custom Perl/Python scripts based on coverage estimates and newly predicted splice junctions. Splice junction predictions were generated by pooling all RNA-Seq expression data that aligned to the *Ps. cubensis* genome uniquely and formatting these alignments to make them usable by the 'stacker' program in SuperSplat requiring two separate read sequences. The "update" and "gene discovery" scripts were run independently, and a final script was used to merge the final results, remove duplicates, and format the final output using the Generic Feature Format Version 3 (GFF3) format specification (http://www.sequenceontology.org/resources/gff3.html).

The update script used RNA-Seq data in three steps to improve gene annotations using the genome coverage generated from RNA-Seq alignments, as well as newly predicted splice junctions, to infer how a gene model should be updated. 1) Regions that immediately flanked an annotated genemodel were analyzed to see if the coverage warranted extending the untranslated regions (UTRs) of a gene. 2) A new transcript and open reading frame (ORF) were predicted incorporating any newly predicted splice junctions to update the gene model. 3) If the length of the UTR indicated the possibility of a new unannotated gene's existence, the UTR was examined for an ORF to check for the possibility of a new gene, and annotated as such if an ORF was found.

Gene UTRs were extended by examining the coverage of every base adjacent to the existing gene annotation to see if it could be included in the original annotation, until no new bases could be included on the 5' or 3' ends. These extensions were made if the

adjacent base met one of two criteria: the base had ≥ 10X coverage, or the base extended into a newly predicted splice junction consistent with the orientation of the gene as defined in the original annotation. The orientation of a splice junction was inferred from the donor/acceptor dinucleotide sequences of its predicted intron. The canonical 'GT-AG' and 'CT-AC' dinucleotide sequences referred to the sense and anti-sense orientations respectively, with non-canonical sequences permitted for use with either strand. Additionally, the adjacent base could not overlap a currently annotated gene.

After a gene annotation was extended, all splice junctions that were currently or newly predicted and encompassed by the newly extended UTR were incorporated. Splice junctions were only incorporated if they were consistent with the gene's orientation used to predict transcripts and subsequent ORFs encoding a protein sequence. For some genes, the addition of new splice junctions allowed for multiple isoforms to be predicted since any splice junction of an overlapping set of splice junctions could be used to predict a transcript. To address this problem, we iteratively processed all isoforms using all potential combinations of splice junctions that overlapped to predict transcript isoforms. If ≥ four isoforms could be predicted based on this iterative process, the gene was extended, the list of splice junctions were annotated, and no other updates were made. If < four isoforms could be predicted, each isoform was analyzed to examine whether or not a new ORF could be predicted based on the updated transcripts' sequence. In order for an ORF to be predicted, it must have met three criteria: a new ORF must have a predicted start and stop codon, predict a protein ≥ 30 aa or have a protein sequence longer than the original ORF predicted, and must overlap ≥ 80% of the originally predicted ORF based

on the genome coordinates of the respective ORFs. If no ORF fitting these criteria could be identified, the original gene model was kept, and no other updates were made. These criteria were implemented with the assumption that most updates would lengthen the predicted protein, rather then alter it completely.

For gene updates that predicted an UTR > 600 nt based on the location of the predicted ORF, the UTR sequence was examined to check if an additional ORF could be found along with the initial ORF that was annotated. If an ORF ≥ 30aa could be found in the UTR, a new gene was added to the annotation and the original gene UTR was truncated to 300 nt. This process was repeated for the 5' and 3' UTRs until no new genes could be found.

The gene discovery script was used to scan RNA-Seq alignments for expressed, but unannotated genome features based on exceeding a coverage of 10X. These expressed regions were joined to adjacent expressed loci if they were within 5 nt of each other or based on support by a gapped read indicative of a splice junction. These final merged sets of loci were then filtered to remove loci <= 300 nt in length or those that overlapped an annotated gene. For some genes, there existed overlapping splice junctions that allowed for the prediction of multiple isoforms since any splice junction of an overlapping set of splice junctions could be used to predict a transcript. In this case, the same process for updated genes was repeated here (see above). For genes that contained splice junctions, an ORF was predicted based on the orientation of the splice junctions inferred from the donor/acceptor dinucleotide sequences of the predicted introns following the same criteria as stated above. If splice junctions existed that were

consistent with either strand, two separate genes with the same start and stop, but different sets of splice junctions were annotated. For each isoform predicted, all isoform ORFs ≥ 30aa were annotated. If no ORF was found, an isoform was labeled as a non-coding RNA (ncRNA). If a gene had no splice junctions, the largest ORF that was found on either strand was annotated and the gene was annotated as having an orientation consistent with the predicted ORF. If no ORFs ≥ 30 aa could be found, the gene was annotated as a ncRNA with the positive strand indicated for the gene by default. IPR Scan (Zdobnov and Apweiler, 2001) version 4.8, GFAM (Sasidharan et al., 2012) version 1.1, and BLAST (Altschul et al., 1990) were used to predict protein domains, combine gene annotation for genes, and find similarly annotated genes respectively.

Gene Expression vs. Contig Length Analysis

Paired-end reads were pooled from all samples, and only those that aligned to the *Ps. cubensis* genome, which were filtered through the in-house Matchmaker pipeline, were used to count the number of fragments that aligned to each gene in the updated genome annotation. Fragments were assigned to a given gene if they aligned uniquely to a gene locus. Genes were then binned into four categories: newly annotated genes, unexpressed genes (0 raw counts), lowly expressed genes (1-9 raw counts), and expressed genes (10+ raw counts). For each gene in a given category the length of the contig it belonged to was enumerated; these values were then log transformed and plotted using the 'density()' function in R.

Analysis of Sequenced Fragments

All reads were pooled based on whether they aligned to the *C. sativus* or *Ps.*

cubensis reference genomes, and binned according to whether they aligned uniquely to one, both, or neither genome. All reads that aligned uniquely to the *Ps. cubensis* genome were then filtered through the Matchmaker pipeline, and grouped based on their alignment to an exon, intron, splice junction, or multiple categories for both updated and newly annotated genes. All reads that did not align to either the *C. sativus* or *Ps. cubensis* reference genomes were enumerated and a unique list of sequences was generated. These unique sequences were then binned according to the number of reads that represented that unique sequence. These unique sequences were then aligned using BLAST and NCBI's nt database using a representative sample of all sequences within a bin and annotated according to top hits found in the alignments. For those sequences with 1 or 2-10 reads, 10,000 random sequences were used, for those with 11-100 or 100-1,000 reads, 1,000 random sequences were used, and for those with more than 1000 reads, there were few enough sequences that they were all aligned using BLAST.

RT-PCR Validation

RNA was prepared using the Qiagen RNAeasy kit according to the manufacturers protocols from infiltrated plants using previously described protocols and was collected during a time course of Vlaspik and MSU-1 (Savory et al., 2012). cDNA was prepared in a final concentration of 10ng/ul using the oligodT primers in the USB cDNA kit according to manufacturer's protocols. The products were amplified using GoTaq. PCR conditions for amplification were a 55° C annealing temp., two minute extension, with 50 cycles. Primers for each new gene were as follows for lane two pcu_gene_340 F) CAAAGACCGCAGTCCAAGGATATTG, R) CTGGTGTGGCGGTACGAACGAAG,

lane three pcu_gene_360 F) GAAAGACCGATAGCAAGTGACTCGAAG, R) GTAGATATGGTGCAGGCATTGCATGC, lane four pcu_gene_403 F) GACCTACTGAAGAAGCTCTATCGACATG, R) GCAAATCGACCGTCAATCTGTTCTAC, lane five pcu_gene_840 F) CAGGCGACAAGAAGCGAAAGAAAGC, R) GTTGCCGTGTTGGCGTAACTTGGA, lane six pcu_gene_1584 F) CTGGAGTAAAGCATGGCGTATTAGG, R) GTACGGAAGGAAATGACAGGAGACATC, lane seven pcu_gene_2203 F) CGAAGTCGACGGGTTGGATTGAC, R) CCTCAACTCTCTCTTCTCGTGAC, lane eight pcu_gene_3095 F) GAATTCTCATTGTGTCGATATCGGC, R) CGAAGTAGCGCAGTCCTCTCG.

KEGG Pathway Analysis

Interpro domains were found for all genes in the original annotation and the expressed genes in the updated annotation using IPR SCAN (Zdobnov and Apweiler, 2001). The InterPro XML database that was downloaded with the IPR SCAN command line tool ‘iprscan’ version 4.8 was then used to extract all KEGG terms associated with the corresponding IPR domains listed in the ‘dbxref” keys found for each IPR domain entry in the XML database. Custom Perl scripts were used to extract and associate the data for each domain found in the XML database.

Intron Bearing Gene Analysis

All GTF/GFF3 files for the respective organisms were downloaded and then converted to GFF3 format with custom Perl scripts if needed. Custom Perl scripts were then used to count the number of genes and parse out how many of these genes contained

introns.

RESULTS AND DISCUSSION

Description of *Ps. cubensis* RNA-Seq datasets

Transcriptome changes at different developmental and infection stages of *Ps. cubensis* on leaves of its compatible *C. sativus* host were previously reported (Savory et al., 2012, Adhikari et al., 2012). To generate the depth necessary for identifying splicing in *Ps. cubensis*, we used paired-end sequencing on an Illumina HiSeq to re-sequence a total of ~735 million cDNA fragments (170~210 million fragments per sample) from bar-coded RNA-Seq libraries derived from two biologically replicated samples of 2-4 (early stage) and 8 (late stage) days post inoculation (DPI). The libraries were sequenced twice in separate sequencing runs, two channels each time. The un-replicated library from sporangia was not re-sequenced but the paired-end RNA-Seq reads from the > 20.5 million previously sequenced fragments were included in this study (Savory et al., 2012).

Assessing the quality of the *Ps. cubensis* RNA-Seq datasets

The RNA-Seq reads were first examined to determine the quality of the preparations and sequencing for informed decisions on implementation of appropriate filters and parameters for downstream analysis. Based on previous reports, the majority of the RNA-Seq reads were expected to correspond to the host (Savory et al., 2012). As such, the reads were first mapped to the *C. sativus* reference genome sequence with an allowance of up to two mismatches, while the sporangia-derived reads were aligned to the *Ps. cubensis* genome with the same parameters (Fig. 3.1). The majority of the mismatches clustered toward the ends of the RNA-Seq reads, consistent with known

biases in sequencing errors. Also, due to a reported machine malfunction in the first sequencing run, there was a higher incident of mismatches in the R2 reads, starting at position 76. Moreover, we observed an unusually high error rate at position 30 in the R1 reads of the second sequencing run and the overall quality of the second run, based on the frequent spikes in mismatches throughout the length of the reads, was less than optimal. To account for the clustering of assumed sequencing errors, we trimmed 5 and 31 nucleotides from the ends of RNA-Seq reads generated from the first sequencing run and 5 and 10 nucleotides from those of the second sequencing run and sporangia reads.

To assess the quality of library preparations, we calculated the distances between pairs of corresponding R1 and R2 RNA-Seq reads that aligned to the *Ps. cubensis* reference sequence. Surprisingly, there was substantial variation within and between library preparations (Fig. 3.2). The fragments for the libraries of replicate one were less variable in size within each preparation, but varied more than fourfold between library preparations. In contrast, the fragments for the libraries of biological replicate two were significantly more variable in size within samples but nearly identical between library preparations. To account for the variation in library sizes, we had to relax the parameters we set for defining confident paired-end alignment of RNA-Seq reads to allow -60 to 200 nt to account for this variation, with a negative integer indicating overlapping reads.

As another assessment of sequencing quality, we calculated the percent of trimmed R1 and R2 reads that could align to the *Ps. cubensis* reference genome but not as pairs based on the parameters defined previously (Fig. 3.3). Of the set of RNA-Seq reads from the first sequencing run, 70%-80% of the single, unpaired RNA-Seq reads that

could align, corresponded to the R1 read. In contrast, for the second sequencing run, we observed an approximate 50/50 split between the alignments of the R1 and R2 reads. These results are consistent with previous observations, supporting the conclusion that the R2 reads of run one had a disproportionately higher rate of sequencing errors. Additionally, though sequencing run two was of lower quality, the rate of errors was equivalent between the R1 and R2 reads.

Depth of sequencing affects the power in measuring differential expression of genes and detecting alternative splicing events (see Chapter IV). This is particularly relevant to this study since the RNA-Seq reads are derived from a mixed tissue sample with the majority of reads corresponding to the host. To determine whether single, unpaired RNA-Seq reads could be included in the study, we estimated and compared gene expression levels derived from counts of paired versus single, unpaired RNA-Seq reads (Fig 3.4). If the two sets of RNA-Seq reads could be aligned with similar levels of accuracy, we would expect a high correlation in the gene expression values. It was apparent that the correlation was higher for the reads from run one compared to run two (r-values ≥ 0.91 in run one compared to r-values ≥ 0.82 in run 2). Because of the higher level of correlation, and the higher quality of the R1 read in run one, we concluded that we could use these R1 reads from run one in subsequent analyses.

The quality of library preparations and sequencing may have profound effects on the data and conclusions if not properly addressed. The analyses described herein address some of the concerns in quality and, more importantly, highlight the importance of employing simple informatics tools to assess data quality to help guide downstream

analyses and understand the potential limitations on conclusions. Based on the analyses, we concluded that technical biases were introduced by independently preparing the biological replicates for *Ps. cubensis*. Additionally, it appears that there were channel and run effects, as we observed variation in the quality of the RNA-Seq datasets. To account for these biases, the RNA-Seq data sets were processed differently and the analyses hereafter only included RNA-Seq reads that aligned appropriately as pair-end reads and the single, unpaired R1 RNA-Seq reads of run one.

Improvement of the reference genome sequence of *Ps. cubensis*

Only 6.5% of the 735 million sequenced fragments aligned uniquely to the *Ps. cubensis* reference sequence (Fig. 3.5). This low percent of unique RNA-Seq reads highlights both the challenges in working with an obligate pathogen as well as the power of contemporary methods in overcoming these limitations. With the ~20.5 million fragments previously sequenced from sporangia preparations and the necessary implementation of filters to address variations in quality, only ~38 million of the uniquely aligned sequenced fragments in total were considered usable. Nonetheless, ~45% of the *Ps. cubensis* reference genome sequence was covered with an average of 218 RNA-Seq reads per sequenced nucleotide. Approximately 29.3 million of the sequenced fragments aligned to an originally annotated gene, with 13,483 of the 23,519 originally annotated genes (57%) considered expressed based on the minimal criterion of $\geq$ 10 aligned sequenced fragments from the pooled RNA-Seq datasets.

The draft genome sequence of *Ps. cubensis* is fragmented across a large number of small contigs and the number of genes exceeds the upper bounds estimated for

oomycetes (Jiang and Tyler 2012). In addition, light microscopy images show that sporangia preparations from infected *C. sativus* leaves can be contaminated with prokaryotic microorganisms and plant debris, which may contribute to the excess of contigs and mis-annotation of genes as corresponding to *Ps. cubensis* (Fig. 3.6). The RNA-Seq datasets could be used to address concerns with contamination since eukaryotic mRNA is preferentially selected for *via* the oligo-dT enrichment step thereby reducing prokaryotic mRNAs. Polyadenylation occurs with lower frequency in chloroplasts, are often heterogeneous in sequence, and tend to promote mRNA degradation (Schuster et al., 1999). As such, these RNA-Seq datasets are also expected to be biased against chloroplast-derived mRNAs from the host.

We plotted the ~27.5K (a combination of the 23.5K originally annotated genes as well as the 4K newly annotated genes as described in the next section) annotated genes as a factor of their expression and the size of the contig on which they were located (Fig. 3.7). The expressed 57% were distributed across contigs of all sizes, but were visibly biased towards the larger contigs. In contrast, the ~43% of annotated genes that were not classified as expressed, were biased to the smallest subset of contigs. Genes with no aligned RNA-Seq reads were found primarily on smaller contigs. The expressed genes that failed to meet the threshold of $\geq$ 10 RNA-Seq reads were also biased on the smaller contigs but had two shoulders. Finally, the genes that were newly annotated, as described in the next section, presented a bimodal distribution. We conclude that in general, the larger contigs are more likely to be representative of the *Ps. cubensis* genome whereas the smaller contigs have a greater probability of corresponding to contaminating

organisms due to difficulty assembling with the rest of the genome.

Introns are a genome feature that can be used to distinguish eukaryotic genes from prokaryotic genes. Indeed, analysis of the expression patterns of intron-containing genes is consistent with the conclusion that the *Ps. cubensis* draft genome sequence is derived from a mixed sample. Like many eukaryotic microbes, *Ps. cubensis* has fewer introns and intron-bearing genes than many multicellular eukaryotes, with an estimated average of only 0.7 introns per gene and only 35% of its genes annotated with at least one intron. Assuming no bias in expression, we therefore expected that ~35% of the expressed genes would encode for an intron and ~57% of the intron-bearing genes would be expressed. The observations, however, were not consistent with expectations as ~50% of the expressed genes had at least one annotated intron and a remarkable 78% of the intron-bearing genes were expressed. It is thus likely that a fair fraction of the single exonic and transcriptionally silent genes are not of *Ps. cubensis* origin.

Finally, we analyzed the 62.8 million fragments that did not align to either *C. sativus* or *Ps. cubensis* reference sequences to explore the possibility that the pathogen draft genome sequence does not sufficiently represent its inventory of genes (Fig. 3.8). More than 60% of the unaligned RNA-Seq reads were unique in sequence and represented >90% of the unaligned sequences. Another 16.5% of the RNA-Seq reads comprised ~8% of the very low abundant sequences (between 2-10 RNA-Seq reads per sequence). As such, ~75% of the unaligned RNA-Seq reads were rare, suggesting they had an excess of sequencing errors. Of the remaining sequences, most had significant homology to library adapter sequences and thus result from technical artifacts associated

with the library preparations. The other common categories we identified were plant, prokaryotic, or rRNA sequences. At best, no more than ~7% of the sequences have the potential to correspond to *Ps. cubensis* based on any level of similarity to an oomycete sequence after exceeding a BLASTN threshold score of 1×10^{-5}. We therefore concluded that most of the genes are present in the *Ps. cubensis* reference sequence.

Analysis of RNA-Seq data suggests the *Ps. cubensis* draft genome sequence likely contains contigs representing the host or other organisms inhabiting the phyllosphere. The inclusion of these contaminating sequences could explain the higher number of predicted genes relative to other genomes of oomycetes and the lower percent of expressed genes in *Ps. cubensis*. RNA-Seq analysis of *Pythium ultimum*, for example, identified 76% of the genome as expressed, albeit under eight different *in vitro* conditions (Levesque et al., 2010). In *Colletotrichum* fungi, more than 90% of their genomes are expressed during *in planta* growth (OConnell et al., 2012). The converse analysis of the unaligned RNA-Seq read sequences failed to yield significantly more *Ps. cubensis* sequences, indicating that the draft genome sequence likely includes most of the genes expressed during the conditions studied herein. Though we could correlate gene expression with contig size, there was no definitive threshold that could be safely used to eliminate contaminating contigs without risking loss of *bona fide*, but transcriptionally silent *Ps. cubensis* genes.

Refinement of *Ps. cubensis* gene annotations

The goal of this work is to characterize the extent and impact of alternative splicing in *Ps. cubensis* (see Chapter IV). It was thus essential to develop an accurate

representation of the gene models. To this end, the RNA-Seq dataset were next used to refine annotations of features in the draft sequence (Fig. 3.9). A total of 73% of the read sequences aligned to a gene that was improved. Based on the gapped alignment and/or paired-end information of RNA-Seq reads, we could confidently associate newly identified exons to 1,950 adjacent annotated genes (Fig. 3.10; Fig. 3.11). The remainder of the refinements was annotations of 5' and 3' untranslated regions (UTRs) to 10,779 genes resulting from sequenced fragments aligning to unannotated regions proximal to annotated genes.

As will be described in greater details in chapter IV, we also used perfect, but gapped alignments of the sequenced fragments to infer alternative splicing events. A subset of the gap aligned reads mapped with high confidence (exceeding threshold) within ~1.7K annotated exons (Fig. 3.10). In addition to the gapped alignment, these regions exhibited substantial drop offs in fragment coverage. Together, these observations are consistent with an intron misannotated as an exon. As a consequence, the coding sequences and/or functions for 24% of these genes were affected.

Another 13% of the sequenced fragments aligned to regions of the genome that are devoid of annotated features and could not be associated with previously annotated genes. Based on clustering of sequenced fragments that exceeded both length and coverage thresholds, we identified 3,939 new candidate gene loci (Fig. 3.11). Most (97.2%) had an identifiable ORF and were predicted to be coding. Inspection of the translated sequences suggested many were putative small peptides, as 53% were between 30 - 70 amino acids in length. Moreover, 37% were predicted to encode secretion signals

suggesting a potential role in *Ps. cubensis* virulence. Over 80% of the randomly selected candidate genes were confirmed in reverse transcriptase PCR (RT-PCR), demonstrating the efficacy of the employed methods (Fig. 3.11).

To provide a high-level perspective of the effect the improvements had on the genome annotation, we quantified the changes to KEGG pathway identifiers associated with the expressed genome versus the original annotation (Fig. 3.12). As expected, most of the identifiers exhibited a dramatic decrease in representation due to a net loss in the total number of genes; only ~10% of the identifiers were found >75% of the time in the expressed genome relative to the annotated genome. The most insightful changes can be seen at the extreme ends. The original genome had identifiers indicative of contaminating samples, including "carbon fixation pathways in prokaryotes" (found 151 times), "methane metabolism" (633), "photosynthesis", "chlorophyll metabolism" or related identifiers (>650) that were dramatically reduced in the expressed genome. Some of the genes with these assigned KEGG identifiers are likely *bona fide Ps. cubensis* genes but the dramatic decrease in numbers between the improved and original annotation is consistent with results suggesting the genome annotation was derived from a mixed sample. Similarly, the number of identifiers in several pathways associated with amino acid, nitrogen, and metabolism in the expressed genome exhibited steep declines to < 30% the number in the original annotation. In contrast, only a limited number of identifiers in pathways associated with biosynthesis of pigments, lipids, and complex carbohydrates remained unchanged.

CONCLUSION

In this chapter, we described the use of transcriptome sequencing datasets to improve the *Ps. cubensis* draft genome sequence. The original draft genome sequence was generated for the primary purpose of identifying its inventory of candidate RxLR effector genes. The genome was distributed across a large number of small contigs (Savory et al., 2012). The use of contemporary methods in transcriptome sequencing proved to adequately address many of the challenges in working with this agriculturally important obligate biotrophic pathogen and contributed to a dramatic advancement of the genome sequence, a necessary step for the work described in the following chapter.

The work described in this chapter highlights the importance of quality controlling the data. The RNA-Seq libraries were prepared and sequenced by a core facility, on two separate occasions because the R2 reads of the first run failed the core's quality metrics. Nonetheless, our analyses revealed undesirable variations in fragment sizes both within and between library preparations as well as a peculiar error in one cycle of one channel in the sequencing of run two. The former affected parameters that were used to assess the alignment of RNA-Seq read pairs whereas the latter likely reduced the number of usable reads. Awareness of such errors was instrumental for the implementation of proper *in silico* filters to balance the desire to maximize the number of usable reads while offsetting the negative effects on downstream analyses.

RNA-Seq datasets confirmed the expression of ~13.5K of the annotated genome, led to the discovery of approximately 4K new genes, and contributed to the improvement of ~9.4K genes by annotating UTRs, extending coding sequences, and correcting exon/introns. We note however, that we took a cautious approach and were reluctant in

making improvements to genes in which the evidence was deemed too ambiguous. The *Ps. cubensis* genome is dense, with short intergenic regions. This compactness often made it difficult to unambiguously assign sequenced fragments to genes.

Analysis of the RNA-Seq dataset led to the important conclusion that few, if any genes are missing from the *Ps. cubensis* genome sequence. In total, our data lead us to conclude that 17.5K represents the lower estimates of the total number of genes in the genome of *Ps. cubensis*. Undoubtedly, some of the transcriptionally silent or lowly expressed genes that failed to meet thresholds for consideration are *bona fide* genes, since transcriptomes of *Ps. cubensis* during important developmental stages, such as sexual reproduction and in oospores were not studied. However, results also suggest the original annotation may also consist of contaminating contigs that contribute to an overestimation of genes. We are reluctant to remove sequences from the draft reference sequence and will simply focus on the "expressed genome". For the work described in the following chapter, it is important to note changes to the genome statistics based on the expressed genes, particularly the number of intron-bearing genes, how *Ps. cubensis* now compares to other oomycetes and eukaryotic microbes (Fig. 3.13).

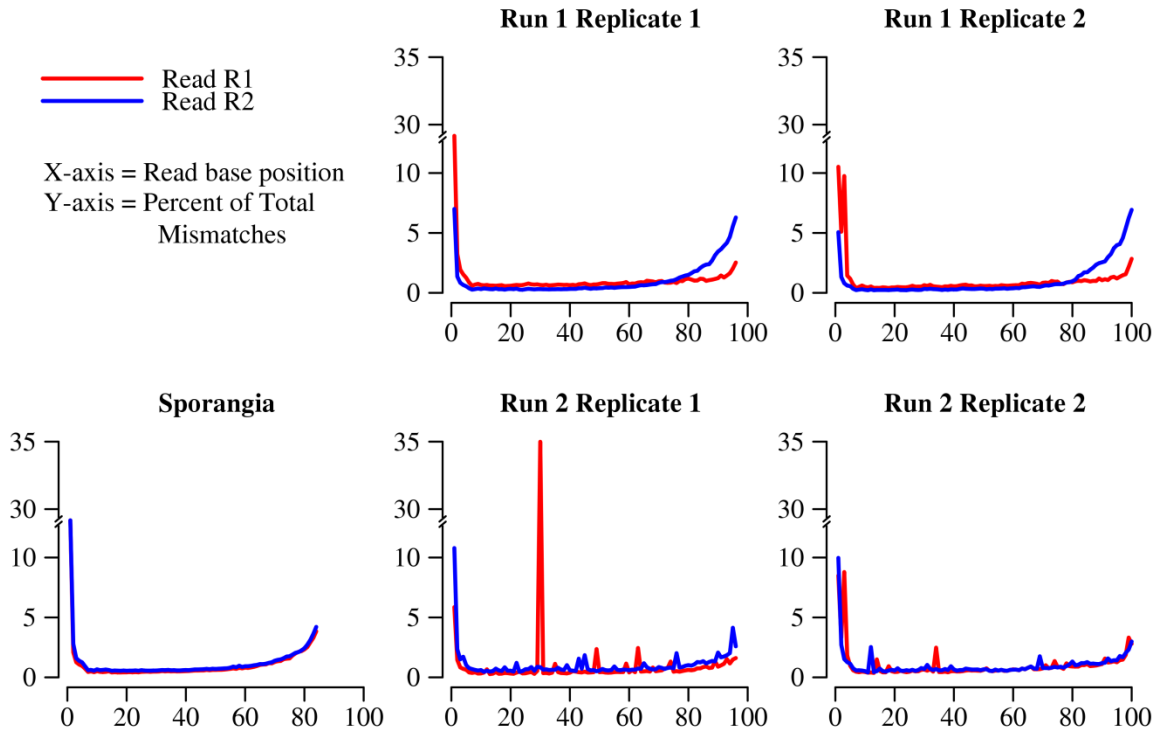


Figure 3.1. Mapping mismatches as a function of nucleotide position along the length of the RNA-Seq read.

The paired-end RNA-Seq reads from re-sequencing of libraries prepared from *C. sativus* infected with *Ps. cubensis* for 2-4 and 8 days post inoculation (dpi) were grouped based on replicate and sequencing run; sporangia was grouped separately. The read sequences were aligned to the *Ps. cubensis* reference genome sequence with an allowance of up to two mismatches. The percent of mismatches were calculated based on the total number of mismatches and plotted for each nucleotide position (R1 = red; R2 = blue).

Figure 3.2. Mapping of PE RNA-Seq reads as a function of fragment spacer size.

The paired-end RNA-Seq reads from re-sequencing of libraries prepared from *C. sativus* infected with *Ps. cubensis* for 2-4 and 8 days post inoculation (DPI) or sporangia were aligned to the *Ps. cubensis* reference genome sequence. The length of the spacer sequence between read pairs that aligned with one read per Watson and Crick strand and the correct orientation relative to each other was calculated and plotted. RNA-Seq reads were grouped according to run/sample/replicate. The distribution of these lengths were plotted with R's 'density()' function.

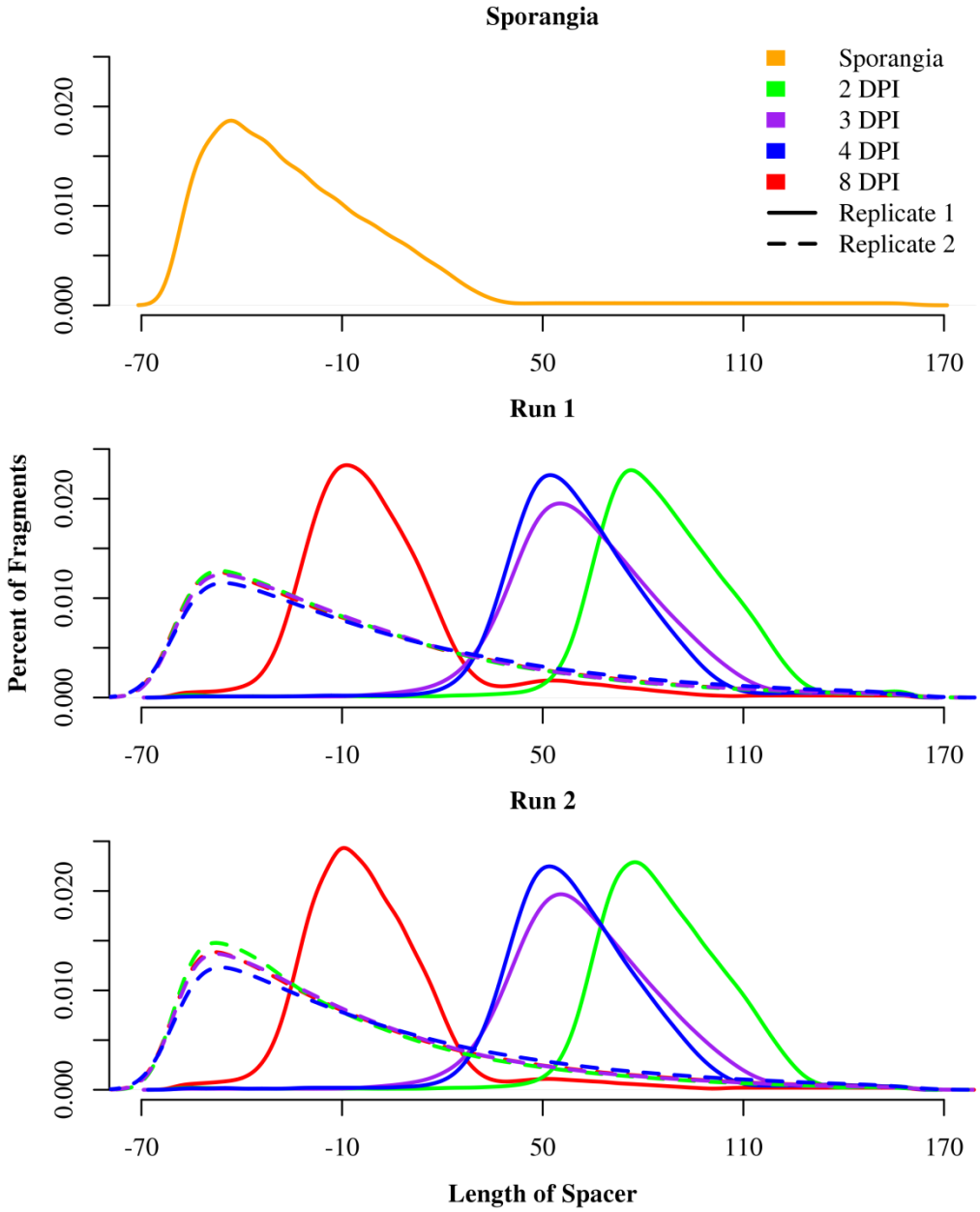


Figure 3.2. Mapping of PE RNA-Seq reads as a function of fragment spacer size.

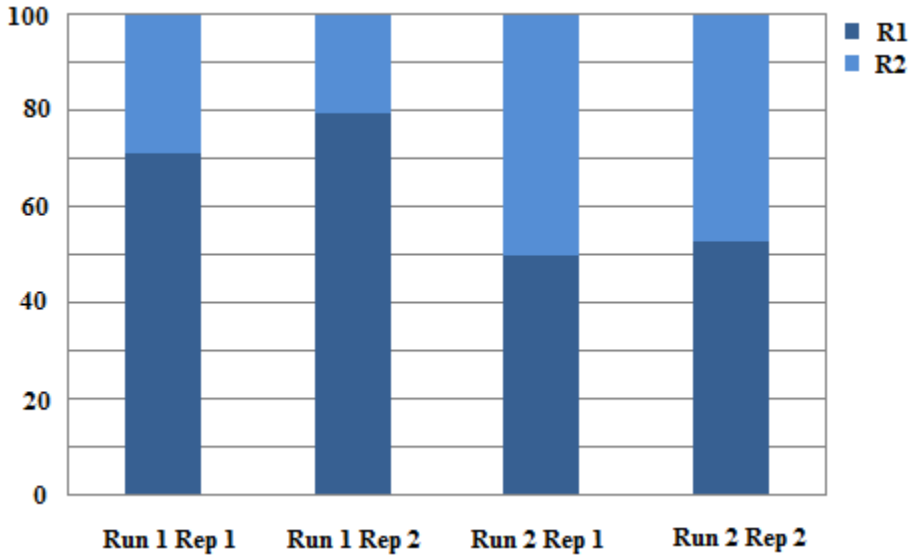


Figure 3.3. Alignment of single, unpaired RNA-Seq reads as a factor of sequencing read.

The RNA-Seq reads were aligned to the *Ps. cubensis* reference genome sequence. The percent of R1 and R2 reads that failed to meet the criteria of pairs was calculated based on the total number of single, unpaired RNA-Seq reads that aligned using only the R1 or R2 read. RNA-Seq reads were binned according to replicate and run.

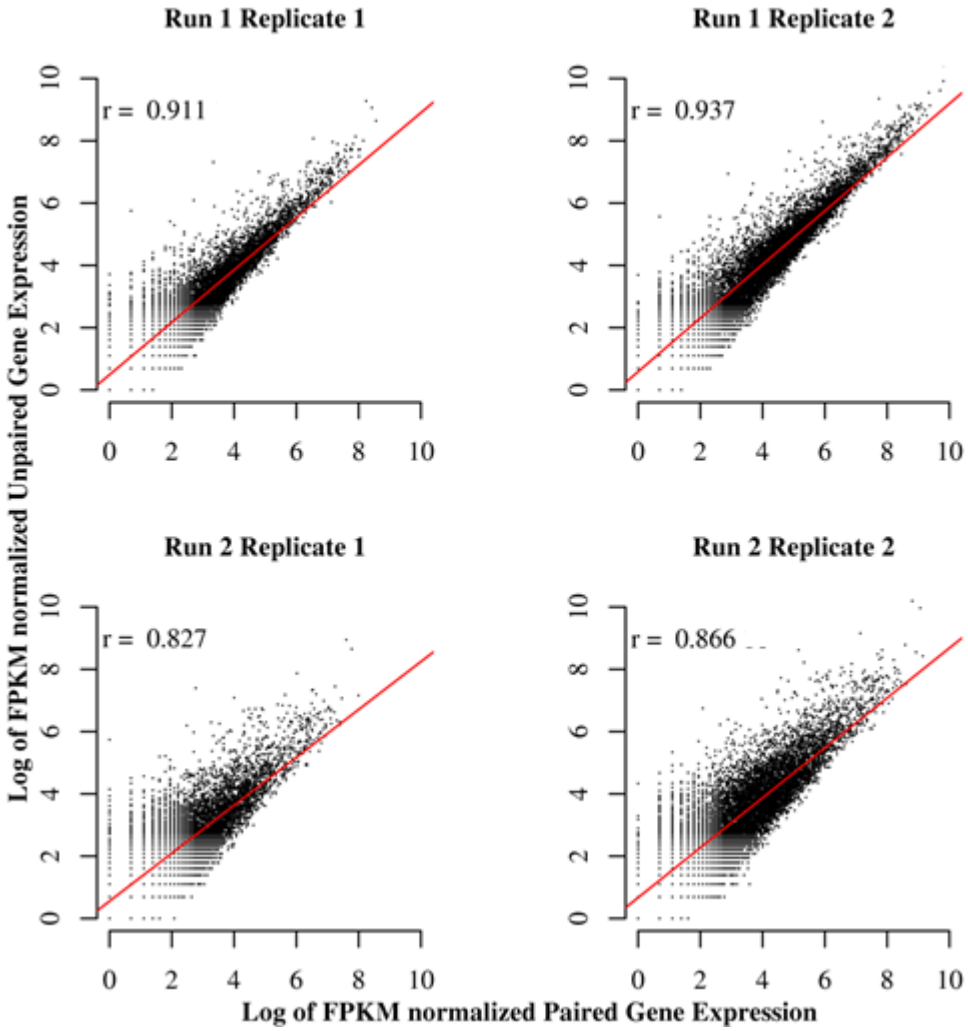


Figure 3.4. Scatter plot comparing gene counts derived from PE versus single, unpaired RNA-Seq reads.

Normalized gene counts were determined by calculating fragments per kilobase per million (FPKM) and log transformed. Counts derived from Paired-end RNA-Seq reads (x-axis) were plotted versus those derived from single, unpaired RNA-Seq reads (y-axis). The data were binned according to replicates and runs. Regression lines were plotted and Pearson's r-values are shown.

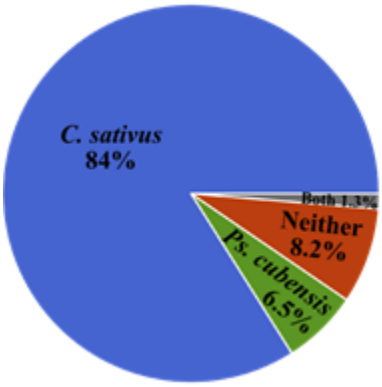


Figure 3.5. Distribution of aligned RNA-sequenced fragments.

The RNA-Seq reads of the ~735 million fragments from all re-sequenced paired-end sequenced libraries derived from pathogen infected *C. sativus* tissue were mapped *in toto* to reference sequences of cucumber and *Ps. cubensis* with up to 2 mismatches or as perfect gapped alignments. Alignments were done using Bowtie ver. 0.12.7 and SuperSplat.

Figure 3.6. Light microscopy image of purified *Ps. cubensis* sporangia.

Visualized in this microscopy image are *Pseudoperonospora cubensis* sporangia and sporangia differentiating into zoospores, showing potential for bacterial contamination. The clearish rod/round covering most of the image is bacteria, most noticeable where the sporangia are not. Scale bar = 10 um.

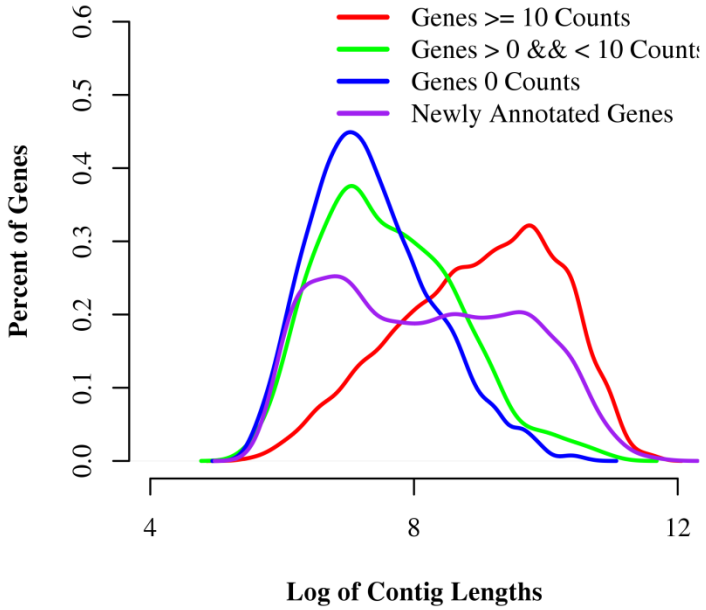


Figure 3.7. Gene expression as a factor of log of contig length.

The approximately 27,500 annotated genes of *Ps. cubensis* were categorized based on the number of RNA-Seq fragments assigned to them or if they were newly annotated genes; the size of the contig each gene belonged to was then tabulated, log transformed, and the density of these sizes was plotted. RNA-Seq fragments which aligned uniquely the *Ps. cubensis* genome and that were filtered through the Matchmaker pipeline were used to generate gene counts. Counts were calculated by enumerating the number of paired-end fragments which uniquely aligned to a given gene. (blue = no counts; green = 1-9 counts, red ≥ 10 counts, and purple = newly identified genes).

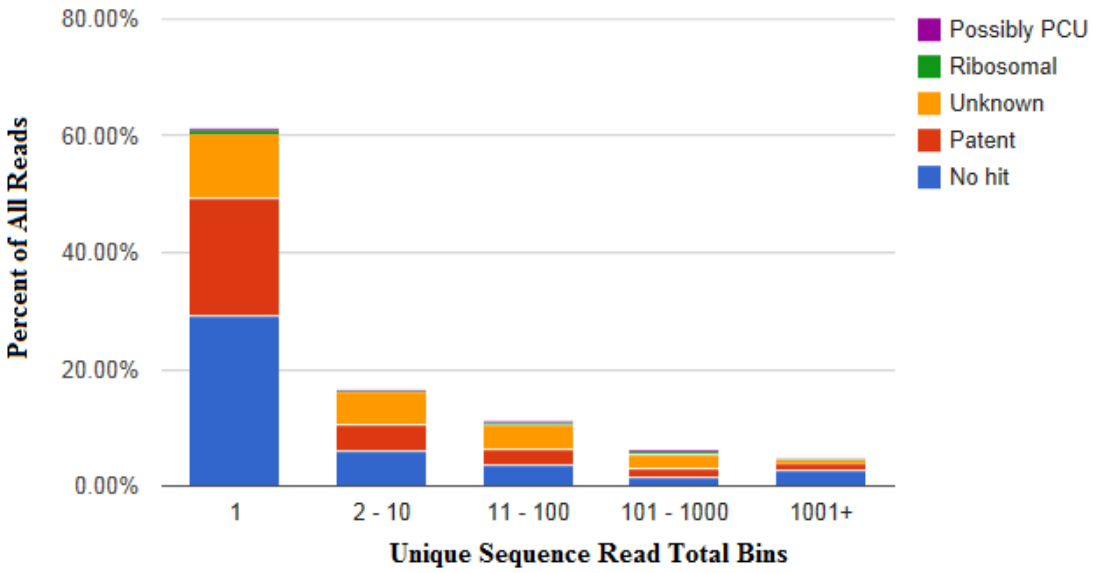


Figure 3.8. Distribution of RNA-Seq reads that did not align to either reference sequence.

Unique RNA-Seq sequences were binned based on the number of reads found for each unique sequence (1x, 2-10x, 11-100x, 101-1000x, and >1000x). The percent representation (y-axis) was calculated as a percent of all reads within a given bin. Bins were further subdivided based on the percent of each bin that was represented by each type of BLASTN homology hit based on a representative sample of sequences within each bin.

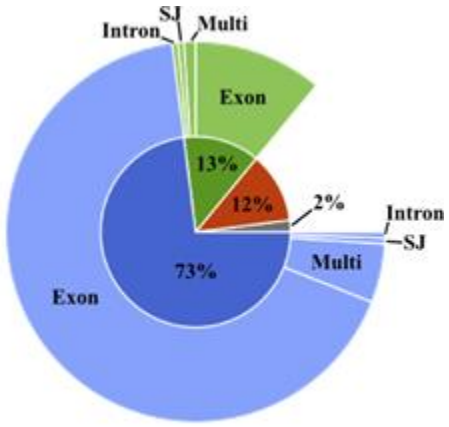


Figure 3.9. Distribution of sequenced fragments that uniquely aligned to the *Ps. cubensis* reference sequence.

The inner circle represents the percent of sequenced fragments that were used to improve a previously annotated gene (blue), identify a new gene (green), support the expression of a previously annotated gene without improving its annotation (red), and aligned spuriously to the intergenic region (gray). The outer circle represents the percent of reads that were used to improve the feature of genes: exon, intron, splice junction (SJ), or multiple features (multi).

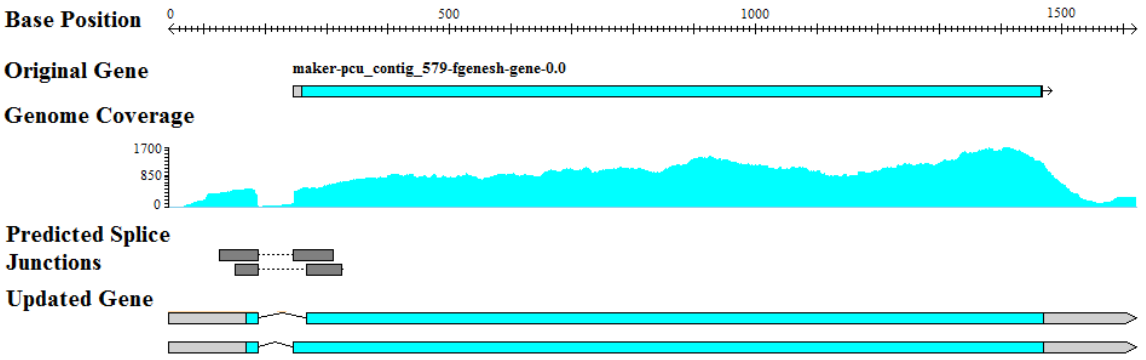


Figure 3.10. A Gbrowse screenshot of a representative improved *Ps. cubensis* gene.

A screenshot from Gbrowse of an updated gene that added a new exon, expanded the coding sequence (CDS) of the gene, and extended the gene to improve the definition of the 5' and 3' untranslated regions (UTRs) of the given gene. On the top row the bas position along the length of the gene relative to the new start is given, with nt distances indicated above. The second row indicates the original gene model (maker-pcu_contig_579-fgenesh.gene-0.0) predicted in the annotation. The third and fourth row show a histogram of the coverage produced by RNA-Seq alignments and predicted splice junctions respectively. On the bottom row are the updated gene models for this gene. Light blue bars indicate a CDS, pink regions denote the UTR, an connecting line indicates an intron, a dashed line indicates a gap in the alignment show by the purple bars, with the orientation of the gene indicated by the pointed ends of the indicated gene feature.

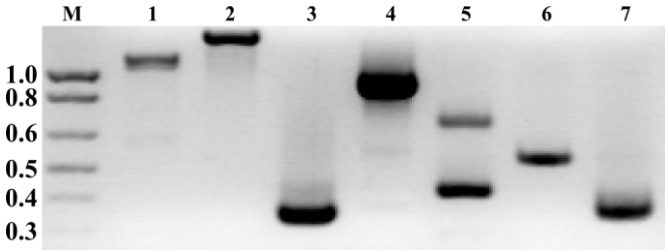


Figure 3.11. Reverse-transcriptase PCR validation of newly predicted genes.

Gene-specific primers and first-strand reverse-transcribed template from sporangia RNA were used in RT-PCRs. All bands were of the expected size. The expected size for the gene five was 420 bp. Products were resolved on a 1%, 1XTAE agarose gel. Picture is a reverse image. M = marker, 1 kb+ (Invitrogen).

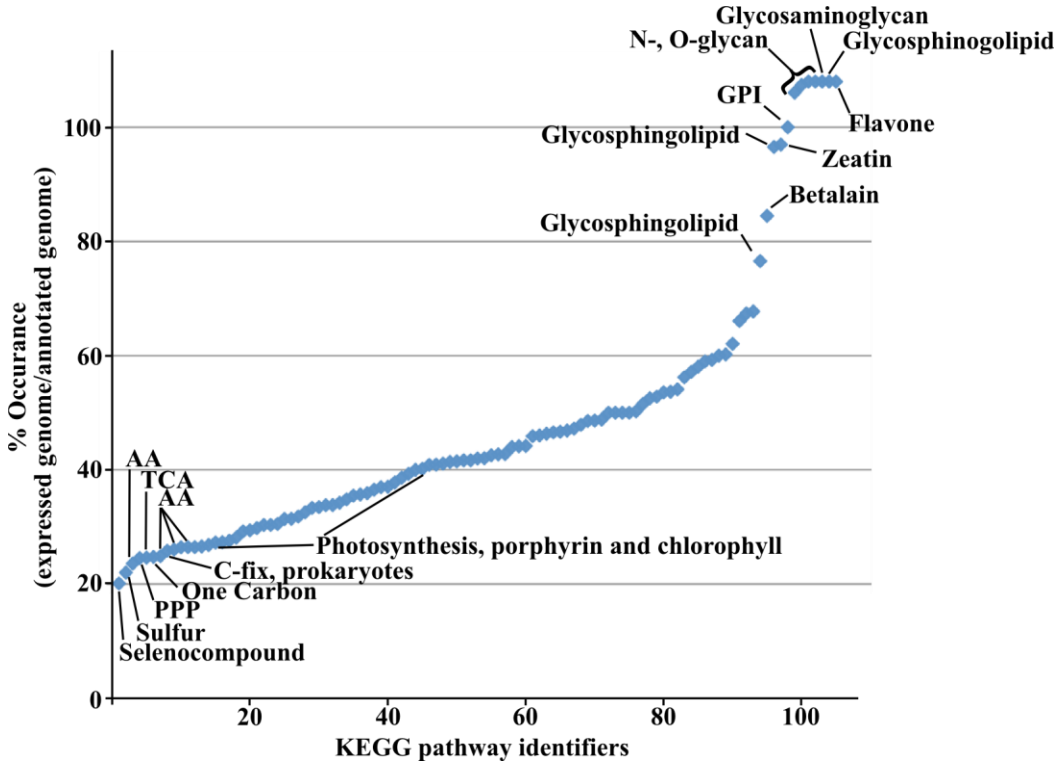


Figure 3.12. Overall reduction of KEGG pathway identifiers in the expressed genome of *Ps. cubensis*.

KEGG pathway identifiers were categorized and enumerated for the 17.5K expressed genes and plotted as a percentage based on the number found for the 23.5 annotated genes of *Ps. cubensis*. The identifiers that were found ≥ 25 times are presented and ranked from most to fewest change relative to annotated genome. Pathway identifiers at the extremes are presented. Those that exhibited a dramatic decrease in representation were most often associated with metabolism (AA = amino acid; PPP = pentose phosphate pathway; TCA = tricarboxylic acid cycle; C-fix = carbon fixation) whereas the pathway identifiers that were least affected were associated with biosynthesis (GPI = glycosylphosphatidylinisotol).

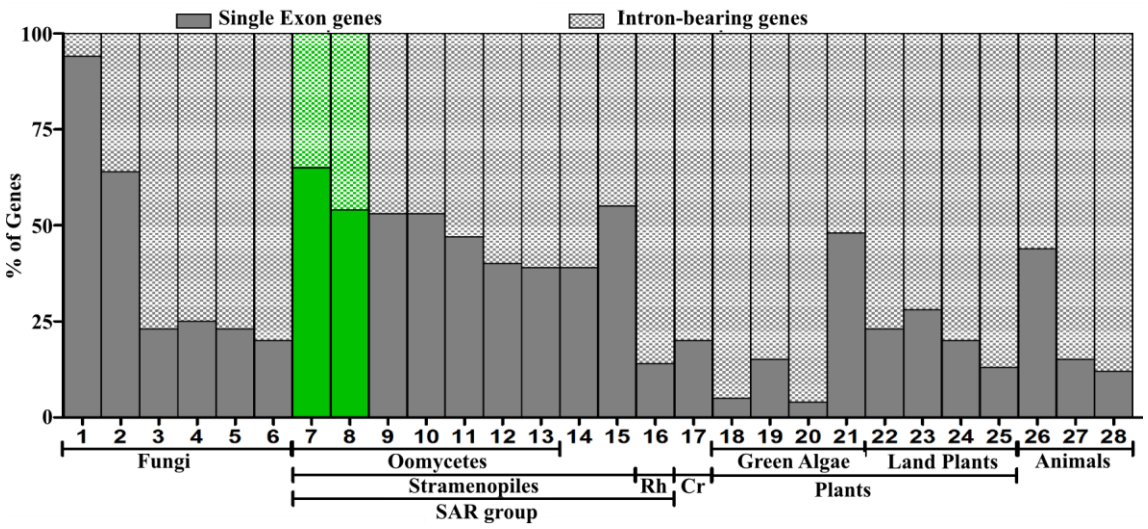


Figure 3.13. Comparison of the inventory of intron-bearing genes in *Ps. cubensis* to other eukaryotes.

The percent of single exonic genes (dark bars) and intron-bearing genes (hatched bars) from transcriptomes of representative eukaryotes were calculated and plotted. Organisms are clustered according to their phylogenetic relationship; 1= *Saccharomyces cerevisiae*, 2 = *S. pombe*, 3 = *Fusarium graminearum*, 4 = *F. oxysporum*, 5 = *F. verticillioides*, 6 = *Magnaporthe grisea*, 7 = *Ps. cubensis* (annotated), 8 = *Ps. cubensis* (expressed), 9 = *Hyaloperonospora arabidopsidis*, 10 = *P. infestans*, 11 = *P. ramorum*, 12 = *P. sojae*, 13 = *Pythium. ultimum*, 14 = *Thalassiosira pseudonana*, 15 = *Phaeodactylum tricornutum*, 16 = *Bigelowiella natans*, 17 = *Guillardia theta*, 18 = *Chlamydomonas reinhardtii*, 19 = *Volvox carteri*, 20 = *Coccomyxa subellipsoidea*, 21 = *Micromaonas pusilla*, 22 = *Phycomitrella patens*, 23 = *Arabidopsis thaliana*, 24 = *Oryza sativa*, 25 = *Cucumis sativus*, 26 = *Caenorhabditis elegans*, 27 = *Mus musculus*, 28 = *Homo sapiens*.

Alternative Splicing in *Pseudoperonospora cubensis*

Jason S. Cumbie, Alyssa Burkhardt, Elizabeth A. Savory, E. Alex Buchanan, Jeff H. Chang

In preparation for submission to The New Phytologist

INTRODUCTION

Oomycetes are biflagellate eukaryotic microbes in the lineage of stramenopiles that together with alveolates and Rhizaria, constitute the SAR super group (Hackett et al., 2007, Beakes et al., 2012; Burki et al., 2012). Members of oomycetes are notorious pathogens, responsible for some of the most devastating diseases on plants. *Pseudoperonospora cubensis* is an obligate biotrophic oomycete and the causative agent of cucurbit downy mildew, the most economically important foliar disease of cucurbits (Savory et al., 2011). Because of recent populations that can compromise natural host resistances, coupled with its ability to develop resistance to fungicides, *Ps. cubensis* threatens the long-term viability of cucumber production (Blum et al., 2011, Quesada-Ocampo et al., 2012). One of the primary ways oomycetes overcome host defense to cause disease is through the use of effectors (Schornack et al., 2009).

Oomycetes secrete these proteins via general secretory pathways, which are then subsequently taken up through ill-defined processes into host cells where the effectors are hypothesized to modulate the host immune response (Torto et al., 2003; Kamoun, 2007, Schornack et al., 2009; Savory et al., 2012). Moreover, in oomycetes, three broad classes of effectors have been identified. The two that are "true" effectors are the crinkler and RxLR effectors, named based on the phenotype the founding member caused and conserved translocation motif, respectively. The third class are the so-called apoplastic effectors that are reported to function in the apoplast (Schornack et al., 2009). In total, these effectors have many different functions such as protease inhibitors, cell wall degradation enzymes (CAzymes), and genes with less well characterized functions such

as the crinkler and RxLR families of effector molecules (Schornack et al., 2009; Savory et al., 2012). A crucial characteristic of these classes of effector proteins is an N-terminal signal peptide, indicative of their delivery via general secretory pathways (Torto et al., 2003; Kamoun, 2007; Schornack et al., 2009). Recently developed computational tools, such as Signal-P, have shown great success in predicting and scoring these putative signal peptides; Signal-P was developed using machine learning techniques to predict these signal peptides, and was used quite successfully to aid *ab initio* predictions for putatively secreted molecules (Nielsen et al., 1997; Win et al., 2007).

Splicing of pre-mRNAs is an important regulatory mechanism of eukaryotes that influences gene expression and transcriptome complexity. This fundamental process is mediated by the spliceosome, an enormous megaDalton macromolecular complex that consists of five small nuclear ribonucleoprotein particles (snRNPs) and an impressive number of more than 100 ~ 300 proteins in yeast and humans, respectively (Hoskins et al., 2012). The spliceosome assembles *de novo* in a step-wise fashion to catalyze two transesterification reactions to splice out introns and join flanking exons. Its assembly on pre-mRNAs is defined by a small number of sequences that include the 5' and 3' splice sites, the branch site, and polypyrimidine tract (Ast, 2004). Often additional short and degenerate cis-regulatory sequences are present in exons and introns that can enhance or silence splicing (Eperon et al., 1993; Zuo and Manley, 1994; McCullough and Berget, 2000; Lam and Hertel, 2002).

The mechanisms by which exons and introns are selected affect several characteristics of gene architecture and gene regulation (Keren et al., 2010). In exon

selection, the spliceosome recognizes and places the initial machinery across the exon. Intron selection, where recognition and placement occurs across introns, is regarded as the more ancient and main mechanism of unicellular eukaryotes. The difference in mechanisms imposes constraints that are consistent with observed trends in exon and intron sizes, as higher metazoans tend to have small exons and large introns whereas unicellular eukaryotes tend towards smaller introns. Plants do not conveniently group with metazoans, as the difference in average length of introns compared to exons is considerably smaller relative to humans (Reddy, 2007).

The alternative splicing of introns increases the complexity of the transcriptome to regulate gene expression and increase proteome plasticity (Keren et al., 2010; Filichkin et al., 2010). The transcriptome-wide effect of alternative splicing has been a subject of many investigations since the introduction of contemporary methods of interrogating transcriptomes. Interestingly, the prevalence and preferred type of alternative splicing tend to group according to mechanism of exon/intron selection, with plants varying yet again. In humans and Arabidopsis, approximately 95% and up to 56% respectively, of intron-containing genes are predicted to be alternatively spliced (Pan et al., 2008; Filichkin et al., 2010). The most prominent form of alternative splicing in vertebrates and invertebrates is exon skipping, a type suggested to contribute the most to proteome plasticity (Keren et al., 2010). In Arabidopsis, intron retention was the most commonly observed type of alternative splicing but the majority of events gave rise to premature termination codons (PTCs) (Filichkin et al., 2010). Alternative 5’ and 3’ splice site selection, the two other common alternative splicing types, are considered a potential

intermediate in the evolution of alternative splicing (Koren et al., 2007).

The relevance of the extraordinarily high numbers of estimated alternatively spliced transcripts has been repeatedly challenged. At the core of the debate is the fate of the alternatively spliced transcript. Many isoforms are predicted to encode nonfunctional proteins and regardless of functionality, most lack support from expressed protein variants in databases (Hegyi et al., 2011). As a consequence, sloppiness in splicing (stochastic noise) is often invoked to explain alternative splicing and the generation of large numbers of seemingly nonproductive transcripts (Melamud and Moult, 2009; Pickrell et al., 2010). Alternatively, the presence of PTCs could be an indication of gene regulation. For example, in vertebrates and plants, alternative splicing can be coupled to nonsense mediated decay (NMD) to shunt nonproductive transcripts to rapidly downregulate expression of genes (Nicholson et al., 2010; Kalyna et al., 2011; Rayson et al., 2012; Filichkin et al., 2010). The position of PTCs relative to splice junctions and the length of the 3' UTR appear to be important characteristics that influence targeting to NMD (Silva et al., 2008).

Eukaryotic microbes tend to have a lower percent of intronic genes with stronger 5' splice site sequences, and lower frequencies of alternative splicing (Irimia et al., 2007). For example, in *Saccharomyces cerevisiae* less than 5% of the genes encode an intron and alternative splicing is rare, an observation consistent with the small number of splicing factors (Meyer and Vilardell, 2009). In *S. pombe*, the fraction of intronic genes is higher, but the prevalence of alternative splicing is not (Wilhelm et al., 2008). Similarly, less than 10% of the intronic genes of *Aspergilllus oryzae* are predicted to be

alternatively spliced (Wang et al., 2010). In these microbial eukaryotes, intron retention predominates and functions as a mechanism for regulating and coordinating gene expression (Kim et al., 2008; Meyer and Vilardell, 2009).

There have been very limited studies of alternative splicing in members of the stramenopiles. Analysis of ~30K expressed sequenced tags (ESTs) from *Phytophthora sojae* revealed the potential for alternative splicing in only 122, or ~2.3%, of the expressed genes (Shen et al., 2011). All four types of alternative splicing were detected, with the majority being intron retention events. However, given limitations with analysis of ESTs and the potential for developmentally as well as environmentally influenced alternative splicing, the extent to which it occurs was likely underestimated. In fact, the four major types of alternative splicing occur in *Bigelowiella natans* at levels remarkably higher than observed in other unicellular eukaryotes (Curtis et al., 2012). This observation is particularly relevant because *B. natans* and oomycetes are more closely related as members of the SAR super-kingdom (Hackett et al., 2007; Burki et al., 2012).

To gain insights into the role of alternative splicing in oomycetes, we used deep RNA sequencing of *Ps. cubensis*-infected *Cucumis sativus* (cucumber) tissue to improve the *Ps. cubensis* genome annotation and provide a transcriptome-wide characterization of alternative splicing in an obligate biotrophic pathogen during different stages of host infection. Analyses suggest that alternative splicing in oomycetes is more similar to that of plants than other eukaryotic microbes, with the potential for alternative splicing in more than half of the intron-containing genes and a substantial portion being PTC-introducing intron retention events. The remaining events were 5’ and 3’ alternative

splice site selection with little to no evidence for exon skipping. Thus, potentially reflective of their closer evolutionary relationship, oomycetes and plants share several commonalities in alternative splicing relative to metazoans. Based on this evidence, we suggest that alternative splicing in oomycetes is likely functioning as a mechanism for both regulating gene expression and conferring proteome plasticity.

MATERIAL AND METHODS

Identification of Splice Junctions

RNA-Seq reads were aligned against the *Ps. cubensis* genome as previously described (see Chapter III), and those reads that did not align were subsequently used for gapped alignments using SuperSplat (Bryant et al., 2010). All reads that aligned using SuperSplat were then filtered using the stacker tool in SuperSplat requiring a minimum of two unique sequences with support from a minimum of one read each, to support a splice junction. These set of splice junctions were then further filtered to remove ambiguous splice junctions that had multiple alignments to different genomic loci, or if the multiple gapped alignments were produced at one locus with no unique gap supporting a canonical dinucleotide donor/acceptor sequence (GT-AG/CT-AC).

Analysis of Coverage Ratios

All RNA-Seq reads were pooled and their respective alignments were used to generate a genome-wide coverage of the *Ps. cubensis* genome. This coverage was then used to analyze all of the predicted isoforms in the updated annotation. For every intron found, the average coverage across the length of the intron was calculated. Then, the 30 nt flanking either side from the adjacent exons were examined and their average coverage

was calculated. The ratio of the intron coverage over the exon coverage was then expressed as a percentage to generate the coverage ratio. All introns with a coverage ratio that exceeded 80% of the surrounding exons were excluded from the analysis to avoid the inclusion of potentially misannotated introns. The distribution of coverage ratios was then plotted using the 'density()' function in R.

Validation using RT-PCR and qRT-PCR

The expression of each isoform was measured using isoform-specific primers (below). The expression was calculated using the delta-delta Ct method with ITS (internal transcribed spacer) region as a housekeeping/internal control gene. The total expression was calculated for each time point by adding the expression of each isoform together. The ratio of each isoform was calculated by dividing the isoform by the total. The qRT-PCR plot shows both ratios (stacked) for each time point so that the total expression for each time point should equal 100%. Total RNA was collected to make cDNA as previously described (Savory et al., 20120). A total of 25 ng of cDNA and 7.5 ul of Sybr green were used per reaction.

Forty cycles of: 95°C for 15 sec melting, 56°C for 15 sec annealing, and 72°C for 30 sec elongation was used for the quantitative real-time polymerase chain reactions (qRT-PCR). Melt curves and/or separation of cDNA samples using gel electrophoresis experiments indicated one main product.

Reverse transcriptase PCR gels were made as previously described (Chapter III). Primers used for the real-time are listed below: For maker-pcu_contig_2364-snap-gene-0.2, RT 2364 SJ1 CGTTTGCGCTGAGCGATACAC, RT 2364 Intron

GCTTGCATCGTGCGGAAGTCG, RT 2364 Rev CAGCATGCACACTACTCGATAAG. For maker-pcu_contig_04986-snap-gene-0.1, 4986_0.1_1 Intron 2 Forward AGAGCAATAGGGATGAGATGC, 4986_0.1_1 SJ2 Forward, GCTCTTTTCTAAGGTGGATGTG, RT 4986 Reverse GAAGTTCCACTTGATATCGTTGGTAGC. For pcu_gene_1703 CDS, F) CACCATGGTGAAGCTCTTCTGCGC R) GTTCTTCTCAAACAACCAGTG, for the UTR, same forward primer, R) ACCGCCAGACACATATCAAGAC.

Stage-dependent changes in coverage ratios

All introns that had evidence for intron retention were put into three separate groups: introns which were biased towards retention in sporangia, introns that were biased towards retention during host association (2-4 and 8 DPI), and introns that showed no obvious bias for either stage of infection. Bias was calculated by estimating the rate of retention for an individual splice junction in sporangia versus host associated stages, and if the ratio was at least 7x greater in one stage or the other it was set aside as biased. Coverage ratios were calculated as previously described by using RNA-Seq reads that were derived from their respective stages of infection. Once each group was enumerated, the average coverage ratio for all introns in each group was calculated for each stage of infection (sporangia, early (2-4 DPI), and late (8 DPI)), and plotted as a line graph using R.

Intron splicing efficiency

Introns with evidence for retention were grouped based on biases in sporangia or host associated states as described previously. Introns that did not belong to genes that

had evidence for differential gene expression were then removed. Raw counts for all genes with evidence for differential gene expression were then calculated for sporangia or host-associated stages combined. They were then normalized to fragments per kilobase per million (FPKM) and $\log_2$ transformed to generate the $\log_2$ fold changes for each gene relative to sporangia. Coverage ratios for each intron were calculated as previously described by pooling all reads for either sporangia or host associated stages (2-4 and 8 DPI). The ratio of sporangia to host-associated coverage ratio was then calculated to determine the change in coverage ratios relative to sporangia with a positive ratio being indicative of an increase relative to sporangia and negative value indicative of the reverse.

Analysis of RNA-Seq for differential expression

RNA-Seq read alignments were pooled for each sample to generate their respective gene counts (sporangia, 2-4 and 8 DPI) independently. Gene counts were generated by counting the number of unique alignments for each gene. These count numbers were then used to compare sporangia to both early (2-4 DPI) and late (8 DPI) stage infection using GENE-counter (see chapter II). All three statistical packages were used to test for differential expression using a q-value cutoff of 0.01 for each test. A gene was considered differentially expressed if it was identified by all three statistical tests as significant, in any pair-wise comparison of sporangia and early/late stage infection. This core set then comprised the main list of all differentially expressed genes. The counts for this gene set were then normalized to FPKMs to perform hierarchical clustering analysis to generate a heat map for all genes using the 'heatmap.2()' function in the gplots

package in the R statistical language (http://cran.r-project.org/web/packages/gplots/index.html).

Distribution of the log of coverage ratios

Coverage ratios for each intron with evidence for intron retention were calculated as previously described and pooled for sporangia, early (2-4 DPI), and late (8 DPI) stages of infection. The distribution of the log of coverage ratios for all introns during sporangia, early, or late stage was then plotting using the 'density()' function in R.

Distribution of protein lengths containing premature termination codons (PTCs)

All genes that had two predicted isoforms with one of them producing a PTC were enumerated. The length of the protein found in the PTC containing transcript and the non-PTC containing transcript was then expressed as a ratio of the former over the latter, and binned in multiples of 10, e.g., 10-20, 20-30, 30-40, etc. The total genes within each bin was calculated and plotted.

IPRScan domain changes

IPR Scan (Zdobnov and Apweiler, 2001) version 4.8 was used to predict the protein domains of all genes as previously described (see Chapter III). The list of protein domains was then analyzed using custom Perl scripts for genes with two, three, or four isoforms predicted. For each comparison, the number of domains was analyzed to test if the total number of domains changed. For those comparisons that did not change the number of domains, the type of domain was inspected to test if there were any differences in the domains listed. Domain changes were then manually inspected to interrogate which were the most commonly changed domains resulting from alternative splicing.

Defining putatively secreted proteins

Translated sequences were scanned using Signal-P version 4.0 (Nielsen et al., 1997), to predict putative signal peptides. Proteins with a signal peptide ending between 10-30 aa from the N-terminal portion of the sequence were labeled as putatively secreted peptides as was done in (Win et al., 2007). These proteins were then further analyzed using homology searches, and identifying RxLR and RxLR-like sequences as previously describe (Win et al., 2007; Savory et al., 2012), with the only modification being a relaxation of the search distance from the cleaveage site of the signal peptide (40-80 aa), to categorize the different type of effector protein molecules. These genes were then plotted on a heat map as was previously described.

Identification of EER and WY domains in RxLR effectors

To identify EER domains, 25 aa following an identified RxLR sequence were searched to define a prospective EER motif. To identify WY domains, a hidden-markov model was built using a previously defined 'WY' domain sequence found in *Phytophthora* species, and searched against our sequences using an e-value cut-off of 0.12 as previously described (Boutemy et al., 2007).

RESULTS AND DISCUSSIONS

Identifying alternative splicing events in *Ps. cubensis*

Approximately 735 million cDNA fragments (170~210 million fragments per lane) from bar-coded RNA-Seq libraries derived from two biologically replicated samples of 2-4 (early stage) and 8 (late stage) dpi were re-sequenced on an Illumina HiSeq. More than 20.5 million previously sequenced fragments from sporangia were

previously reported and included in this study (Savory et al., 2012). The fragments were processed as described (chapter III) to identify usable RNA-Seq fragments that aligned uniquely to the reference sequence of *Ps. cubensis*.

We analyzed the RNA-Seq datasets for read sequences that are consistent with a splicing or intron retention event. For the former, sequences that failed to uniquely align to the *Ps. cubensis* reference sequence with ≤ 2 mismatches were aligned as perfect, but gapped alignments to find support for splicing events that include constitutive and alternative splicing events (5' and 3' alternative splice site selection, and exon skipping). To confidently categorize the sequences as an alternative splicing event as opposed to the identification of an annotation error, we further required correspondence of the gapped alignment to a region that had empirical support for the annotated spliced sequence from at least two seperate sequences supporting the gapped alignment.

Approximately 1.9 million RNA-Seq fragments mapped as gapped, but perfect alignments to the *Ps. cubensis* reference genome sequence. A high confident list of ~20,000 putatively spliced sequences was derived by filtering out those that were not supported by multiple RNA-Seq fragments that differed in their sequences (Fig. 4.1). The majority of the predicted spliced sequences were very well supported, with approximately half supporting 65% of the junctions inferred from the original genome annotation (9,964 of the 15,369 splice junction sequences present in 6,366 expressed genes annotated with at least one intron). Additionally, not only were the newly discovered splice junction sequences well supported, but the majority had canonical intron donor and acceptor dinucleotide sequences (GT-AG or CT-AC).

To confidently identify and analyze expressed introns, we implemented three criteria based on the coverage of the 30 nt immediately flanking the exon/intron border. Firstly, in order for an intron to be considered, the flanking sequences were required to have ≥ 10X coverage as a binary measure of gene expression. Secondly, the "local intron coverage ratio" (coverage of the intron expressed as a ratio of the coverage of the flanking 30 nt) must exceed a threshold of 13%. This criterion was developed based on the assumption that most introns are constitutively spliced but potentially subject to random splicing errors, which will be revealed as stochastic noise. The average local intron coverage ratio was indeed low and the 13% threshold was set based on it being nearly 2 standard deviations to the right of the mean (Fig. 4.2). Finally, the local intron coverage ratio cannot exceed 80% to minimize the categorization of potentially misannotated exonic sequences as intron retention events.

All together, the spliced sequences and expressed introns, that we will refer to as retained introns, contribute to an estimated ~10.5K alternative splicing events in ~5.2K genes. Therefore, 29.5% of the expressed genome and 57.9% of the intron-containing genes of *Ps. cubensis* are potentially alternatively spliced. Similar to land plants, the predominant type of alternative splicing is intron retention (Figure 4.3A). Interestingly, unlike mammals (REF) and the more closely related algae (REF), we found a lack of exon skipping. To ensure that our filters were not overly stringent, we checked for exon skipping in a more rigorous approach utilizing all of the potential splice junctions predicted by our dataset without any filtering. To do this, we found all overlapping splice junctions with the same donor/acceptor sequences (excluding non-canonical sequences),

and iterated through all combinations to find examples of exon skipping. Only two examples were identified: one overlapping maker-pcu_contig_397-snap-gene-0.10, and the other overlapping gene fgenesh_masked-pcu_contig_790-abinit-gene-0.17. In both examples, at least one of the overlapping splice junctions failed to meet criteria of SuperSplat, and in the former example the exon skipping event occurred in a gene that had too many splice junctions to easily resolve the different transcript isoforms predicted. This could imply two potential conclusions: either exon skipping does not appear to occur in *Ps. cubensis* or our data did not have sufficient coverage for interrogating the potential for exon skipping. Considering that even with deep sequencing many of our reads aligned to the host genome and the fact that we did not have samples for all potential conditions and developmental stages, we felt it would be best to err on the side of caution and conclude that our data could not sufficiently interrogate exon skipping or that it was exceptionally rare. Because of this, exon skipping is not further examined in this study. Examples supporting the different types of alternative splicing events are provided (Fig. 4.3B). Furthermore, to validate our predictions, we used qRT-PCR and isoform specific primers to quantify expression of a few spliced isoforms (Fig. 4.4).

Not all of the alternative splicing events necessarily contribute to a functional protein isoform. Rather, the inventory provides an itemization of putative alternatively spliced genes that is useful for generating hypotheses that need to be experimentally validated. In fact, the sum total alternative splicing events likely consist of biologically relevant events that enhance the proteome and reflect forms of gene regulation, as well as stochastic noise and incompletely processed pre-mRNAs. In the following, we attempt to

tease apart the contributions of noise, pre-mRNAs, and biologically relevant events, to the conclusions on alternative splicing in *Ps. cubensis*.

It is assumed that high levels of expression are correlated with essential functions and consequently, more accurate splicing. Low abundant transcripts, in contrast, are less likely to be toxic if mis-spliced and noise is thus often correlated to lowly expressed genes in both mammals and plants (Pickrell et al., 2010; Jiao and Meyerowitz , 2010). In humans, for example, the noise inferred based on unannotated spliced sequences was poorly supported, represented by only 1.7% of the gap-aligned RNA-Seq read sequences. For *Ps. cubensis*, a number of filters were implemented to limit poorly supported alternative splicing events and reduce the misclassification of noise as relevant alternative splicing events (Fig. 4.1). In fact, the unannotated spliced sequences were well supported by ~30-40% of the gap-aligned RNA-Seq read sequences.

We examined the trend in expression for the introns. Those that passed thresholds were categorized based on sharing similar patterns in normalized intron coverage ratios across the three different stages (Fig. 4.5). As shown, approximately 1.3% and 21.6% of the introns showed a dramatic increase or decrease in retention during growth of *Ps. cubensis* on host tissue, as compared to sporangia, respectively. The vast majority, of retained introns (77.1%) had little to no significant change in retention over the course of infection. In regards to the former two categories, the average stage-dependent changes from sporangia to its host-associated stages were substantial in difference. We also examined the normalized intron coverage ratio as a factor of stage-dependent differential expression relative to sporangia, expressed as $\log_2$ fold change, and found no obvious

trends (Fig. 4.6). The normalized intron coverage ratio is effectively a measure of splicing efficiency since it is calculated relative to the coverage of the 30 nt immediately flanking the exon/intron junction. Thus, splicing efficiency does not appear to correlate with transcriptional activity for the conditions examined in *Ps. cubensis*.

Other than stochastic noise, which is addressed in an earlier section, the behavior of introns with stage-dependent changes can be explained by one of several regulatory mechanisms. For the genes that showed an increase in intron retention, alternative splicing to generate premature termination codons can be coupled to a quality control process called nonsense mediated decay (NMD) as a post-transcriptional process to down regulate transcript levels (Nagy and Maquat., 1998; Maquat, 2004; Hori and Watanabe, 2007). This is indeed a very likely possibility for many genes as a substantial portion of intron retention events (~35%+) introduce a PTC.

Alternatively, intron retention could lead to functional protein variants, as is the case for *Psc_RXLR1* (Savory et al., 2012). For the genes that showed a decrease in intron retention, there was no global coupling of transcription and splicing, as was described for *S. pombe*, but by no means does this exclude the potential for a coupling of alternative splicing and gene regulation (Wilhelm et al., 2008). It has also been suggested that sporangia, like oocytes in metazoans, synthesize, store, and mask mRNAs (Walker et al., 2008; Vasudevan et al., 2006; Richter and Lasko, 2011). Upon perception of a signal, the mRNAs are "unmasked" for rapid expression to undergo developmental changes. In *Xenopus* for example, it has been suggested that the inactivation of a *bona fide* DEAD-box helicase unmasks transcripts followed by polyadenylation that together promote

translation (Minshall et al., 2001). In *P. infestans*, a highly induced putative DEAD-box RNA-helicase encoding gene was discovered that when silenced, led to morphological defects in zoospore, thereby linking the unmasking of mRNA in oomycete sporangia (Walker et al., 2008).

There is however, little support for the unmasking model in oomycetes and even if relevant, is not likely to have consequence to conclusions on alternative splicing. Firstly, there has been no characterization of helicase or RNA binding activity or demonstration that RNA is indeed masked in sporangia. Furthermore, in order for unmasking to explain intron retention, it necessarily implies that unprocessed pre-mRNAs are stored in sporangia. However, the majority of splicing events in eukaryotes occur in concert with transcription (Han et al., 2011). Moreover, the presence of an intron in *in vitro* transcribed mRNAs did not influence the targeting of transcripts to the masking pathway when microinjected into *Xenopus* oocyetes and in fact, the majority of introns were spliced (Meric et al., 1996). Thus, in general, the possibility that pre-mRNAs are stored and masked in sporangia of oomycetes is unlikely. Finally, in our experiments, the mRNA was purified using oligo-dT enrichment from sporangia triggered for germination. Polyadenylation is a critical regulatory step in the unmasking of mRNAs, masked mRNAs are either deadenylated or have very short polyA tails that are elongated upon unmasking in oocytes (Richter and Lasko, 2011). Though, in the absence of experimental data, we cannot exclude the possibility that pre-mRNAs are masked, there are several inconsistencies that compel us to suggest that masking of pre-mRNAs is an unlikely explanation to generalize the observed high number of expressed

introns in sporangia.

Differential expression

To provide support for the biological relevance of alternative splicing in the development and virulence of *Ps. cubensis*, we identified genes that exhibited stage-dependent changes in alternative splicing and differential expression. Estimated gene expression values were calculated by counting the number of raw sequence fragments which aligned uniquely to both updated and newly annotated genes for all samples sequenced. Relative to sporangia, approximately 3.7K genes were identified as differentially expressed, with 1.1K expressed to higher levels (q-value threshold = 0.01; Fig. 4.7). Of the ~5.2K alternatively spliced genes, 968 were differentially expressed (177 up and 791 down relative to sporangia). As was found previously, there were a substantial number of genes whose expression patterns were highly correlated to stage dependent changes (Savory et al., 2012). More importantly, however, the fact that a number of these genes have evidence for alternative splicing implicates a substantial role for alternative splicing in their regulation. Since over 177 genes were much more likely to contain an intron compared to their host-associated stages, and additionally were shown to be upregulated it would seem possible that NMD could be important to their regulation. However, many more genes showed a decrease in gene expression following the reduction of intron retention, again consistent with the argument that NMD would not be a sufficient explanation for the regulation of many alternatively spliced transcripts.

We next plotted the change in intron coverage ratios for those genes identified as differentially expressed and alternatively spliced. For those genes identified, there was a

clear bias towards intron retention events in sporangia compared to early and late stages of infection (Fig. 4.8). A total of 458 genes with evidence for alternative splicing and differential expression, and their respective introns, were analyzed with 78 genes having higher levels of expression and 380 having lower levels of expression relative to sporangia. Again, as was established before, intron retention is highly biased towards sporangia. Considering that many of these genes seem to be down regulated and hence are not likely to be regulated by NMD, it may be that a change in function, and not a simple change in the level of expression, is important as a general mechanism for gene regulation via alternative splicing as was the case with *Psc_RXLR1* (Savory et al., 2012).

In addition to examining intron retention events, we also analyzed predicted splice junctions for stage-dependent expression. Ideally, statistical models would be used to test for differential expression analysis to evaluate stage-dependent changes in the use of specific splice junctions. The fact that individual samples had a low abundance of splice junctions sequenced, coupled with the fact that only one biological replicate was sequenced for sporangia, made us reluctant to test individual splice junctions for differential expression. However, a binary present/absent analysis could at the very least illustrate the potential for stage-dependent expression for the predicted splice junctions, albeit a less sensitive and statistically rigorous approach. To that end, all the predicted splice junctions in the final genome annotation (see Chapter III) were analyzed and examined for their presence/absence in sporangia vs. early or late stage infection. A total of ~4k genes were found that had at least one splice junction that met this criterion. Additionally, approximately 723 of the genes out of the ~4K genes represented had

evidence for differential expression. To rule out the potential for low abundance transcripts accounting for these differences as a result of low sampling, we filtered out genes that had an FPKM normalized value < 10 in all stages of infection ensuring expression across all time points analyzed. Applying this filter left 190 genes with evidence for differential expression that met our binary cutoff and were expressed to sufficient levels to rule out lowly expressed genes with insufficient splice junction sampling. Considering that a substantial number of genes with this form of alternative splicing were left even after these strict filtering criteria would appear to, again, implicate that it is likely a change in function that plays an important regulatory role of alternative splicing in *Ps. cubensis*.

To continue to address the biological relevance of these genes we also examined some functional characteristics of differentially expressed and putatively alternatively spliced genes. Using the most recent version of Signal-P, ver. 4.0, we analyzed the newly annotated genome for predicted signal peptides. Using the *ab initio* criteria previously developed (Win et al., 2007), we were able to predict ~1K putatively secreted proteins. Of this list, 249 had evidence for alternative splicing, 250 had evidence for differential gene expression, and 59 had evidence for both. This would imply that alternative splicing has a substantial impact on secreted proteins, and considering the fact that differential expression was analyzed using very strict criteria, this probably represents a lower bound for genes that are regulated via their alternatively spliced transcripts. Additionally, homology and protein domain searches with the differentially expressed and alternatively spliced genes highlighted some with functional similarities to protease

inhibitors (2), CAzymes (3), crinkler (1), and RxLR effector-like (11) domains again implicating that these alternatively spliced genes are likely important players in pathogenesis. The majority of these genes (42) could not be easily categorized into functional groups based on homology searches or domain scans. While it must be stressed that stochastic noise and mis-annotations could potentially account for the differences noticed, the fact that such conservative criteria were used in defining alternatively spliced and differentially expressed genes would suggest that there are indeed a number of *bona fide* instances of stage-dependent alternative splicing.

Development-associated alternative splicing

To identify examples of regulated alternative splicing as a counter to stochastic noise, we correlated changes in alternative splicing to each of the developmental stages we investigated for *Ps. cubensis*, early and late infection and sporangia. Additionally, we wanted to establish the similarities in overall expression patterns compared to previous studies and different stages of infection (Savory et al., 2012). To do that, we looked at all genes called differentially expressed in early and late stages of infection, and compared them to the gene 'modules' previously described (Savory et al., 2012) which were correlated with different stages of infection.

These gene modules were generated based on similarities in gene expression patterns over the course of infection, comprising 6 different gene modules. Modules 2-4 correlate best to our early stage definition (2-4 DPI), and modules 5-6 went best with our late stage definition (8 DPI). We excluded module 1 from the analysis since we did not sequence 1 DPI in our data set. When looking at genes called differentially expressed,

we found that 20% of genes in modules 2-4 were found in early stage differentially expressed genes, and that 48% of modules 5-6 were found in our late stage differentially expressed genes.

Since the gene modules were made by comparing similarities in gene expression at different stages of infection, as opposed to direct differential expression analysis, it was assumed that *bona fide* differentially expressed genes would be included along with genes that had similar but different degrees of change and thus would likely include genes not called differentially expressed. To test if this was the case, we did the same analysis above, but used a q-value cutoff of 0.05 and called a gene differentially expressed if it was found by any of the three methods implemented in GENE-counter to find all possible differentially expressed genes. With these more relaxed settings, we found that 53% of genes in modules 2-4 were found in early stage differentially expressed genes, and that 83% of modules 5-6 were found in our late stage differentially expressed genes, consistent with this hypothesis. We thus concluded that these results were similar to earlier studies and that we did indeed capture significant stage-dependent changes in gene expression.

To further look into stage dependent alternative splicing, we looked at those genes which had previously been estimated to have evidence for changes in intron retention events or the use of an alternative 5' or 3' splice site and compared them to the modules previously described. For modules 2-4, we found 22% of genes with evidence for stage-dependent alternative splicing, and 19% of genes in modules 5-6. Additionally, these two groups comprised about 20% of all genes with evidence for stage dependent alternative

splicing (~3K genes). This would indicate that alternative splicing may indeed play a significant role in regulating the different stages of infection.

Potential changes to the *Ps. cubensis* proteome

In silico studies characterizing alternative splicing and protein functions have revealed some tendencies that can be used to help distinguish plausible biologically relevant events from noise. Some general trends include the following: 1) protein structures derived from different isoforms tend to be similar (Hegyi et al., 2011), 2) truncations tend to occur between and not within functional domains (Kriventseva et al., 2003), and 3) active sites tend to be present in protein isoforms (Leoni et al., 2011).

We determined the potential effects that alternative splicing has on the *Ps. cubensis* proteome. The translated sequences of the alternatively spliced coding genes were scanned for changes in putative functional domains as well as extra-cellular and sub-cellular localization sequences. Approximately 2.1K of the alternatively spliced genes encode a predicted domain. Focusing on the 1.4K genes with no more than two predicted gene models, 34% had evidence for a PTC. Plotting their positions relative to the full-length sequence revealed a bimodal distribution distributed around 25% and 80% of the length of the presumed functional sequence (Fig 4.9) showing that there was a wide range in protein lengths that occurred. Regardless of the length of the truncated proteins, 79% had a predicted premature termination codon between putative functional domains, consistent with the potential for being a biologically relevant alternative splicing event. This would again be consistent with and support a regulatory role that emphasizes the change in protein function, rather than expression level, that alternative

splicing plays in *Ps. cubensis*.

The remaining 66% of the alternatively spliced genes are not predicted to give rise to a premature termination codon. Of these, 19% are predicted to vary in the number of functional or targeting domains with a substantial number affected in putative signal peptides, zinc finger domains, and transmembrane domains that can be associated with virulence proteins, DNA-binding proteins, and membrane-localized proteins, respectively (Fig. 4.10)

Members of the often antagonistically acting heterogeneous nuclear ribonucleoprotein (hnRNP) and serine/arginine-rich (SR) protein families contribute to many aspects of gene expression, including constitutive and regulated alternative splicing (Long et al., 2009). Members of both families exhibit a modular structure, with most containing an RNA-binding domain, e.g., RNA recognition motif (RRM), that underscores their functional flexibility (Han et al., 2010). The hnRNPs and SR-encoding genes are themselves subject to alternative splicing (Richardson et al., 2011; Han et al., 2010; Reddy et al., 2004). In Arabidopsis for example, 18 serine/arginine-rich (SR) genes give rise to an estimated 93 different isoforms in response to various developmental cues and stresses, consistent with the view that they coordinate signal and stress perception and regulation of splicing (Palusa and Reddy., 2010). *Ps. cubensis* expresses at least two hnRNPs and five SR genes, with one and three members, respectively, having evidence for alternative splicing (Fig. 4.11).

We also focused on functional categories of genes that are potentially involved in virulence or regulation. These included genes encoding putative transcription factors,

non-coding RNAs, as well as secreted proteins. To highlight their overall prevalence, we generated a heatmap of putatively secreted proteins that were also differentially expressed. As shown in the figure, a substantial portion are alternatively spliced, with many having functions related to virulence (Fig. 4.12). The fact that many of the secreted proteins also had functions indicative of pathogenesis that were alternatively spliced would, again, hammer home the point that the impact on protein cannot be underestimated when looking at alternative splicing in *Ps. cubensis*.

These data show that alternative splicing can influence the functions for many genes belonging to categories associated with development and virulence of *Ps. cubensis*. However, the suggestions that splicing is a stochastic process and some, to many, alternatively spliced transcripts are merely products of "noise", challenge the biological relevance of their functions and dampens estimates of impact on transcriptome plasticity (Melamud et al., 2009). Considering that a substantial portion of alternatively spliced transcripts have evidence for differential expression, are annotated with putative secretion signals, or both would argue against the hypothesis that these events are merely the product of "noise", and indeed are important regulators of pathogenesis. In the latter case, the splice junctions are wholly contained within exonic sequences and presumably correspond to introns. In all, the RNA-Seq reads were used to correct the annotations for 9,456 genes, with the majority of these adding a UTR to the gene (5820).

RxLR and Crinkler effectors show evidence of alternative splicing

To investigate more closely the impacts of alternative splicing, we analyzed the two most well characterized groups of effectors in oomycetes: the RxLR and crinkler

effectors. A total of 114 genes were found that encoded an RxLR motif downstream of a putative N-terminal signal peptide (Win et al., 2007). Our estimates nearly doubled the total number of candidate RxLR effectors starting with an 'R' at the R1 aa position (Tian et al., 2011; Savory et al., 2012). Moreover, a total of 79 of the candidate RxLR encoding genes were also identified as expressed in the RNA-Seq data set (Table 4.1) Of these expressed genes, 12 had evidence for alternative splicing. Two well known domains described for RxLR proteins include an N-terminal EER domain found within 25 nt of the RxLR motif, and a C-terminal WY fold domain (Whisson et al., 2007; Win et al., 2012). After analyzing our RxLR protein sequences for these domains, we found that only 15 had an identifiable EER motif, 12 had evidence for a WY domain, and two had evidence for both. Interestingly, only one of the alternatively spliced RxLR effectors had evidence for a WY domain. This was the case as well for *Psc_RXLR1*, the previously identified alternatively spliced RxLR effector, which contains neither the EER domain or the WY domain, but loses a transmembrane domain as a result of alternative splicing (Savory et al., 2012). It was not immediately clear from our data that this type of domain loss was the same with our RxLR proteins, since none of the proteins showed similar changes in the protein domains identified. The role for alternatively splicing in these RxLR candidates remains to be seen.

We also searched the updated gene list for candidate crinkler effector encoding genes. A homology-based search, using the N-terminal regions of candidate crinkler peptides, was employed since the N-terminal portion has been reported to be the most conserved part of crinkler peptides (Torto et al., 2003; Haas et al., 2009). Two new

genes, pcu_gene_1703 and pcu_gene_2449, were found which we have labeled PsCRN1 and PsCRN2, respectively, based on their homolgy to the *P. infestans* crinklers CRN1 and CRN2 respectively (% identity >= 70% in the N-terrminal region using BLAST) (Torto et al., 2003; Haas et al., 2009). This brings the total of crinkler genes to 140 adding to those previously identified (Savory et al., 2012), of which 83 were identified as expressed in our RNA-Seq data set including the newly identified genes (Table 4.2). Of this expressed set, eight had evidence of alternative splicing, however only 3 of these had identifiable signal peptides, two of which had evidence for alternative splicing. This is far fewer than the 196 crinkler proteins found in *Phtyphthora infestans* of which 60% were shown to possess a signal peptide (Haas et al., 2009), but much closer to 19 and 8 crinkler genes in *Hyaloperonospora arabidopsidis* and *Phytophthora ramorum* respectively (Baxter et al., 2010). The latter two comparisons are more relevant considering H. *arabidopsis* and *P. ramorum* are obligate biotrophic pathogens, like *Ps. cubensis*. One of these two crinklers, PsCRN2, interestingly had evidence for a retained intron in the 3' UTR of the gene. We were able to confirm the existence of this intron retention event (Fig. 4.13), and showed that there was a distinct shift in isoform abundance over our time course via RT-PCR.

CONCLUSION

In this chapter, we inventoried the *Ps. cubensis* transcriptome for potential alternative splicing events. Additionally we identified a number of new potential alternatively spliced RxLR effectors, and were able to confirm the alternative splicing of a crinkler protein with similarities to the previously identified CRN2 crinkler in *P.*

infestans. This work is important for two main reasons. Our efforts to identify alternatively spliced genes represents an important first step in understanding the potential impact that alternative splicing has on transcriptome plasticity of oomycetes and its influence on the adaptability of this important group of pathogens to a wide range of hosts and environmental conditions. The genes we have identified represent a resource for hypothesis generation and functional testing. Secondly, models of alternative splicing are primarily based on mammals, model fungi, and to a lesser extent, higher plants. In mammals, there is a bias for exon skipping and the tendency for genes to have short exons but long introns. In model fungi, introns may be uncommon and alternative splicing is rare, almost to a point of being dismissed. Results from this study revealed clear differences in oomycetes, which are members of a eukaryotic kingdom poorly studied in regards to alternative splicing. Both exons and introns of *Ps. cubensis* are short in length with intron retention the preferred type of alternative splicing. These data also revealed a greater than an order of magnitude increase in the prevalence of alternative splicing relative to the previous reports of 2.3% for *P. sojae*. In hindsight, given that oomycetes are more closely related to plants, the commonalities with plants are not surprising. But relative to possible misconceptions based on fungi, results challenge the thought that alternative splicing is of little consequence in eukaryotic microbes. As such, results from this work will contribute new insights into the evolution and function of alternative splicing.

In contrast to over representing the amount of alternative splicing, we suspect that we have in fact, provided an under estimate. Alternative splicing is often

developmentally regulated or stress-induced and the three stages we investigated are not likely sufficient to represent all conditions that associate with alternative splicing. Additionally, it is clear that because the RNA-Seq libraries were from mixed samples, dominated by the host, the depth of sequencing for *Ps. cubensis* was low and less likely to capture low abundant transcripts, e.g., those from splicing errors and targeted to quality control mechanisms. There were several aspects to the manner in which we analyzed the data that also contribute to under estimating alternative splicing. The most significant limitation was the pooling of the RNA-Seq datasets to account for the fact that *Ps. cubensis* represents a miniscule fraction of the biomass and hence sequenced transcript fragments, particularly during early infection stages. Pooling will artificially depress intron coverage ratios, and in combination with the threshold-based filters that we implemented to address the absence of robust statistical methods for identifying true intron retention events, contribute to a high false negative rate. This is further amplified by the observation that intron retentions are the most common type of alternative splicing in *Ps. cubensis*.

Admittedly, the relative contributions of the biologically relevant events and noise to the total number of events categorized as alternative splicing is difficult to determine. However, a number of important findings stress the potential for the biological relevance of alternative splicing. A number of alternatively spliced genes had putative signal peptides, and many had annotations consistent with pathogenesis. Many genes that were alternatively spliced also had evidence for stage-dependent differential gene expression. These findings indicate that alternative splicing is indeed implicated as a substantial

contributor to gene regulation in the obligate plant pathogen *Ps. cubensis*.

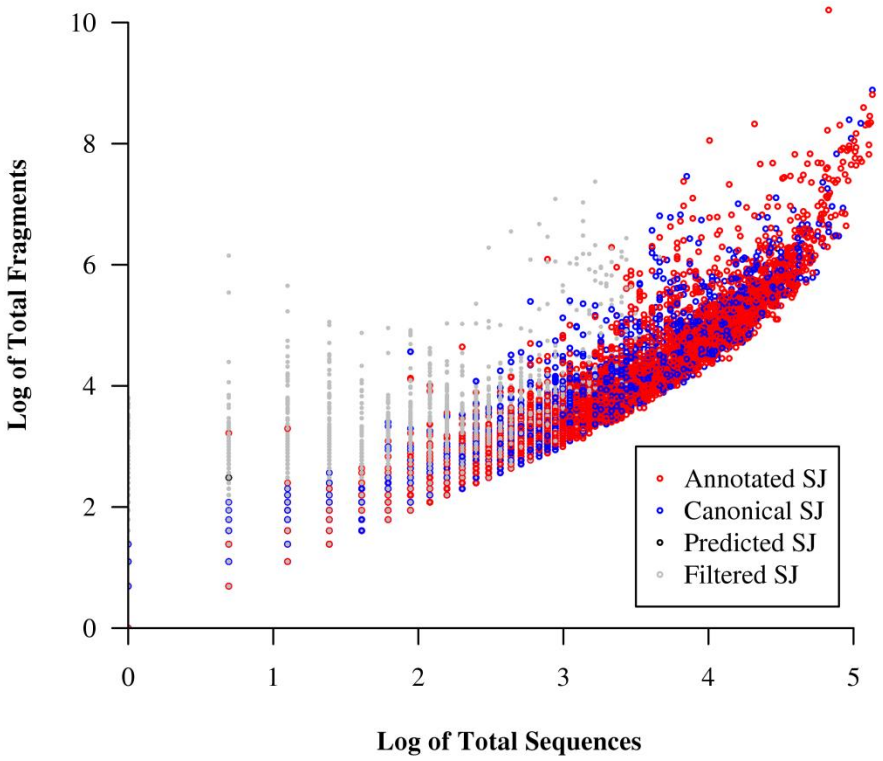


Figure 4.1. Spliced sequences in *Ps. cubensis* are well supported.

Putative splice junction sequences were identified using gapped alignments of RNA-Seq fragments. The total number of read sequences (natural log transformed; x-axis) was plotted as a factor of the number of fragments sequenced (natural log transformed; y-axis). Red = supported a splice junction inferred from the original genome annotation; blue = predicted a new splice junction with either the GT-AG or GC-AG canonical splice site sequences; black = predicted a new splice junction with a novel splice site sequence; gray = filtered out by SuperSplat or lack of sufficient support.

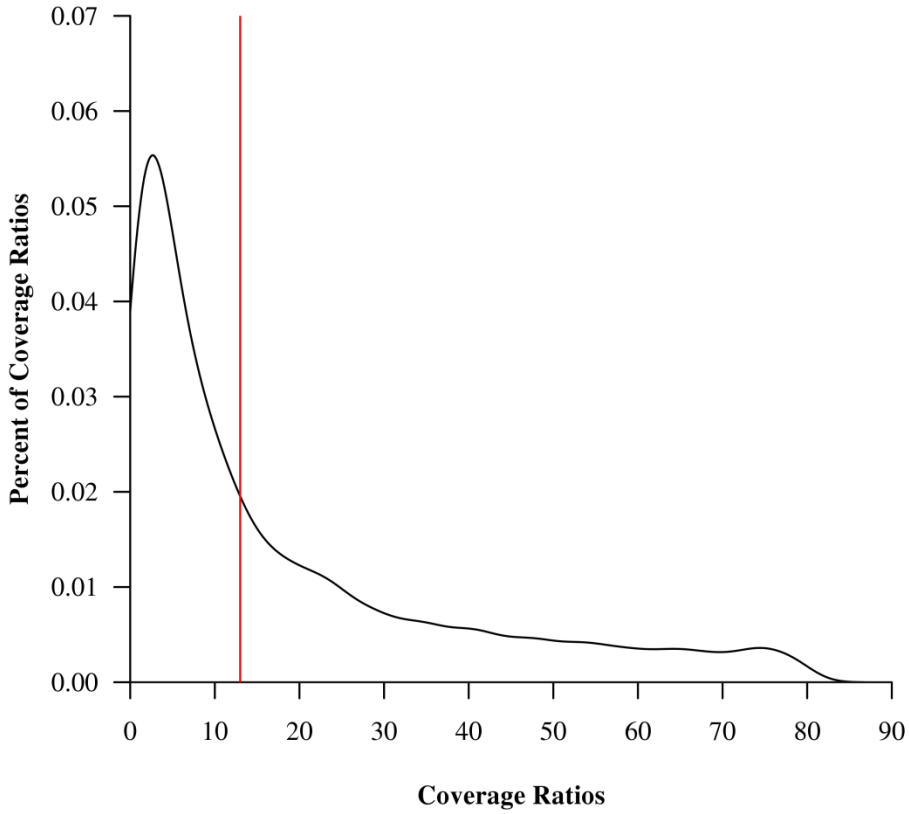


Figure 4.2. Distribution of coverage ratios for *Ps. cubensis*.

Coverage ratios were expressed as a percentage, from 0 - 80%, of the average coverage of an intron over the average coverage of the adjacent exons, plotted on the x-axis above. The distribution of coverage ratios is indicated on the y-axis. The vertical red line indicates the minimum coverage value that was used to call an intron retained. Introns exceeding 80% of the surrounding exon coverage were excluded from the analysis.

Figure 4.3. Distribution of alternative splicing events in *Ps. cubensis* and predicted gene models.

A) Pie chart depicting distribution of alternative splicing types. Percent of alternative splicing events categorized as intron retention, alternative 5′ and 3′ splice site (Alt 5′ SS or Alt 3′ SS, respectively), and exon skipping (0%) of 10,551 alternative splicing events. **B)** Screenshots from GBrowse presenting examples of alternative splicing types in *Ps. cubensis*. Each box represents a different gene (from top to bottom: maker-pcu_contig_579-fgenesh-gene-0.0, maker-pcu_contig_157-snap-gene-0.13, maker-pcu_contig_345-snap-gene-0.10). Within each box, the top row depicts the predicted gene models with pink regions defining UTRs, light blue defining the predicted coding sequence, and connecting lines indicating spliced out introns. The second row is a histogram plot of coverage generated from the alignment of all RNA-Seq reads against the *Ps. cubensis* genome. The last row illustrates the predicted splice junctions found when doing gapped alignments with SuperSplat, with the purple bar representing the sequence alignment, and the dashed line representing the gap in the alignment.

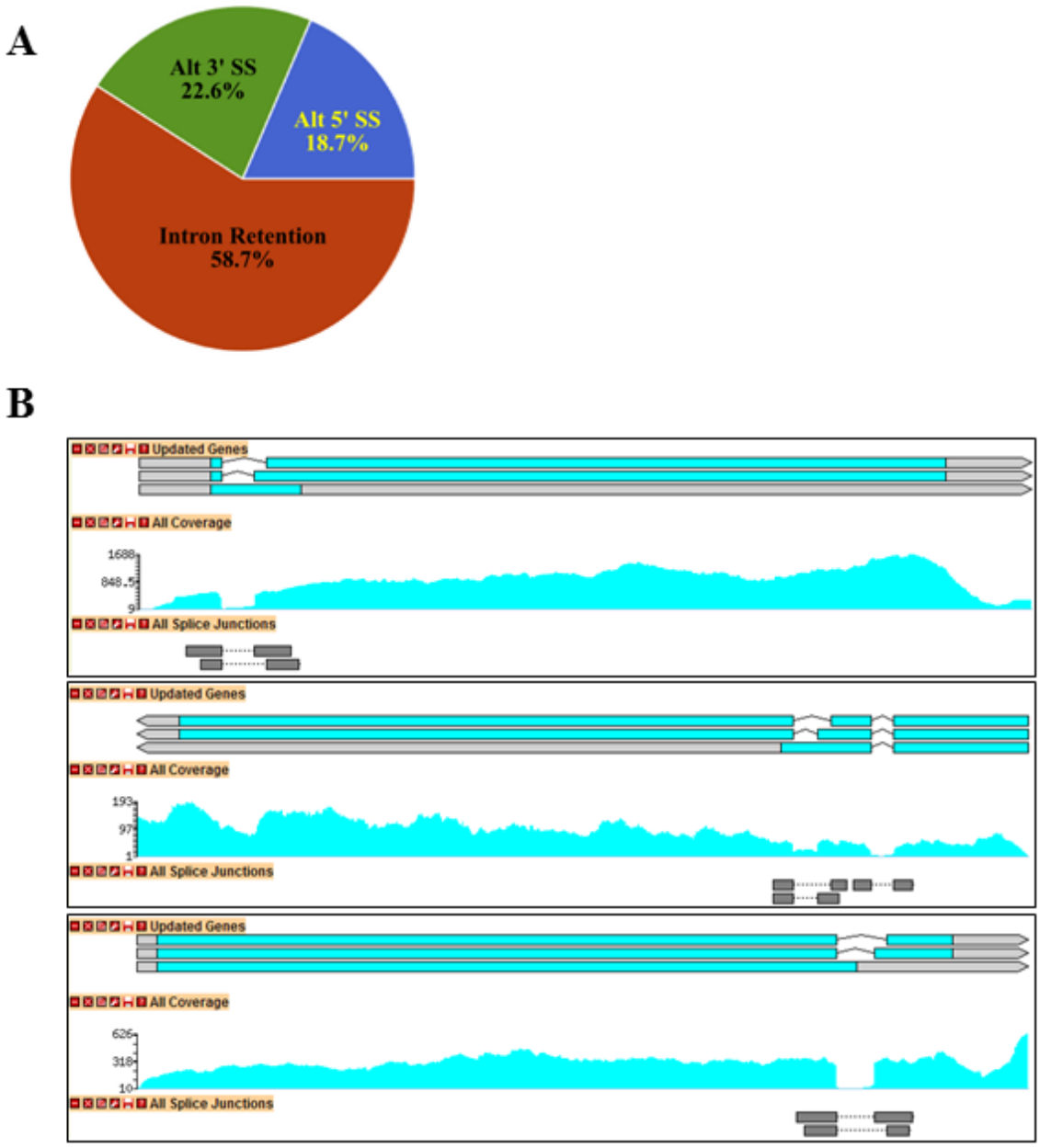


Figure 4.3. Distribution of alternative splicing events in *Ps. cubensis* and predicted gene models.

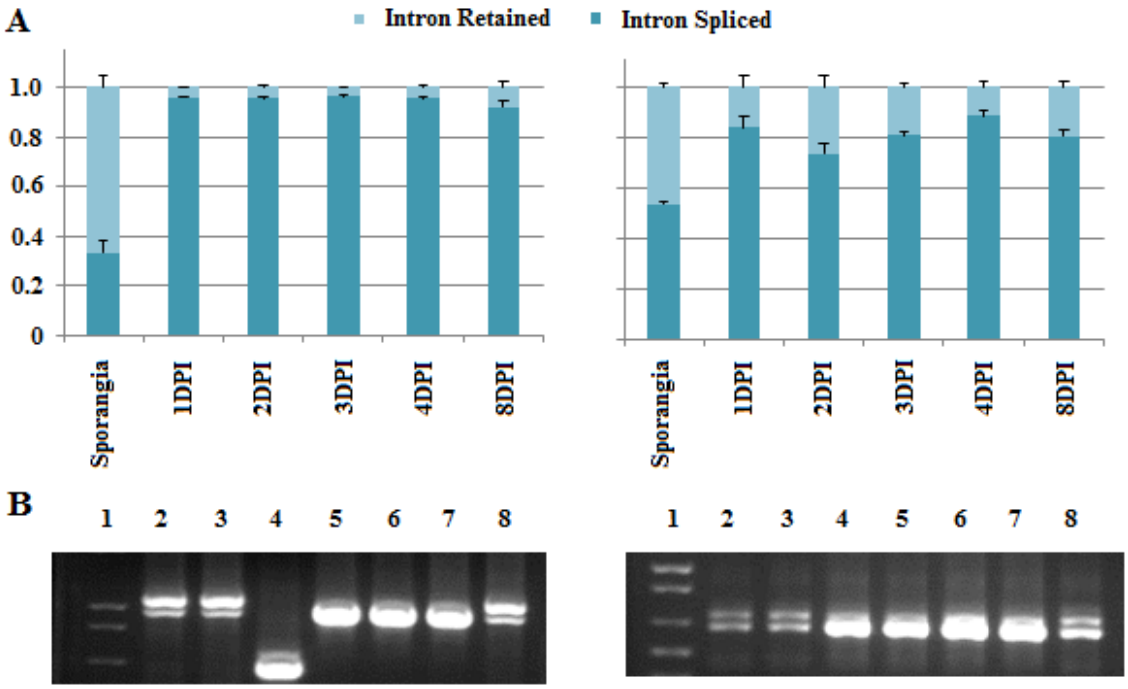


Figure 4.4. qRT-PCR and RT-PCR gel images of verified intron retention events.

A) qRT-PCR was done using isoform specific primers to test for intron retention across multiple time points (sporangia, 1 day post inoculation (DPI), 2DPI, 3DPI, 4DPI, and 8DPI). Depicted are the results for two genes (from left to right: maker-pcu_contig_2364-snap-gene-0.2 a putative endonuclease/exonuclease/phosphatase, and maker-pcu_contig_04986-snap-gene-0.1 a putative glutathione peroxidase). These values were expressed as a percentage of the total product found that either retained the intron or spliced it out. cDNA from 2 separate replicates were used. Error bars are indicated for each time point. **B)** RT-PCR gels were run out for each of the time points for the same genes in the same order as before. For the left gel, lane numbers indicate: 1 is the ladder, 2-3/8 are sporangia replicates, 4/5 are 4 DPI replicates, and 6/7 are 8 DPI replicates. Expected intron sizes are 1045/945 bp for with or without an intron. For the right gel, lane numbers indicate: 1 is the ladder, 2-3/8 are sporangia replicates, 4/5 are 4 DPI replicates, and 6/7 are 8 DPI replicates. All replicates are biological replicates. Expected intron sizes are 759/686 bp for with or without an intron.

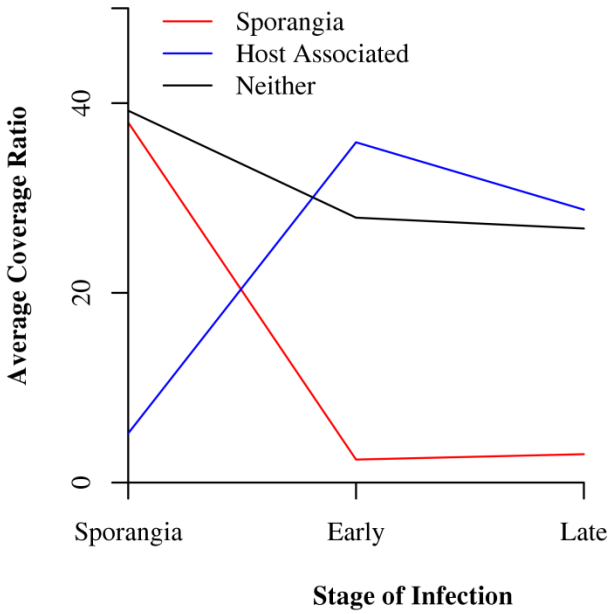


Figure 4.5. Changes in average coverage ratios over stages of infection.

The introns of genes that were differentially expressed and showed evidence of intron retention (~1000) were grouped based on a bias towards intron retention during sporangia (~21.6%); host associated (1.3%) (Early/Late), or neither stage of infection (77.1%). The average coverage ratio for all introns within each group was plotted (y-axis) for each stage of infection (x-axis; Sporangia, Early [2-4 DPI], Late [8 DPI]). Coverage ratios were calculated as a ratio of the intron coverage over the coverage of adjacent exons, and expressed as a percent.

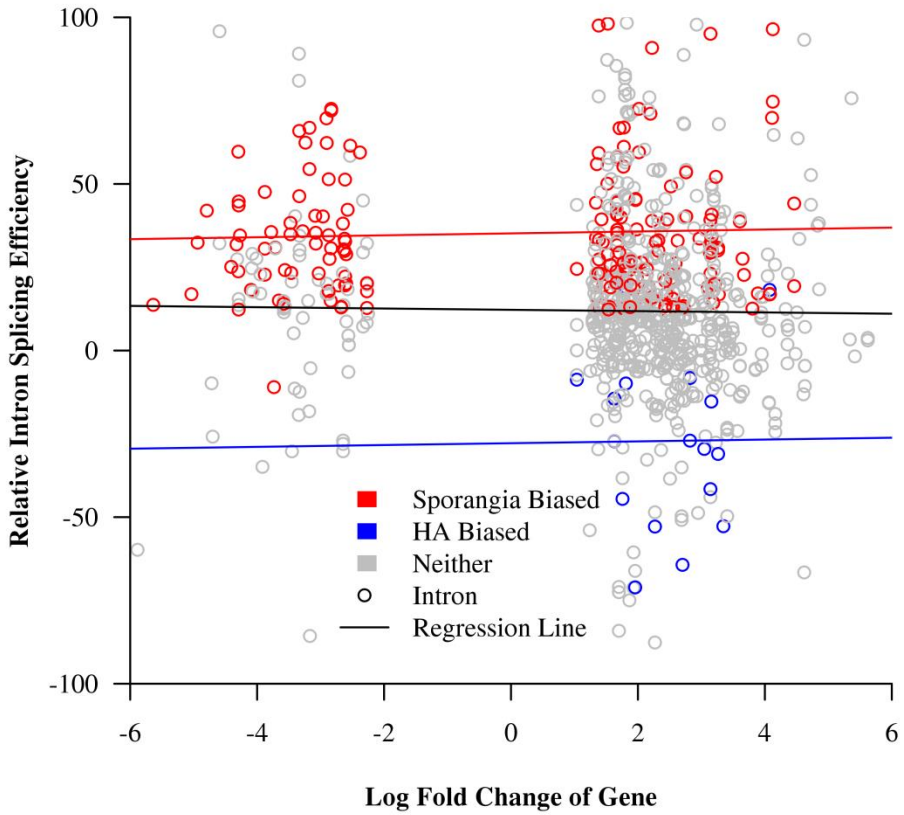


Figure 4.6. Distribution of splicing efficiency as a function of log fold change.

The introns of genes that were differentially expressed and showed evidence of intron retention were group based on a bias towards intron retention during sporangia, host associated (HA; Early/Late), or neither stage of infection. Introns biased for either sporangia, HA, or neither stage of infection are depicted by red, blue, and gray circles respectively. The $\log_2$ fold change of the intron's gene during the HA stage, relative to sporangia, was calculated and plotted (x-axis), against the relative intron splicing efficiency of the intron (y-axis). Splicing efficiency was calculated as a ratio of the coverage ratio in sporangia over the coverage ratio during the HA stage of infection. The coverage ratio was calculated as the ratio of intron coverage over adjacent exon coverage. Regression lines were drawn for each group of data points using the same colors listed above.

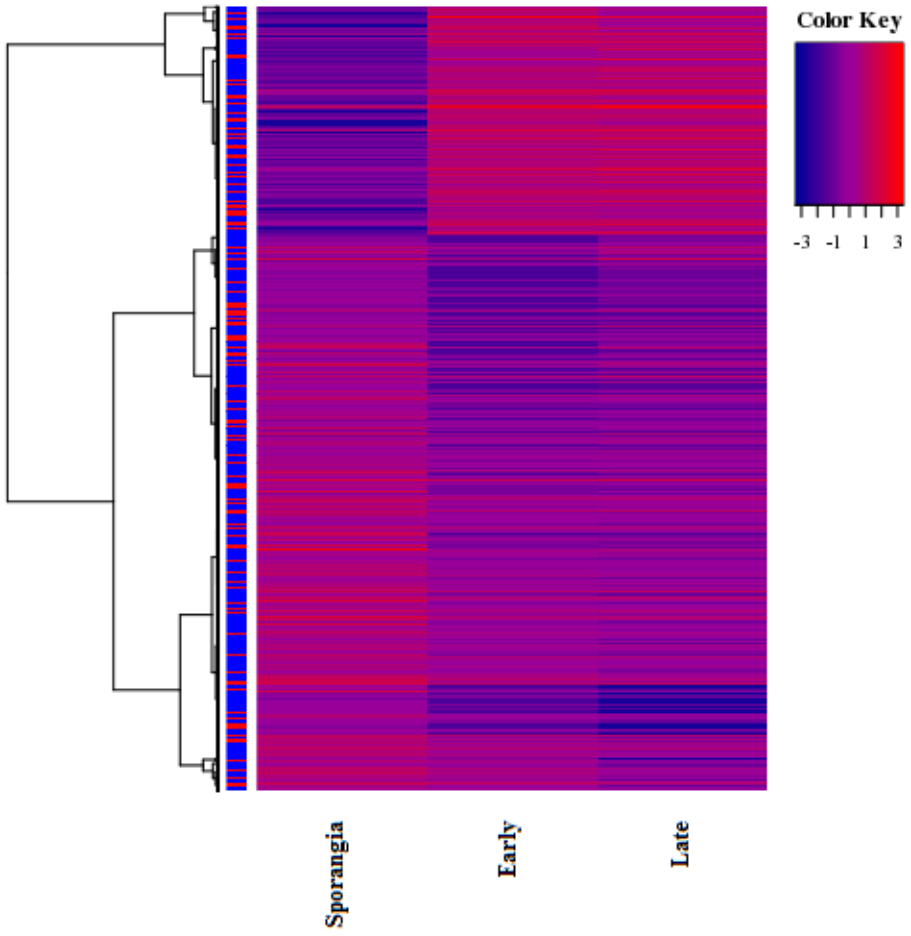


Figure 4.7. Heat map of all differentially expressed genes.

Genes differentially expressed relative to sporangia, were normalized (log-transformed FPKM) and clustered based on expression patterns (dendrogram). The lowest (blue) and highest (red) values define the lower and upper bounds of expression; purple = intermediate expression levels. Next to dendrogram: blue = not alternatively spliced; red = alternatively spliced. Time points are sporangia, 2-4 (early), and 8 (late) dpi.

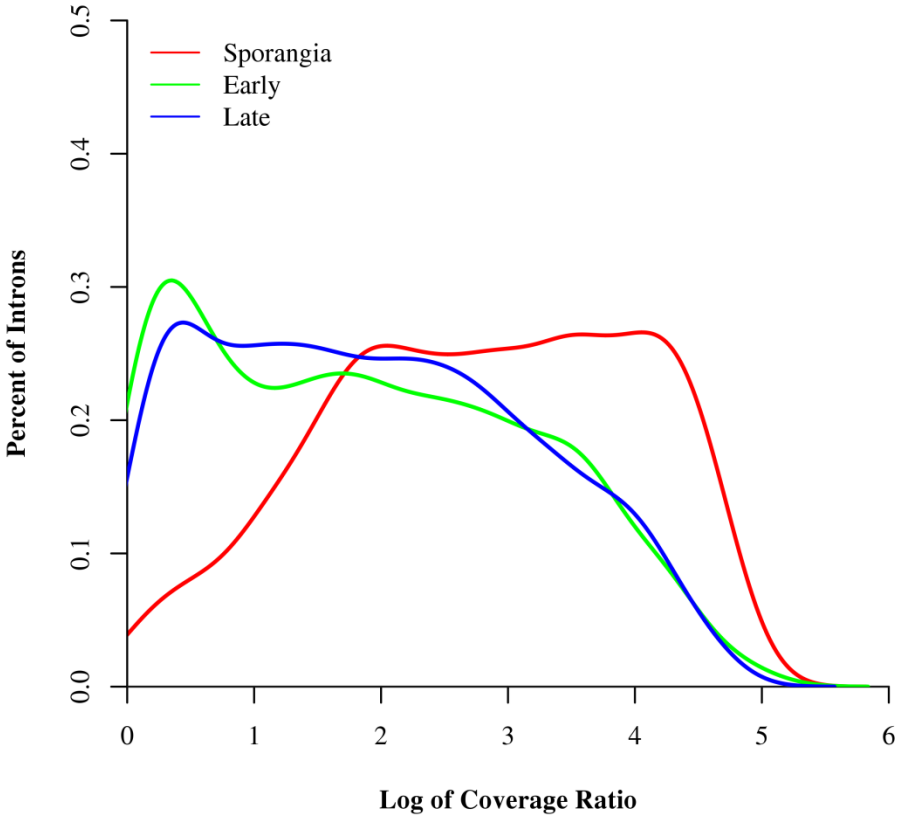


Figure 4.8. Distribution of the log of coverage ratios for all genes with evidence for intron retention within each stage of infection.

Coverage ratios were calculated for each intron that had evidence of intron retention. The coverage ratio was calculated as a ratio of the intron coverage over the coverage of the adjacent exons, expressed as a percent. Coverage ratios were calculated using coverage estimates generated by alignment of RNA-Seq fragments using all the samples for each stage of infection (sporangia, 2-4 DPI for early, and 8 DPI for late). The distribution (y-axis) of the natural log of these coverage ratios were then plotted (x-axis).

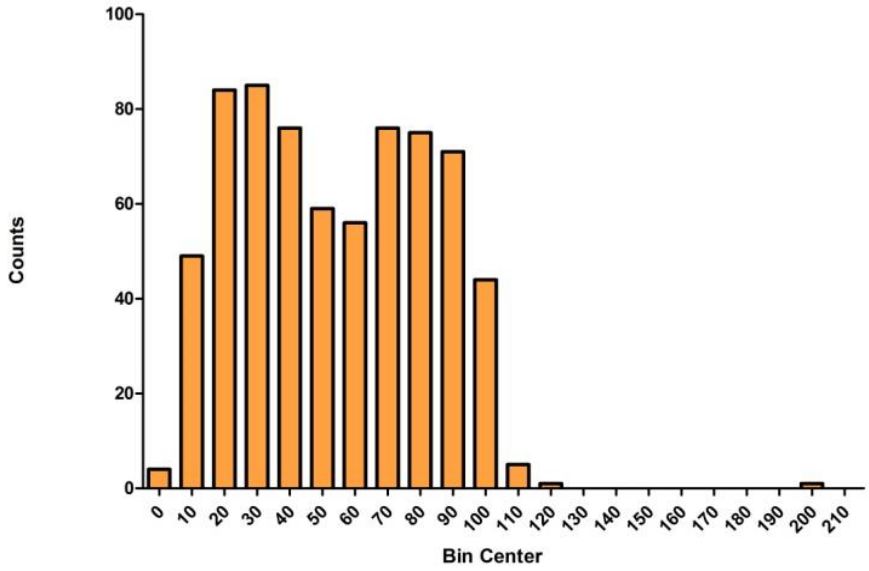


Figure 4.9. Bar chart of ratios of protein length for PTC+/PTC- genes with two predicted isoforms.

All genes predicted to have two isoforms with one containing a PTC were analyzed. The ratio of the lengths of the predicted proteins for PTC+/PTC- isoforms was then binned in multiples of 10 (x-axis). The number of genes within each bin was plotted on the y-axis.

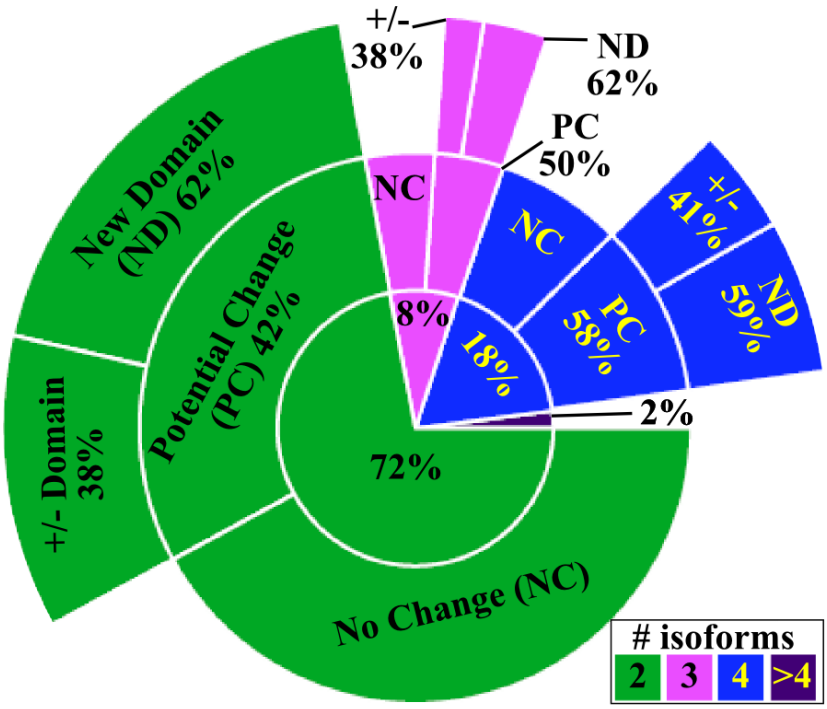


Figure 4.10. Pie Chart representing changes in IPRScan domains found as a result of alternative splicing events in all genes with evidence for alternative splicing.

All genes that showed evidence for alternative splicing were grouped according to the number of transcript isoforms predicted for each gene (green: 2 isoforms, pink: 3 isoforms, blue: 4 isoforms, purple: > 4 isoforms). The domains found for each isoform using IPRSCAN were then compared to each other, and the distribution of differences was enumerated. Shown are pie charts listing the distribution of gene groups amongst all alternatively spliced genes (inner circle), the breakdown of domain changes within each group (middle circle), and a breakdown of the type of potential change found within each group where differences existed(outer circle). 'New Domain' indicates the potential swapping of one domain for another, maintaining the same number of domains. '+/- Domain' indicates a loss or gain in total number of IPRScan domains.

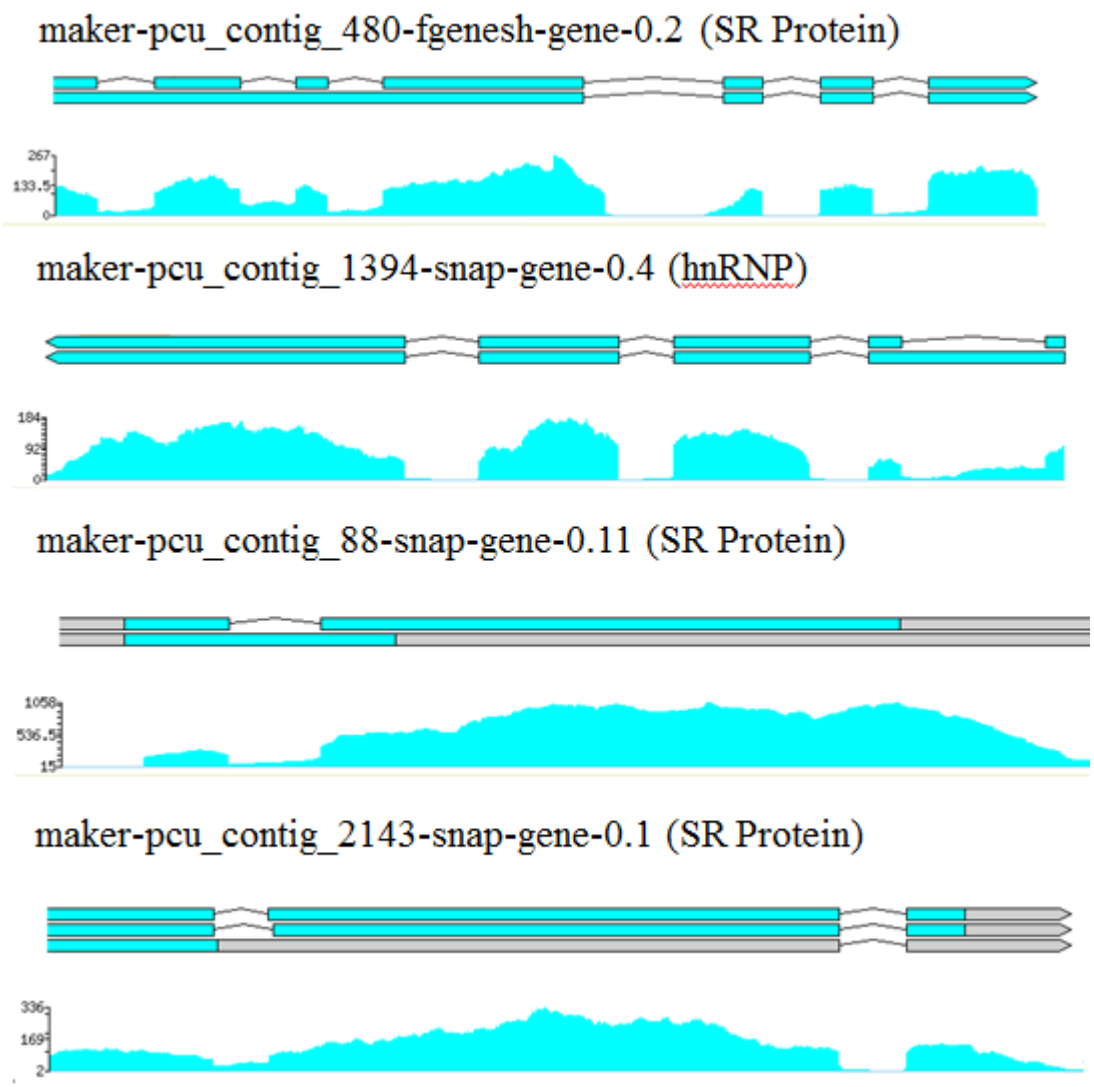


Figure. 4.11. Gbrowse screen-shots of genes important to alternative splicing (SR and hnRNP proteins).

The predicted models for hnRNP and SR proteins predicted to have evidence for alternative splicing. The name and type of each gene are indicated above, with the Gbrowse screen shot below that with all the predicted isoforms and coverage predicted for each gene. Light boxes indicate coding sequences, light pink indicates predicted UTRs, and connecting lines are spliced introns. Coverage is displayed as a histogram plot along the length of the gene. Coverage values were found by pooling all RNA-Seq alignment across the *Ps. cubensis* genome.

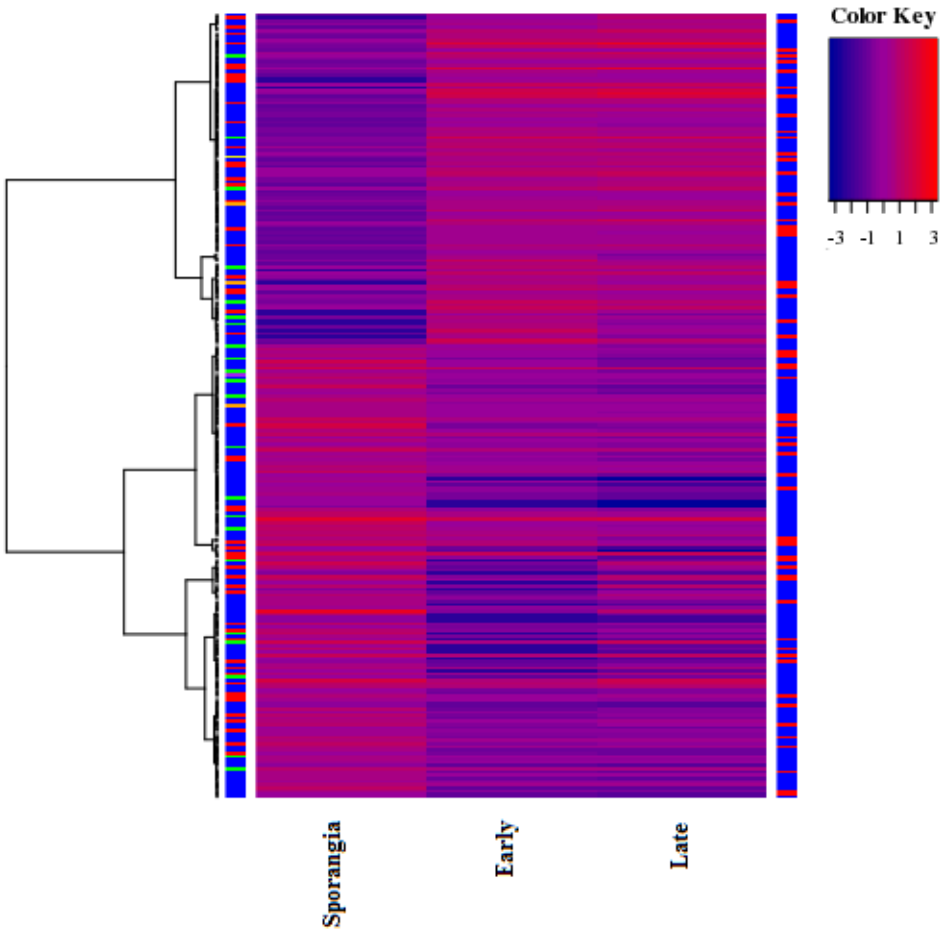


Fig. 4.12. Heat map putatively secreted genes with evidence for differential expression.

Genes differentially expressed relative to sporangia, were normalized (log-transformed FPKM) and clustered based on expression patterns (dendrogram). The lowest (blue) and highest (red) values define the lower and upper bounds of expression; purple = intermediate expression levels. Next to dendrogram on the right: blue = not alternatively spliced; red = alternatively spliced. Time points are sporangia, 2-4 (early), and 8 (late) dpi. Next to dendrogram on the left are the functional categories predicted for each gene (blue = none, red = RXLR, green = CAzyme, purple = crinkler, orange = protease inhibitor, yellow = transcription factor).

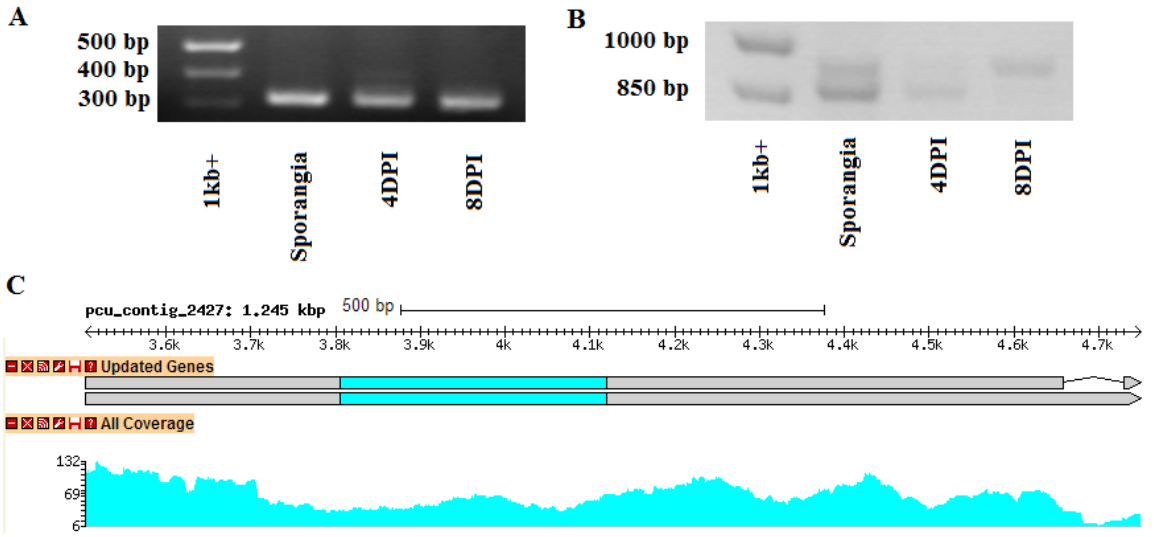


Fig. 4.13. RT-PCR analysis and Gbrowse screenshots of PsCRN2.

A) RT-PCR anaysis of the coding sequence of PsCRN2 (pcu_gene_1703). Representative time points (Sporangia, 4 days post inoculation (DPI), and 8DPI) show expression throughout. **B)** RT-PCR analysis of the 3' untranslated region (UTR) with isoform specific primers which include the intron for representative time points (Sporangia, 4DPI, and 8DPI). **C)** Gbrowse screenshot of the predicted gene. The top row shows the position of the gene in the *Ps. cubensis* genome. The row under 'Updated Genes' shows the isoform structures. The light grey areas represent the UTR, the light blue represents the coding sequence, and the connecting line represents spliced out introns. The row below that shows a histogram plot of the genome coverage found by aligning all read sequences against the genome.

Tabel 4.1. Candidate RxLR effectors.

Gene ID	¥RXLR	€EER	θWY	λIso. #	βSPOR	γHA
maker-pcu_contig_671-snap-gene-0.5	RDLR	0	0	4	4884	50
pcu_gene_2303	RFLR	0	1	2	227	369
maker-pcu_contig_3369-snap-gene-0.1	RTLR	0	0	2	7667	8
pcu_gene_2371	RSLR	0	0	2	2673	78
maker-pcu_contig_764-snap-gene-0.3	RFLR	0	0	2	2434	223
maker-pcu_contig_104-snap-gene-0.21	RLLR	0	0	2	813	5
maker-pcu_contig_1221-snap-gene-0.8	RVLR	0	0	2	567	134
maker-pcu_contig_1892-snap-gene-0.3	RLLR	0	0	2	546	54
maker-pcu_contig_781-snap-gene-0.4 (*Psc_RXLR1*)	RFLR	0	0	2	187	102
pcu_gene_2471	RSLR	0	0	2	37	215
maker-pcu_contig_3126-snap-gene-0.1	RRLR	0	0	2	21	155
snap_masked-pcu_contig_342-abinit-gene-0.19	RLLR	0	0	2	9	27
maker-pcu_contig_3292-snap-gene-0.0	RYLR	1	1	1	370	27
maker-pcu_contig_838-snap-gene-0.2	RFLR	1	1	1	207	14
pcu_gene_377	RSLR	1	0	1	1437	29
pcu_gene_3024	RLLR	1	0	1	779	57
pcu_gene_940	RSLR	1	0	1	708	75
pcu_gene_2286	RRLR	1	0	1	186	865
pcu_gene_3032	RRLR	1	0	1	149	167
maker-pcu_contig_1374-snap-gene-0.4	RFLR	1	0	1	137	19
pcu_gene_783	RYLR	1	0	1	73	209
pcu_gene_164	RSLR	1	0	1	71	84
pcu_gene_2457	RSLR	1	0	1	60	51
pcu_gene_2896	RSLR	1	0	1	44	29
pcu_gene_824	RYLR	1	0	1	24	33
pcu_gene_2897	RHLR	1	0	1	6	112
pcu_gene_1388	RMLR	1	0	1	5	266
pcu_gene_2664	RSLR	1	0	1	3	139
pcu_gene_258	RSLR	1	0	1	0	17
maker-pcu_contig_2647-snap-gene-0.0	RLLR	0	1	1	2252	55
maker-pcu_contig_5041-snap-gene-0.1	RYLR	0	1	1	180	15
maker-pcu_contig_12074-snap-gene-0.0	RHLR	0	1	1	150	5
maker-pcu_contig_5147-snap-gene-0.1	RHLR	0	1	1	90	4
pcu_gene_2639	RSLR	0	1	1	62	26
pcu_gene_2815	RFLR	0	1	1	58	150

maker-pcu_contig_9754-snap-gene-0.0	RSLR	0	1	1	22	58
maker-pcu_contig_12150-snap-gene-0.0	RHLR	0	1	1	18	39
maker-pcu_contig_3681-snap-gene-0.1	RFLR	0	1	1	10	4
maker-pcu_contig_3341-snap-gene-0.1	RFLR	0	1	1	7	10
maker-pcu_contig_10463-snap-gene-0.0	RGLR	0	1	1	7	1
maker-pcu_contig_709-snap-gene-0.8	RRLR	0	0	1	3368	45
maker-pcu_contig_2392-snap-gene-0.2	RRLR	0	0	1	3097	209
maker-pcu_contig_10792-snap-gene-0.1	RPLR	0	0	1	1979	50
maker-pcu_contig_409-snap-gene-0.7	RLLR	0	0	1	1264	109
maker-pcu_contig_2696-snap-gene-0.2	RVLR	0	0	1	1121	19
maker-pcu_contig_118-fgenesh-gene-0.2	RLLR	0	0	1	722	569
maker-pcu_contig_58-fgenesh-gene-0.4	RTLR	0	0	1	587	642
maker-pcu_contig_361-snap-gene-0.6	RFLR	0	0	1	515	217
maker-pcu_contig_6018-fgenesh-gene-0.0	RSLR	0	0	1	328	30
maker-pcu_contig_2566-snap-gene-0.2	RSLR	0	0	1	303	78
maker-pcu_contig_647-snap-gene-0.8	RHLR	0	0	1	286	111
maker-pcu_contig_4158-snap-gene-0.0	RTLR	0	0	1	268	87
maker-pcu_contig_82-snap-gene-0.19	RKLR	0	0	1	234	20
maker-pcu_contig_210-snap-gene-0.17	RHLR	0	0	1	218	68
maker-pcu_contig_830-snap-gene-0.4	RDLR	0	0	1	197	300
maker-pcu_contig_3553-snap-gene-0.1	RFLR	0	0	1	195	107
maker-pcu_contig_3525-snap-gene-0.0	RTLR	0	0	1	161	81
maker-pcu_contig_6909-snap-gene-0.2	RALR	0	0	1	146	27
maker-pcu_contig_186-snap-gene-0.10	RSLR	0	0	1	127	48
pcu_gene_2283	RSLR	0	0	1	120	0
maker-pcu_contig_2913-snap-gene-0.1	RELR	0	0	1	103	60
snap_masked-pcu_contig_3981-abinit-gene-0.4	RALR	0	0	1	93	87
pcu_gene_2812	RSLR	0	0	1	76	121
maker-pcu_contig_3359-snap-gene-0.1	RMLR	0	0	1	53	3
maker-pcu_contig_12733-snap-gene-0.0	RRLR	0	0	1	45	13
maker-pcu_contig_6167-snap-gene-0.1	RSLR	0	0	1	37	4
pcu_gene_1680	RFLR	0	0	1	35	48
pcu_gene_2433	RPLR	0	0	1	24	3
maker-pcu_contig_4588-fgenesh-gene-0.0	RALR	0	0	1	23	172
maker-pcu_contig_3836-snap-gene-0.3	RTLR	0	0	1	22	4
maker-pcu_contig_1451-snap-gene-0.2	RTLR	0	0	1	21	7
maker-pcu_contig_363-snap-gene-0.10	RALR	0	0	1	20	28
maker-pcu_contig_14864-snap-gene-0.0	RMLR	0	0	1	14	41
maker-pcu_contig_1066-snap-gene-0.3	RDLR	0	0	1	11	13

pcu_gene_2562	RFLR	0	0	1	10	34
pcu_gene_2237	RLLR	0	0	1	8	233
pcu_gene_2844	RALR	0	0	1	3	17
pcu_gene_1879	RRLR	0	0	1	1	54
pcu_gene_2128	RMLR	0	0	1	1	52

* [¥] RXLR lists the identified RxLR sequence in the protein. [€] EER – denotes whether an EER domain was found (1 found, 0 not found). [θ] WY denotes whether a WY C-terminal domain was found (1 found, 0 not found). [λ] Iso. # specifies the number of predicted transcripts found for this gene. [β] SPOR and [γ] HA show the number of average counts for their respective time points (SPOR = Sporangia, HA = 2, 3, 4, and 8 days post inoculation). Genes beginning with 'pcu_gene_' indicate newly predicted genes.

Table 4.2. Predicted and candidate Crinkler Proteins.

Gene ID	$^{\lambda}$ Iso. #	$^{\theta}$Sig.	$^{\beta}$SPOR	$^{\gamma}$HA
maker-pcu_contig_14858-snap-gene-0.0	2	YES	1065	419
pcu_gene_1703	2	YES	427	39
pcu_gene_2449	1	YES	572	6
maker-pcu_contig_408-snap-gene-0.2	4	NO	3188	3893
maker-pcu_contig_5713-fgenesh-gene-0.0	2	NO	105	76
maker-pcu_contig_18337-snap-gene-0.0	2	NO	46	0
maker-pcu_contig_1326-snap-gene-0.0	2	NO	41	1
maker-pcu_contig_15161-snap-gene-0.0	2	NO	40	24
maker-pcu_contig_500-snap-gene-0.1	2	NO	36	45
maker-pcu_contig_8453-snap-gene-0.0	1	NO	17013	12863
maker-pcu_contig_13266-snap-gene-0.0	1	NO	7383	578
maker-pcu_contig_4377-snap-gene-0.0	1	NO	2169	341
maker-pcu_contig_109-snap-gene-0.17	1	NO	1757	587
maker-pcu_contig_6508-snap-gene-0.1	1	NO	788	258
maker-pcu_contig_19985-snap-gene-0.0	1	NO	695	197
maker-pcu_contig_12177-snap-gene-0.0	1	NO	675	32
maker-pcu_contig_9941-snap-gene-0.0	1	NO	639	1
maker-pcu_contig_1780-snap-gene-0.3	1	NO	605	265
maker-pcu_contig_13898-snap-gene-0.0	1	NO	599	1
maker-pcu_contig_9086-snap-gene-0.0	1	NO	483	3
maker-pcu_contig_23172-snap-gene-0.0	1	NO	320	12
maker-pcu_contig_2379-snap-gene-0.2	1	NO	312	112
maker-pcu_contig_9750-snap-gene-0.0	1	NO	293	260
maker-pcu_contig_18050-snap-gene-0.0	1	NO	280	28
maker-pcu_contig_9328-snap-gene-0.0	1	NO	254	327
maker-pcu_contig_2347-snap-gene-0.3	1	NO	248	10
maker-pcu_contig_12900-snap-gene-0.0	1	NO	231	108
maker-pcu_contig_4839-snap-gene-0.1	1	NO	187	13
maker-pcu_contig_10631-snap-gene-0.1	1	NO	184	28
maker-pcu_contig_15908-snap-gene-0.0	1	NO	182	152
maker-pcu_contig_24064-snap-gene-0.0	1	NO	175	10
maker-pcu_contig_2473-snap-gene-0.4	1	NO	153	52
maker-pcu_contig_2903-snap-gene-0.0	1	NO	149	51
maker-pcu_contig_20766-snap-gene-0.0	1	NO	127	5
maker-pcu_contig_13687-snap-gene-0.0	1	NO	97	173
maker-pcu_contig_35203-snap-gene-0.0	1	NO	94	28
maker-pcu_contig_12774-snap-gene-0.0	1	NO	91	77

maker-pcu_contig_8208-snap-gene-0.0	1	NO	90	3
maker-pcu_contig_74-snap-gene-0.15	1	NO	89	30
maker-pcu_contig_23757-snap-gene-0.0	1	NO	87	43
maker-pcu_contig_2473-snap-gene-0.5	1	NO	71	49
snap-pcu_contig_7144-abinit-gene-0.2	1	NO	70	17
snap_masked-pcu_contig_8943-abinit-gene-0.2	1	NO	65	17
maker-pcu_contig_29447-snap-gene-0.0	1	NO	59	12
maker-pcu_contig_4198-snap-gene-0.0	1	NO	57	19
maker-pcu_contig_24302-snap-gene-0.0	1	NO	57	55
maker-pcu_contig_520-snap-gene-0.3	1	NO	51	96
maker-pcu_contig_12701-snap-gene-0.0	1	NO	49	52
maker-pcu_contig_4839-snap-gene-0.2	1	NO	45	6
maker-pcu_contig_23913-snap-gene-0.0	1	NO	39	17
maker-pcu_contig_7680-snap-gene-0.1	1	NO	37	22
maker-pcu_contig_27752-snap-gene-0.0	1	NO	36	28
maker-pcu_contig_19470-snap-gene-0.0	1	NO	35	26
maker-pcu_contig_2047-snap-gene-0.7	1	NO	34	3
maker-pcu_contig_5713-snap-gene-0.3	1	NO	30	49
maker-pcu_contig_25909-snap-gene-0.0	1	NO	30	23
maker-pcu_contig_2712-snap-gene-0.2	1	NO	28	2
maker-pcu_contig_2047-snap-gene-0.1	1	NO	27	15
maker-pcu_contig_21035-snap-gene-0.0	1	NO	27	63
maker-pcu_contig_16810-snap-gene-0.1	1	NO	26	1
maker-pcu_contig_1642-snap-gene-0.2	1	NO	24	37
maker-pcu_contig_12576-snap-gene-0.1	1	NO	24	11
maker-pcu_contig_14518-snap-gene-0.1	1	NO	24	21
maker-pcu_contig_14445-snap-gene-0.0	1	NO	23	9
maker-pcu_contig_52-snap-gene-0.12	1	NO	21	11
maker-pcu_contig_1817-snap-gene-0.3	1	NO	21	0
maker-pcu_contig_10727-snap-gene-0.0	1	NO	19	7
maker-pcu_contig_473-snap-gene-0.3	1	NO	18	6
maker-pcu_contig_571-snap-gene-0.8	1	NO	18	27
maker-pcu_contig_24289-snap-gene-0.0	1	NO	17	4
maker-pcu_contig_673-snap-gene-0.4	1	NO	14	4
maker-pcu_contig_2473-snap-gene-0.3	1	NO	14	4
maker-pcu_contig_10797-snap-gene-0.0	1	NO	13	18
maker-pcu_contig_24384-snap-gene-0.0	1	NO	13	1
maker-pcu_contig_83-snap-gene-0.14	1	NO	12	3
maker-pcu_contig_925-snap-gene-0.6	1	NO	12	0

maker-pcu_contig_7416-snap-gene-0.0	1	NO	12	0
maker-pcu_contig_33003-snap-gene-0.0	1	NO	11	0
maker-pcu_contig_25814-snap-gene-0.0	1	NO	10	10
maker-pcu_contig_6396-snap-gene-0.0	1	NO	8	30
maker-pcu_contig_24502-snap-gene-0.0	1	NO	6	13
maker-pcu_contig_6331-snap-gene-0.3	1	NO	4	20

* $^{\lambda}$ Iso. # specifies the number of predicted transcripts found for this gene. $^{\theta}$ Sig. specifies whether a signal domain was found or not. $^{\beta}$ SPOR and $^{\gamma}$ HA show the number of average counts for their respective time points (SPOR = Sporangia, HA = 2, 3, 4, and 8 days post inoculation). Genes beginning with 'pcu_gene_' indicate newly predicted genes.

Conclusions and Future Directions

Jason Cumbie

With the advent of high-throughput sequencing and the subsequent development of RNA-Seq, genome-enabled studies have become the new standard by which to interrogate molecular interactions. They have allowed the rapid development of draft genomes, single base resolution of transcript isoforms for whole transcriptomes, and made it possible to test thousands of genes simultaneously for differential expression under a variety of biologically relevant conditions producing numerous new hypotheses to test. The cost of this new technology has also made it difficult to produce a large number of biological replicates, necessitating ever more sophisticated statistical approaches to analyzing the data. Fortunately both multi-plexing and increases in the depth of sequence are quickly making this less of a limitation. In the meantime, however, more sophisticated computational tools and statistical models will be needed to address the challenges of using highthroughput sequencing. To that end, the work described herein has established the development of a scale-able computational program, GENE-counter (Chapter II), to aid scientists to simplify this process.

GENE-counter includes a number of simple methods to automate the process of data processing that allow the end-user to more easily generate count data for their RNA-Seq experiments. It has also made possible the use of the newly developed statistical method, NBPSeq, as well as both the edgeR and DESeq set of packages to test for differential expression without needing a strong background in the R statistical language.

However, GENE-counter still requires some simple command line skills to operate. In the future, an even further simplified graphical user interface (GUI) could address this, and would help biologists with experimental design and analysis. GENE-

counter was also initially built to handle single-end data, and still needs to have fully developed methods to handle pair-end sequencing data. In the future it would be prudent to push GENE-counter to further process this next stage in high-throughput sequencing to allow even more precise measurements of gene expression and isoform abundances to interrogate the expression of alternatively spliced transcripts. The incorporation of paired-end alignments to handle unique characteristics of paired-end data such as the spacer sequence between reads, the unique orientation of paired-end reads, and the ability to capture a wide range of variability in transcript isoforms will greatly improve resolution of multiple transcript isoforms.

Genome-enabled experiments have improved our ability to interrogate the transcriptomes of obligate plant pathogens organisms. This is especially beneficial since direct experiments via transformation of these organisms cannot be done, so the ultra-deep sequencing of transcriptional changes greatly enhances the power of experimental inferences in the context of disease. The fact that RNA populations derived from these organisms will necessarily make up a small fraction of total RNA samples due to the greater abundance of host tissue has made the solution of ultra-deep sequencing via RNA-Seq a highly attractive prospect. While at first glance this would appear to make microarrays a more attractive solution, the fact that RNA-Seq does not need a reference sequence and can allow for prediction of transcription isoforms without any *a priori* assumptions makes this an ideal technology for interrogating a multitude of species. In the work characterized within, we deeply resequenced an RNA population of the obligate

oomycete plant pathogen *Ps. cubensis* to improve the genome annotation as well as interrogate alternative splicing (Chapter III & IV).

An RxLR effector of *Ps. cubensis* was recently shown to be alternatively spliced. However, interrogation of alternative splicing has received little attention and appears to be an underappreciated mode of regulation in oomycetes pathogens. To that end, we developed new computational methods to improve the genome models of the draft genome of *Ps. cubensis* and aid the interrogation of alternatively spliced transcripts.

This work highlights the potential that alternative splicing has in oomycete pathogenesis. We have found a large number of candidate effector proteins which have evidence for stage dependent alternative splicing over the course of infection. A number of these genes also were shown to be differentially expressed further implicating their importance in pathogenesis. These genes will need to be cloned to test the alternative isoforms of these candidate effector proteins, as well as functionally characterize these genes test for secretion, subcellular localization, and the elicitation of defense responses to understand the biological relevance of these candidate effectors. Future work will need to establish the effects alternative splicing has on protein function and to what degree the change in function or expression is critical to the use of alternative splicing in *Ps. cubensis*. It will be critical to further study these mechanisms to generate a broader picture of pathogenesis in oomycetes and to aid our understanding of how these pathogens cause disease; it would allow us to find new potential resistance targets by which to aid the control of these devastating plant pathogens.

BIBLIOGRAPHY

Adhikari, B. N., E. A. Savory, et al. (2012). "Expression profiling of Cucumis sativus in response to infection by Pseudoperonospora cubensis." PLoS One 7(4): e34954.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol 215(3): 403-10.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biol 11(10): R106.

Arabidopsis Genome Initiative (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." Nature 408(6814): 796-815.

Arauz, L. F., K. N. Neufeld, et al. (2010). "Quantitative models for germination and infection of Pseudoperonospora cubensis in response to temperature and duration of leaf wetness." Phytopathology 100(9): 959-67.

Armstrong, M. R., S. C. Whisson, et al. (2005). "An ancestral oomycete locus contains late blight avirulence gene Avr3a, encoding a protein that is recognized in the host cytoplasm." Proc Natl Acad Sci U S A 102(21): 7766-71.

Ast, G. (2004). "How did alternative splicing evolve?" Nat Rev Genet 5(10): 773-82.

Bauer, S., S. Grossmann, et al. (2008). "Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration." Bioinformatics 24(14): 1650-1.

Baxter, L., S. Tripathy, et al. (2010). "Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome." Science 330(6010): 1549-51.

Beakes, G. W., S. L. Glockling, et al. (2012). "The evolutionary phylogeny of the oomycete "fungi"." Protoplasma 249(1): 3-19.

Bhattacharyya, M. K., N. N. Narayanan, et al. (2005). "Identification of a large cluster of coiled coil-nucleotide binding site--leucine rich repeat-type genes from the Rps1 region containing Phytophthora resistance genes in soybean." Theor Appl Genet 111(1): 75-86.

Birch, P. R. J., P. C. Boevink, et al. (2008). "Oomycete RXLR effectors: delivery, functional redundancy and durable disease resistance." Current Opinion in Plant Biology 11(4): 373.

Bittner-Eddy, P. D., I. R. Crute, et al. (2000). "RPP13 is a simple locus in Arabidopsis thaliana for alleles that specify downy mildew resistance to different avirulence determinants in Peronospora parasitica." Plant J 21(2): 177-88.

Blum, M., M. Waldner, et al. (2011). "Resistance mechanism to carboxylic acid amide fungicides in the cucurbit downy mildew pathogen Pseudoperonospora cubensis." Pest Manag Sci 67(10): 1211-4.

Boller, T. and G. Felix (2009). "A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors." Annu Rev Plant Biol 60: 379-406.

Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics 19(2): 185-93.

Boutemy, L. S., S. R. King, et al. "Structures of Phytophthora RXLR effector proteins: a conserved but adaptable fold underpins functional diversity." J Biol Chem 286(41): 35834-42.

Bryant, D. W., Jr., H. D. Priest, et al. (2012). "Detection and quantification of alternative splicing variants using RNA-seq." Methods Mol Biol 883: 97-110.

Bryant, D. W., Jr., R. Shen, et al. (2010). "Supersplat--spliced RNA-seq alignment." Bioinformatics 26(12): 1500-5.

Buell, C. R., V. Joardar, et al. (2003). "The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000." Proc Natl Acad Sci U S A 100(18): 10181-6.

Bullard, J. H., E. Purdom, et al. (2010). "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments." BMC Bioinformatics 11: 94.

Burki, F., P. Flegontov, et al. (2012). "Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin." Genome Biol Evol 4(6): 626-35.

Chang, J. H., J. M. Urbach, et al. (2005). "A high-throughput, near-saturating screen for type III effector genes from Pseudomonas syringae." Proc Natl Acad Sci U S A 102(7): 2549–2554.

Choi, J., K. Cheong, et al. (2012). "CFGP 2.0: a versatile web-based platform for supporting comparative and evolutionary genomics of fungi and Oomycetes." Nucleic Acids Res 41(Database issue): D714-9.

Civelek, M., R. Hagopian, et al. (2013). "Genetic regulation of human adipose microRNA expression and its consequences for metabolic traits." Hum Mol Genet.

Cohen, Y., I. Meron, et al. (2003). "A new pathotype of Pseudoperonospora cubensis

causing downy mildew in cucurbits in Israel." Phytoparasitica 31: 458–466.

Curtis, B. A., G. Tanifuji, et al. (2012). "Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs." Nature 492(7427): 59-65.

Dangl, J. L. and J. D. Jones (2001). "Plant pathogens and integrated defence responses to infection." Nature 411(6839): 826-33.

Deng, W. L., G. Preston, et al. (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.

Denoux, C., R. Galletti, et al. (2008). "Activation of defense response pathways by OGs and Flg22 elicitors in Arabidopsis seedlings." Mol Plant 1(3): 423-45.

Di, Y., D. W. Schafer, et al. (2011). "The NBP negative binomial model for assessing differential gene expression from RNA-seq." Stat Appl Genet Mol Biol 10: Article 24.

Dodds, P. N. and J. P. Rathjen (2010). "Plant immunity: towards an integrated view of plant-pathogen interactions." Nat Rev Genet 11(8): 539-48.

Eperon, I. C., D. C. Ireland, et al. (1993). "Pathways for selection of 5' splice sites by U1 snRNPs and SF2/ASF." Embo J 12(9): 3607-17.

Eulgem, T. and I. E. Somssich (2007). "Networks of WRKY transcription factors in defense signaling." Curr Opin Plant Biol 10(4): 366-71.

Fahlgren, N., C. M. Sullivan, et al. (2009). "Computational and analytical framework for small RNA profiling by high-throughput sequencing." Rna 15(5): 992-1002.

Fan, H., Y. Xiao, et al. (2013). "RNA-Seq Analysis of Cocos nucifera: Transcriptome Sequencing and De Novo Assembly for Subsequent Functional Genomics Approaches." PLoS One 8(3): e59997.

Feil, H., W. S. Feil, et al. (2005). "Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000." Proc Natl Acad Sci U S A 102(31): 11064-9.

Feng, F. and J. M. Zhou (2012). "Plant-bacterial pathogen interactions mediated by type III effectors." Curr Opin Plant Biol 15(4): 469-76.

Filichkin, S. A., H. D. Priest, et al. (2010). "Genome-wide mapping of alternative splicing in Arabidopsis thaliana." Genome Res 20(1): 45-58.

Fox, S., S. Filichkin, et al. (2009). "Applications of ultra-high-throughput sequencing." Methods Mol Biol 553: 79-108.

Garber, M., M. G. Grabherr, et al. (2011). "Computational methods for transcriptome annotation and quantification using RNA-seq." Nat Methods 8(6): 469-77.

Glazebrook, J., W. Chen, et al. (2003). "Topology of the network integrating salicylate and jasmonate signal transduction derived from global expression phenotyping." Plant J 34(2): 217-28.

Goncalves, A., A. Tikhonov, et al. (2011). "A pipeline for RNA-seq data processing and quality assessment." Bioinformatics 27(6): 867-9.

Grabherr, M. G., B. J. Haas, et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol 29(7): 644-52.

Granke, L. L. and M. K. Hausbeck (2011). "Dyanmics of airborne Pseudoperonospora cubensis sporangia in commercial cucurbit fields in Michigan." Plant Dis 95: 1392-1400.

Grant, G. R., J. Liu, et al. (2005). "A practical false discovery rate approach to identifying patterns of differential expression in microarray data." Bioinformatics 21(11): 2684-90.

Graveley, B. R., A. N. Brooks, et al. (2010). "The developmental transcriptome of Drosophila melanogaster." Nature 471(7339): 473.

Grossmann, S., S. Bauer, et al. (2007). "Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis." Bioinformatics 23(22): 3024-31.

Grünwald, N. J., M. Garbelotto, et al. (2012). "Emergence of the sudden oak death pathogen Phytophthora ramorum." Trends Microbiol 20(3): 131-8.

Gunaratne, P. H., C. Coarfa, et al. (2012). "miRNA data analysis: next-gen sequencing." Methods Mol Biol 822: 273-88.

Haas, B. J., S. Kamoun, et al. (2009). "Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans." Nature 461(7262): 393-8.

Hackett, J. D., H. S. Yoon, et al. (2007). "Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates." Mol Biol Evol 24(8): 1702-13.

Han, J., J. Xiong, et al. (2011). "Pre-mRNA splicing: where and when in the nucleus." Trends Cell Biol 21(6): 336-43.

Hegyi, H., L. Kalmar, et al. (2011). "Verification of alternative splicing variants based on

domain integrity, truncation length and intrinsic protein disorder." Nucleic Acids Res 39(4): 1208-19.

Henderson, L. (2006). "The Irish Famine: A Historiagraphical Review." Historia 15: 133-140.

Hori, K. and Y. Watanabe (2007). "Context analysis of termination codons in mRNA that are recognized by plant NMD." Plant Cell Physiol 48(7): 1072-8.

Hoskins, A. A., J. Gelles, et al. (2011). "New insights into the spliceosome by single molecule fluorescence microscopy." Curr Opin Chem Biol 15(6): 864-70.

Hou, S., Y. Yang, et al. (2011). "Plant immunity: evolutionary insights from PBS1, Pto, and RIN4." Plant Signal Behav 6(6): 794-9.

Huang, S., R. Li, et al. (2009). "The genome of the cucumber, Cucumis sativus L." Nat Genet 41(12): 1275-81.

Huang, S., E. A. van der Vossen, et al. (2005). "Comparative genomics enabled the isolation of the R3a late blight resistance gene in potato." Plant J 42(2): 251-61.

Huitema, E., V. G. Vleeshouwers, et al. (2003). "Active defence responses associated with non-host resistance of Arabidopsis thaliana to the oomycete pathogen Phytophthora infestans." Mol Plant Pathol 4(6): 487-500.

Irimia, M., J. L. Rukov, et al. (2007). "Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing." BMC Evol Biol 7: 188.

Irizarry, R. A., B. M. Bolstad, et al. (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Res 31(4): e15.

Islam, S., U. Kjallquist, et al. (2011). "Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq." Genome Res 21(7): 1160-7.

Jiang, R. H. and B. M. Tyler (2012). "Mechanisms and evolution of virulence in oomycetes." Annu Rev Phytopathol 50: 295-318.

Jiao, Y. and E. M. Meyerowitz (2010). "Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control." Mol Syst Biol 6: 419.

Jones, J. D. and J. L. Dangl (2006). "The plant immune system." Nature 444(7117): 323-9.

Kalyna, M., C. G. Simpson, et al. (2011). "Alternative splicing and nonsense-mediated

decay modulate expression of important regulatory genes in Arabidopsis." Nucleic Acids Res 40(6): 2454-69.

Kamoun, S. (2007). "Groovy times: filamentous pathogen effectors revealed." Curr Opin Plant Biol 10(4): 358-65.

Kanetis, L., G. Holmes, et al. (2010). "Survival of Pseudoperonospora cubensis sporangia exposed to solar radiation." Plant Pathol 59(313-323).

Keren, H., G. Lev-Maor, et al. (2010). "Alternative splicing and evolution: diversification, exon definition and function." Nat Rev Genet 11(5): 345-55.

Kim, E., A. Goren, et al. (2008). "Alternative splicing: current perspectives." Bioessays 30(1): 38-47.

Koren, E., G. Lev-Maor, et al. (2007). "The emergence of alternative 3' and 5' splice site exons from constitutive exons." PLoS Comput Biol 3(5): e95.

Kriventseva, E. V., I. Koch, et al. (2003). "Increase of functional diversity by alternative splicing." Trends Genet 19(3): 124-8.

Kunjeti, S. G., T. A. Evans, et al. (2012). "RNA-Seq reveals infection-related global gene changes in Phytophthora phaseoli, the causal agent of lima bean downy mildew." Mol Plant Pathol 13(5): 454-66.

Lam, B. J. and K. J. Hertel (2002). "A general role for splicing enhancers in exon definition." Rna 8(10): 1233-41.

Lamour, K. H., R. Stam, et al. (2012). "The oomycete broad-host-range pathogen Phytophthora capsici." Mol Plant Pathol 13(4): 329-37.

Langmead, B., K. D. Hansen, et al. (2010). "Cloud-scale RNA-sequencing differential expression analysis with Myrna." Genome Biol 11(8): R83.

Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol 10(3): R25.

Lebeda, A., J. Pavelková, et al. (2011). "Distribution, Host Range and Disease Severity of Pseudoperonospora cubensis on Cucurbits in the Czech Republic." Journal of Phytopathology(159): 589–596.

Lebeda, A. and M. P. Widrlechner (2003). "A set of Cucurbitaceae taxa for differentiation of Pseudoperonospora cubensis pathotypes." J. Plant Dis. Prot. 110: 337–349.

Leoni, G., L. Le Pera, et al. (2011). "Coding potential of the products of alternative splicing in human." Genome Biol 12(1): R9.

Levesque, C. A., H. Brouwer, et al. (2010). "Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire." Genome Biol 11(7): R73.

Li, H. and R. Durbin (2010). "Fast and accurate long-read alignment with Burrows-Wheeler transform." Bioinformatics 26(5): 589-95.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics 25(16): 2078-9.

Li, R., C. Yu, et al. (2009). "SOAP2: an improved ultrafast tool for short read alignment." Bioinformatics 25(15): 1966-7.

Lindeberg, M., S. Cartinhour, et al. (2006). "Closing the circle on the discovery of genes encoding Hrp regulon members and type III secretion system effectors in the genomes of three model Pseudomonas syringae strains." Mol Plant Microbe Interact 19(11): 1151-8.

Long, J. C. and J. F. Caceres (2009). "The SR protein family of splicing factors: master regulators of gene expression." Biochem J 417(1): 15-27.

Luna, E., V. Pastor, et al. (2011). "Callose deposition: a multifaceted plant defense response." Mol Plant Microbe Interact 24(2): 183-93.

Mahalingam, R., A. Gomez-Buitrago, et al. (2003). "Characterizing the stress/defense transcriptome of Arabidopsis." Genome Biol 4(3): R20.

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." Mol Plant Pathol 10(6): 721-34.

Maquat, L. E. (2004). "Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics." Nat Rev Mol Cell Biol 5(2): 89-99.

Marioni, J. C., C. E. Mason, et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome Res 18(9): 1509-17.

McCullough, A. J. and S. M. Berget (2000). "An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5' splice sites." Mol Cell Biol 20(24): 9225-35.

McGraw, K. O. and S. P. Wong (1996). "Forming inferences about some intraclass correlation coefficients." Psychological Methods 1(1): 30-46.

McIntyre, L. M., K. K. Lopiano, et al. (2011). "RNA-seq: technical variability and sampling." BMC Genomics 12: 293.

Melamud, E. and J. Moult (2009). "Structural implication of splicing stochastics." Nucleic Acids Res 37(14): 4862-72.

Meric, F., A. M. Searfoss, et al. (1996). "Masking and unmasking maternal mRNA. The role of polyadenylation, transcription, splicing, and nuclear history." J Biol Chem 271(48): 30804-10.

Meyer, M. and J. Vilardell (2009). "The quest for a message: budding yeast, a model organism to study the control of pre-mRNA splicing." Brief Funct Genomic Proteomic 8(1): 60-7.

Minshall, N., G. Thom, et al. (2001). "A conserved role of a DEAD box helicase in mRNA masking." Rna 7(12): 1728-42.

Monaghan, J. and C. Zipfel (2012). "Plant pattern recognition receptor complexes at the plasma membrane." Curr Opin Plant Biol 15(4): 349-57.

Morgan, W. and S. Kamoun (2007). "RXLR effectors of plant pathogenic oomycetes." Current Opinion in Microbiology 10(4): 332.

Nagy, E. and L. E. Maquat (1998). "A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance." Trends Biochem Sci 23(6): 198-9.

Navarro, L., C. Zipfel, 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.

Neufeld, K. and P. S. Ojiambo (2012). "Interactive effectrs of temperature and leaf wetness duration on sporangia germination and infection of cucurbit hosts by Pseuedoperonospora cubensis." Plant Dis.

Nicaise, V., M. Roux, et al. (2009). "Recent Advances in PAMP-Triggered Immunity against Bacteria: Pattern Recognition Receptors Watch over and Raise the Alarm." Plant Physiology 150(4): 1638-1647.

Nicholson, P., H. Yepiskoposyan, et al. (2010). "Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors." Cell Mol Life Sci 67(5): 677-700.

Nielsen, H., J. Engelbrecht, et al. (1997). "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites." Protein Eng 10(1): 1-6.

O'Connell, R. J., M. R. Thon, et al. (2012). "Lifestyle transitions in plant pathogenic Colletotrichum fungi deciphered by genome and transcriptome analyses." Nat Genet 44(9): 1060-5.

Palusa, S. G. and A. S. Reddy (2010). "Extensive coupling of alternative splicing of pre-mRNAs of serine/arginine (SR) genes with nonsense-mediated decay." New Phytol 185(1): 83-9.

Pan, Q., O. Shai, et al. (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing." Nat Genet 40(12): 1413-5.

Pandelova, I., M. F. Betts, 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.

Pickrell, J. K., A. A. Pai, et al. (2010). "Noisy splicing drives mRNA isoform diversity in human cells." PLoS Genet 6(12): e1001236.

Quesada-Ocampo, L. M., L. L. Granke, et al. (2012). "The genetic structure of Pseudoperonospora cubensis populations." Plant Dis 96: 1459-1470.

Rayson, S., L. Arciga-Reyes, et al. (2012). "A role for nonsense-mediated mRNA decay in plants: pathogen responses are induced in Arabidopsis thaliana NMD mutants." PLoS One 7(2): e31917.

Reddy, A. S. (2007). "Alternative splicing of pre-messenger RNAs in plants in the genomic era." Annu Rev Plant Biol 58: 267-94.

Reddy, V. S. and A. S. Reddy (2004). "Proteomics of calcium-signaling components in plants." Phytochemistry 65(12): 1745-76.

Richardson, D. N., M. F. Rogers, et al. (2011). "Comparative analysis of serine/arginine-rich proteins across 27 eukaryotes: insights into sub-family classification and extent of alternative splicing." PLoS One 6(9): e24542.

Richter, J. D. and P. Lasko (2011). "Translational control in oocyte development." Cold Spring Harb Perspect Biol 3(9): a002758.

Robinson, M. D. and G. K. Smyth (2007). "Moderated statistical tests for assessing differences in tag abundance." Bioinformatics 23(21): 2881-7.

Robinson, M. D. and G. K. Smyth (2008). "Small-sample estimation of negative binomial dispersion, with applications to SAGE data." Biostatistics 9(2): 321-32.

Roine, E., W. Wei, 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.

Rosenberg, B. R., C. E. Hamilton, et al. (2011). "Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs." Nat Struct Mol Biol 18(2): 230-6.

Rosenthal., A. Z., E. G. Matson, et al. (2011). "RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture." Isme J 5(7): 1133-42.

Roy, S., M. Kagda, et al. (2013). "Genome-wide Prediction and Functional Validation of Promoter Motifs Regulating Gene Expression in Spore and Infection Stages of Phytophthora infestans." PLoS Pathog 9(3): e1003182.

Salzberg, S. L. (2010). "Recent advances in RNA sequence analysis." F1000 Biol Rep. 2: 64.

Sasidharan, R., T. Nepusz, et al. (2012). "GFam: a platform for automatic annotation of gene families." Nucleic Acids Res 40(19): e152.

Savory, E. A., B. N. Adhikari, et al. (2012). "mRNA-Seq analysis of the Pseudoperonospora cubensis transcriptome during cucumber (Cucumis sativus L.) infection." PLoS One 7(4): e35796.

Savory, E. A., L. L. Granke, et al. (2011). "The cucurbit downy mildew pathogen Pseudoperonospora cubensis." Mol Plant Pathol 12(3): 217-26.

Savory, E. A., C. Zou, et al. (2012). "Alternative splicing of a multi-drug transporter from Pseudoperonospora cubensis generates an RXLR effector protein that elicits a rapid cell death." PLoS One 7(4): e34701.

Schornack, S., E. Huitema, et al. (2009). "Ten things to know about oomycete effectors." Mol Plant Pathol 10(6): 795-803.

Schuster, G., I. Lisitsky, et al. (1999). "Polyadenylation and degradation of mRNA in the chloroplast." Plant Physiol 120(4): 937-44.

Scofield, S. R., C. M. Tobias, et al. (1996). "Molecular Basis of Gene-for-Gene Specificity in Bacterial Speck Disease of Tomato." Science 274(5295): 2063-5.

Severing, E. I., A. D. van Dijk, et al. (2009). "Comparative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome." BMC Genomics 10: 154.

Shanku, A. G., M. A. McPeek, et al. (2013). "Functional Annotation and Comparative Analysis of a Zygopteran Transcriptome." G3 (Bethesda).

Shao, F., C. Golstein, et al. (2003). "Cleavage of Arabidopsis PBS1 by a bacterial type III effector." Science 301(5637): 1230-3.

Shen, D., W. Ye, et al. (2011). "Characterization of intronic structures and alternative splicing in Phytophthora sojae by comparative analysis of expressed sequence tags and genomic sequences." Can J Microbiol 57(2): 84-90.

Shulaev, V., D. J. Sargent, et al. (2011). "The genome of woodland strawberry (Fragaria vesca)." Nat Genet 43(2): 109-16.

Silva, A. L., P. Ribeiro, et al. (2008). "Proximity of the poly(A)-binding protein to a premature termination codon inhibits mammalian nonsense-mediated mRNA decay." Rna 14(3): 563-76.

Simonich, M. T. and R. W. Innes (1995). "A disease resistance gene in Arabidopsis with specificity for the avrPph3 gene of Pseudomonas syringae pv. phaseolicola." Mol Plant Microbe Interact 8(4): 637-40.

Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Stat Appl Genet Mol Biol 3: Article3.

Sohn, K. H., R. Lei, et al. (2007). "The downy mildew effector proteins ATR1 and ATR13 promote disease susceptibility in Arabidopsis thaliana." Plant Cell 19(12): 4077-90.

Stein, L. D., C. Mungall, et al. (2002). "The generic genome browser: a building block for a model organism system database." Genome Res 12(10): 1599-610.

Takken, F. L. and A. Goverse (2012). "How to build a pathogen detector: structural basis of NB-LRR function." Curr Opin Plant Biol 15(4): 375-84.

Team, R. D. C. (2010). "R: A language and environment for statistical computing. R Foundation for Statistical Computing."

Thilmony, R., W. Underwood, et al. (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.

Thomas, W. J., C. A. Thireault, et al. (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.

Tian, M., J. Win, et al. (2011). "454 Genome sequencing of Pseudoperonospora cubensis reveals effector proteins with a QXLR translocation motif." Mol Plant Microbe Interact 24(5): 543-53.

Torto, T. A., S. Li, et al. (2003). "EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen Phytophthora." Genome Res 13(7): 1675-85.

Toung, J. M., M. Morley, et al. (2011). "RNA-sequence analysis of human B-cells." Genome Res 21(6): 991-8.

Trapnell, C., A. Roberts, et al. (2011). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nat Protoc 7(3): 562-78.

Trapnell, C., B. A. Williams, 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.

Truman, W., M. T. de Zabala, et al. (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.

Tsuda, K., M. Sato, et al. (2008). "Interplay between MAMP-triggered and SA-mediated defense responses." Plant J 53(5): 763-75.

Turro, E., S. Y. Su, et al. (2011). "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads." Genome Biol 12(2): R13.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A 98(9): 5116-21.

Tyler, B. M., S. Tripathy, et al. (2006). "Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis." Science 313(5791): 1261-6.

Van Vliet, G. J. A. and W. D. Meysing (1977). "Relation in the inheritance of resistance to Pseudoperonospora cubensis Rost and Sphaerotheca fuliginea Poll. in cucumber (Cucumis sativus L.)." Euphytica 26: 793-796.

Vasudevan, S., E. Seli, et al. (2006). "Metazoan oocyte and early embryo development program: a progression through translation regulatory cascades." Genes Dev 20(2): 138-46.

Vleeshouwers, V. G., S. Raffaele, et al. (2011). "Understanding and exploiting late blight resistance in the age of effectors." Annu Rev Phytopathol 49: 507-31.

Walker, C. A., M. Koppe, et al. (2008). "A putative DEAD-box RNA-helicase is required for normal zoospore development in the late blight pathogen Phytophthora infestans." Fungal Genet Biol 45(6): 954-62.

Wang, B., G. Guo, et al. (2010). "Survey of the transcriptome of Aspergillus oryzae via massively parallel mRNA sequencing." Nucleic Acids Res 38(15): 5075-87.

Wang, K., D. Singh, et al. (2010). "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery." Nucleic Acids Res 38(18): e178.

Wang, L., R. M. Mitra, 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.

Wang, Y., X. Zeng, et al. (2012). "Exploring the switchgrass transcriptome using second-generation sequencing technology." PLoS One 7(3): e34225.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet 10(1): 57-63.

Wettenhall, J. M., K. M. Simpson, et al. (2006). "affylmGUI: a graphical user interface for linear modeling of single channel microarray data." Bioinformatics 22(7): 897-9.

Whisson, S. C., P. C. Boevink, et al. (2007). "A translocation signal for delivery of oomycete effector proteins into host plant cells." Nature 450(7166): 115-8.

Wilhelm, B. T., S. Marguerat, et al. (2008). "Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution." Nature 453(7199): 1239-43.

Win, J., W. Morgan, et al. (2007). "Adaptive Evolution Has Targeted the C-Terminal Domain of the RXLR Effectors of Plant Pathogenic Oomycetes." The Plant Cell Online 19(8): 2349.

Wu, J., Z. Wang, et al. (2013). "The genome of the pear (Pyrus bretschneideri Rehd.)." Genome Res 23(2): 396-408.

Zdobnov, E. M. and R. Apweiler (2001). "InterProScan--an integration platform for the signature-recognition methods in InterPro." Bioinformatics 17(9): 847-8.

Zhang, J., F. Shao, et al. (2007). "A Pseudomonas syringae effector inactivates MAPKs to suppress PAMP-induced immunity in plants." Cell Host Microbe 1(3): 175-85.

Zipfel, C. and G. Felix (2005). "Plants and animals: a different taste for microbes?" Curr Opin Plant Biol 8(4): 353-60.

Zuo, P. and J. L. Manley (1994). "The human splicing factor ASF/SF2 can specifically recognize pre-mRNA 5' splice sites." Proc Natl Acad Sci U S A 91(8): 3363-7.

APPENDIX

Appendix I: AutoSPOTs: Automated Image Analysis for Enumerating Callose Deposition

Jason S. Cumbie, Rebecca C. Pankow, William J. Thomas, and Jeff H. Chang

In *Genome-Enabled Analysis of Plant-Pathogen Interactions*
pp. 233-242

INTRODUCTION

Computational methods are essential to any genomicist's toolkit. With the continual advances in sequencing technology, there are demands for computational approaches that can keep pace with the different data structures. It is with these in mind that we have developed software programs to further enable integration of genomics with plant-pathogen research. In this chapter, we describe AutoSPOTs, one of the programs that we developed to facilitate high-throughput characterization of bacteria-plant interactions.

The type III secretion system (T3SS) is used by many Gram-negative bacteria to establish interactions with their hosts (Grant et al., 2006). The T3SS is a conduit that deploys bacterial encoded type III effector proteins directly into host cells where they function to manipulate the host for the benefit of the infecting bacterium. In the case of plant pathogenic bacteria, type III effectors are necessary to engage and dampen one layer of plant defense called PAMP-triggered immunity (PTI; Jones and Dangl, 2006). A number of events have been associated with PTI, including the deposition of callose in cell walls (Zipfel, 2009). Callose, a β-1,3 linked glucan, along with cellulose, pectin, lignin, and hydroxyproline-rich proteins, are deposited as an agglomeration believed to function as an apposition to infecting bacteria located in the apoplastic space and to other penetrating-type microbes (Bestwick et al., 1995; Bestwick et al., 1998).

Pseudomonas syringae is an excellent model pathogen of plants. The genome sequences for several strains of *P. syringae* have been completed and mined for candidate type III effector genes (Buell et al., 2003; Feil et al., 2005; Joardar et al., 2005; Almeida

et al., 2009; Reinhardt et al., 2009; Studholme et al., 2009). Functional approaches that relied on the availability of the genome sequence have also been used (Chang et al., 2005). One strain in particular, *P. syringae* pv *tomato* race DC3000 (*Pto*DC3000), is intensively studied because of its ability to infect the model host plant, *Arabidopsis thaliana*. *Pto*DC3000 has approximately 30 type III effector genes (Schechter et al., 2006). The challenge now is to understand the functions of all type III effector proteins and how a system of deployed type III effectors is coordinated in the host cell to dampen PTI for the benefit of the infecting bacterium.

AutoSPOTs – for automated batch enumeration of callose deposition

Enumerating the deposition of callose is an often-used assay for quantifying PTI and perturbations to PTI. The wet-lab manipulations for this assay are relatively straightforward. The robustness of the assay, however, is affected by the variable host response to pathogen challenge and the obvious solution is to simply increase the number of samples. But, this simple solution is often outweighed by the onerous nature of the callose assay and its analyses.

We have therefore developed AutoSPOTs to mitigate the labor-intensive steps associated with image analyses and their potential associated biases. With user-defined criteria based on size and color, AutoSPOTs automates and batch enumerates aniline-stained callose deposits from JPEG images. AutoSPOTs will also automatically execute a series of standard statistical analyses. We have used AutoSPOTs to analyze thousands of images on a laptop computer. AutoSPOTs is an open-source Graphical User Interface

(GUI) written in Perl and C. The software program and user's manual can be downloaded from our website at: http://changlab.cgrb.oregonstate.edu/.

Requirements for AutoSPOTs

Methods for sample preparation have been described (Kim and Mackey, 2008). Yet, some simple steps taken during sample preparation and microscopy can greatly improve the quality of the images for more accurate identification and enumeration of callose deposits. It is important to clear leaves as completely as possible subsequent to sample collection because autofluorescence of the chlorophyll will lead to background fluorescence. Insufficient staining can result in weakly fluorescent callose deposits. We have found that the simple act of staining leaves in aniline blue overnight improves the clarity of callose fluorescence. Proper mounting of leaves is another crucial step in sample preparation; wrinkling of leaves or bubbles in the mounting medium can result in multiple focal planes in a single field of view, making resolution of the entire field difficult. Finally, it is important to use an appropriate exposure time for capturing high-quality images (this may require some trial and error). While the customizable color filter settings make AutoSPOTs functional over a range of exposures, extremes in exposures pose potential problems. Exposure settings that are too low will result in faint or dim callose deposits whereas exposure settings that are too high will wash out fluorescent spots. Both result in a reduction in the accuracy of AutoSPOTs.

We typically take ten JPEG images per leaf and sample fifteen leaves per treatment. A minimum of two treatments is required. For fully automated batch analysis, AutoSPOTs requires the user to properly name and store JPEG images in a recognizable

manner. The two recognized formats are as single numbers (e.g., sample1.jpg, sample2.jpg), or as number-number (e.g., treatment1-1, treatment1-2). Additionally, there must be the same number of JPEG images per sample (leaf) per treatment group. JPEG images should be saved in directories labeled according to treatment groups. If these conditions are not met then some of the automated functions of AutoSPOTs cannot be used.

Defining filters

AutoSPOTs requires the user to define a size filter and one of two types of color filters. In a subsequent section of this chapter, we show the effects that different color filters have on results. AutoSPOTs will apply the filters on a pixel-by-pixel basis to identify callose deposits for each JPEG image to be analyzed. It is therefore important for the user to capture high quality JPEG images and to establish the proper filter settings for the most uniform, sensitive and accurate identification of aniline-blue stained callose deposits across an experiment.

For the size filter, we recommend starting with minimum and maximum sizes of 20 and 100, respectively, and to refine as needed (see discussion on previewing below). We have included two types of color filters: the RGB and ratio filters. To simplify selection, we have included a ‘color selection assistance’ feature. By selecting pixels of callose deposits from several representative images, the color selection assistance feature will provide the user with the values for each of the criteria required of the RGB or ratio filters. Other criteria include ‘Trip’ and ‘Drop’ thresholds. The former is used by AutoSPOTs to determine which pixels will be considered as part of a stained callose

deposit and ‘trips’ AutoSPOTs into expanding a callose deposit. The latter is used by AutoSPOTs to exclude pixels from a stained callose deposit and forces AutoSPOTs to ‘drop’ the pixel from expanding the callose deposit. The user can then determine the average values from multiple pixels of multiple images and set the color filters accordingly.

In most cases, AutoSPOTs performs better with grayscale JPEG images; this may depend on the camera and staining/de-staining of leaves. We have added a feature that enables all images to be automatically converted to grayscale. When defining the color filter, note that the red, green, and blue channels will have the same value so the ratio filter cannot be used. In contrast, the RGB filter must be used and simply becomes an RGB intensity filter.

AutoSPOTs allows the user to preview the sensitivity and accuracy of the filters. A screenshot of a preview and the GUI is presented (Fig. 1). The image will be displayed and each identified callose deposit will be demarked. The total number of callose deposits identified will also be displayed. It is strongly recommended that the user carefully examine several images and adjust the filter settings to find the desired level of sensitivity and accuracy. It is important to preview images with few and many callose deposits (see Fig. 3). We caution the user to pay close attention to identification of leaf features such as veins or trichomes as callose deposits as well as incomplete demarcation or over-extension of callose deposits. Incorrect identification of leaf features as callose deposits suggests the filters are too sensitive, whereas inaccurate demarcation of callose suggests the Trip and Drop distances are not correctly set.

It cannot be stressed enough that the successful use of AutoSPOTS will depend on consistent, high-quality images, control treatments to assess the accuracy of filters, application of filters uniformly on all samples of all treatments being compared, and a sufficient number of JPEG images and samples to obtain good statistical power for analysis. Not all callose deposits will be identified, especially those in different focal planes, but as along as all leaves were prepared in a similar manner and JPEG images were photographed under similar settings, there will not be any biases in the results.

We have provided a detailed step-by-step Users Manual available by download from our website.

Image analysis

Once the user has identified a satisfactory filter setting, AutoSPOTs can automatically batch process all images. Analysis begins by examining each pixel of each image individually to identify those that pass the 'Trip' threshold for a color filter. Once the pixels that pass the 'Trip' threshold are located, all adjoining pixels are analyzed using a 'Drop' threshold, which is usually a more relaxed threshold allowing for spot fading near the edges. Pixels are then continually counted outward until no more adjoining pixels can be found that match the 'Drop' threshold criteria. The number of pixels in a given 'spot' is tallied, and then analyzed using the size threshold. Those that are within the minimum and maximum values set by the user are counted as a single callose deposit.

AutoSPOTs calculates the average number of callose deposits by averaging per JPEG image per leaf per treatment. AutoSPOTS has built-in statistical analysis tools and

will generate a statistical report for all treatments against the user-defined control treatment. AutoSPOTs will also plot the data for visual representation. At each step of analysis all the data is saved to text files and directories specified by the user. Copies of every image analyzed with their demarked callose deposits are also stored so the user can inspect the sensitivity and accuracy of the filters.

Demonstration of AutoSPOTS

We used one size filter setting and six different color filter settings in AutoSPOTs to demonstrate their effects on sensitivity and accuracy in enumerating callose deposits from JPEG images (Fig. 2). Four of the tested color filter settings used RGB (intensity) values to identify callose deposits from JPEG images that were converted to grayscale. In these cases, the color filter setting was set from least sensitive to overly sensitive by using different values – we noted the drop and trip values had the largest effect on sensitivity. Two of the color filter settings used a ratio or RGB filter to analyze the original color JPEG images.

The treatments we tested were Arabidopsis infected with *Pto*DC3000, a T3SS-deficient mutant of *Pto*DC3000 (*hrcC*), a soil bacterium with an integrated T3SS-encoding region (EtHAn), and EtHAn carrying the type III effector gene, *hopM1*. *Pto*DC3000 deploys 30 type III effector proteins into Arabidopsis and sufficiently dampens PTI to cause disease. Its ability to dampen the deposition of callose has been repeatedly demonstrated (Hauck et al., 2003; DebRoy et al., 2004; Nomura et al., 2006; Ham et al., 2007). In contrast, since the *hrcC* mutant is incapable of delivering type III effectors, it cannot dampen the deposition of callose or PTI, nor cause disease on

Arabidopsis (Niepold et al., 1985; Lindgren et al., 1986; Roine et al., 1997; Hauck et al., 2003; Thilmony et al., 2006). EtHAn was engineered from *P. fluorescens* Pf0-1 and is devoid of any endogenous type III effectors (Thomas et al., 2009). EtHAn therefore elicits PTI. The type III effector, HopM1, is sufficient to dampen the deposition of callose (DebRoy et al., 2004; Nomura et al., 2006; Thomas et al., 2009). A total of fifteen leaves were challenged per treatment, and ten images were randomly taken from each leaf. AutoSPOTs took less than 45 minutes to automatically analyze the 600 JPEG images.

In general., the trends were similar for each of the six filter settings (Fig. 2). However, when the automatically generated statistics were analyzed, it is clear that the filter settings do indeed affect interpretation of data. Based on previous findings, we expected significant differences between *Pto*DC3000 versus the *hrcC* mutant and EtHAn + *hopM1* versus EtHAn treatments. The color filter settings 1-3 resulted in no differences in the conclusions – both comparisons within each of the three settings were statistically significant. However, the color filter setting 1 was clearly the poorest of the three in terms of sensitivity. In contrast, increasing the sensitivity of grayscale analysis (setting 4) or use of color JPEG images (setting 5 and 6) resulted in less desirable results. Thus, increased sensitivity to identify the highest number of callose deposits is not necessarily the most recommended approach.

We visually examined the analyzed JPEG images to understand the results of the different color filter settings (Fig. 3). In general., most of the color settings performed fairly well in analyzing areas with few callose deposits. Color filter settings 2, 3 and 6

were the more accurate. In contrast, the different color filter settings resulted in dramatic differences in the analysis of areas in which callose deposits were abundant. Settings 2 and 3 performed fairly well. However, very few callose deposits were identified in JPEG images with dense staining spots when AutoSPOTs used color filter settings 4 - 6. This was a consequence of AutoSPOTs failing to drop pixels and categorizing several callose deposits as one larger spot. These large spots would exceed the maximum of 100 as defined by the size filter and not be counted. Changing the size filter could potentially alleviate this problem to a certain extent. We have analyzed JPEG images provided by another research group and results from analysis of the color images were superior to grayscale images. This could be a consequence of differences in staining/de-staining of leaves or in the microscope camera. It is recommended to try different combinations of filters.

The differences in performance when analyzing JPEG images with sparse and dense callose deposits can lead to very misleading results. For example, we could not detect a significant difference between treatments with *Pto*DC3000 and its *hrcC* mutant under color filter setting numbers 4 and 5. This is because AutoSPOTs was sufficiently accurate in identifying the sparse callose deposits resulting from infection with *Pto*DC3000 but was inadequate in identifying densely distributed callose deposits resulting from infection with the *hrcC* mutant.

CONCLUSION

We developed AutoSPOTs a simple, user-friendly, and open-source software program to facilitate the high-throughput analysis of JPEG images. AutoSPOTs mitigates

labor-intensive data analysis by automating and batch analyzing large sets of JPEG images for callose deposits and comparing results between treatments. AutoSPOTs therefore provides the opportunity to examine larger numbers of type III effectors or host genetic backgrounds for their effects on PTI.

We purposefully developed AutoSPOTs to be a simple program. As a consequence, the filtering scheme that AutoSPOTs uses relies on the user to identify the most suitable combination of filters through careful visual examination of their JPEG images. It is therefore expected that the user will design a properly controlled experiment and capture high-quality and uniform JPEG images for analysis.

AutoSPOTs was developed for identification and enumeration of aniline-stained callose deposits but it has potential uses in other applications in studying plant-pathogen interactions, such as enumerating GFP-expressing bacteria.

ACKNOWLEDGEMENTS

We thank Caitlin A. Thireault, Allison Smith, and Philip Hillebrand for their assistance. We gratefully acknowledge Jim Carrington for use of his light microscope. This research was supported in part by start-up funds from Oregon State University to JHC and by the National Research Initiative Competitive Grant no. 2008-35600-18783 from the USDA National Institute of Food and Agriculture, Microbial Functional Genomics Program. JSC was supported by a Computational and Genome Biology Initiative Fellowship from Oregon State University.

REFERENCES CITED

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.

Bestwick, C.S., Bennett, M.H., and Mansfield, J.W. 1995. *Hrp* mutant of *Pseudomonas syringae* pv *phaseolicola* induces cell wall alterations but not membrane damage leading to the hypersensitive reaction in lettuce. Plant Physiol. 108:503-516.

Bestwick, C.S., Brown, I.R., and Mansfield, J.W. 1998. Localized changes in peroxidase activity accompany hydrogen peroxide generation during the development of a nonhost hypersensitive reaction in lettuce. Plant Physiol. 118:1067-1078.

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., Madupu, R., Daugherty, S., Brinkac, L., Beanan, M.J., Haft, D.H., Nelson, W.C., Davidsen, T., Zafar, N., Zhou, L., Liu, J., Yuan, Q., Khouri, H., Fedorova, N., Tran, B., Russell, D., Berry, K., Utterback, T., Van Aken, S.E., Feldblyum, T.V., D'Ascenzo, M., Deng, W.L., Ramos, A.R., Alfano, J.R., Cartinhour, S., Chatterjee, A.K., Delaney, T.P., Lazarowitz, S.G., Martin, G.B., Schneider, D.J., Tang, X., Bender, C.L., White, O., Fraser, C.M., and Collmer, A. 2003. The complete genome sequence of the Arabidopsis and tomato pathogen *Pseudomonas syringae* pv. *tomato* DC3000. Proc. Natl. Acad. Sci. U S A 100:10181-10186.

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. U S A 102:2549-2554.

DebRoy, S., Thilmony, R., Kwack, Y.B., Nomura, K., and He, S.Y. 2004. A family of conserved bacterial effectors inhibits salicylic acid-mediated basal immunity and promotes disease necrosis in plants. Proc. Natl. Acad. Sci. U S A 101:9927-9932.

Feil, H., Feil, W.S., Chain, P., Larimer, F., DiBartolo, G., Copeland, A., Lykidis, A., Trong, S., Nolan, M., Goltsman, E., Thiel, J., Malfatti, S., Loper, J.E., Lapidus, A., Detter, J.C., Land, M., Richardson, P.M., Kyrpides, N.C., Ivanova, N., and Lindow, S.E. 2005. Comparison of the complete genome sequences of *Pseudomonas syringae* pv. *syringae* B728a and pv. *tomato* DC3000. Proc. Natl. Acad. Sci. U S A 102:11064-11069.

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. Ann. Rev. Microbiol. 60:425-449.

Ham, J.H., Kim, M.G., Lee, S.Y., and Mackey, D. 2007. Layered basal defenses underlie non-host resistance of Arabidopsis to *Pseudomonas syringae* pv. *phaseolicola*. Plant J. 51:604-616.

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. U S A 100:8577-8582.

Joardar, V., Lindeberg, M., Jackson, R.W., Selengut, J., Dodson, R., Brinkac, L.M., Daugherty, S.C., Deboy, R., Durkin, A.S., Giglio, M.G., Madupu, R., Nelson, W.C., Rosovitz, M.J., Sullivan, S., Crabtree, J., Creasy, T., Davidsen, T., Haft, D.H., Zafar, N., Zhou, L., Halpin, R., Holley, T., Khouri, H., Feldblyum, T., White, O., Fraser, C.M., Chatterjee, A.K., Cartinhour, S., Schneider, D.J., Mansfield, J., Collmer, A., and Buell, C.R. 2005. Whole-genome sequence analysis of *Pseudomonas syringae* pv. *phaseolicola* 1448A reveals divergence among pathovars in genes involved in virulence and transposition. J. Bacteriol. 187:6488-6498.

Jones, J.D., and Dangl, J.L. 2006. The plant immune system. Nature 444:323-329.

Kim, M.G., and Mackey, D. 2008. Measuring cell-wall-based defenses and their effect on bacterial growth in Arabidopsis. Methods Mol. Biol. 415:443-452.

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.

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.

Reinhardt, J.A., Baltrus, D.A., Nishimura, M.T., Jeck, W.R., Jones, C.D., and Dangl, J.L. 2009. De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. Genome Res. 19:294-305.

Roine, E., Wei, W., Yuan, J., Nurmiaho-Lassila, E.L., Kalkkinen, N., Romantschuk, M., and He, S.Y. 1997. Hrp pilus: an hrp-dependent bacterial surface appendage produced

by *Pseudomonas syringae* pv. *tomato* DC3000. Proc. Natl. Acad. Sci. U S A 94:3459-3464.

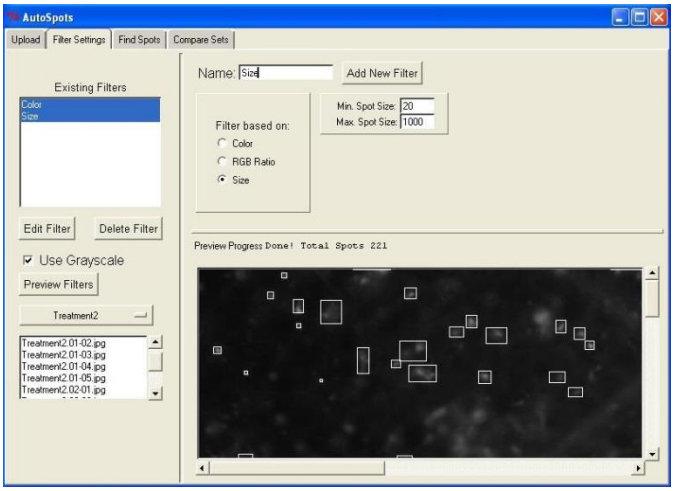
Schechter, L.M., Vencato, M., Jordan, K.L., Schneider, S.E., Schneider, D.J., and Collmer, A. 2006. Multiple approaches to a complete inventory of *Pseudomonas syringae* pv. *tomato* DC3000 type III secretion system effector proteins. Mol. Plant-Microbe Interact. 19:1180-1192.

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.

Thilmony, R., Underwood, W., and He, S.Y. 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. 46:34-53.

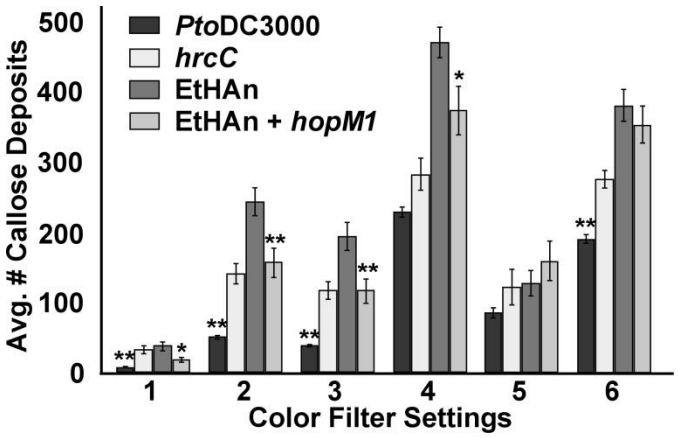
Thomas, W.J., Thireault, C.A., Kimbrel, J.A., and Chang, J.H. 2009. Recombineering and stable integration of the *Pseudomonas syringae* pv. *syringae* 61 *hrp/hrc* cluster into the genome of the soil bacterium *Pseudomonas fluorescens* Pf0-1. Plant J. 60:919-928.

Zipfel, C. 2009. Early molecular events in PAMP-triggered immunity. Curr. Opin. Plant Biol. 12:414-420.



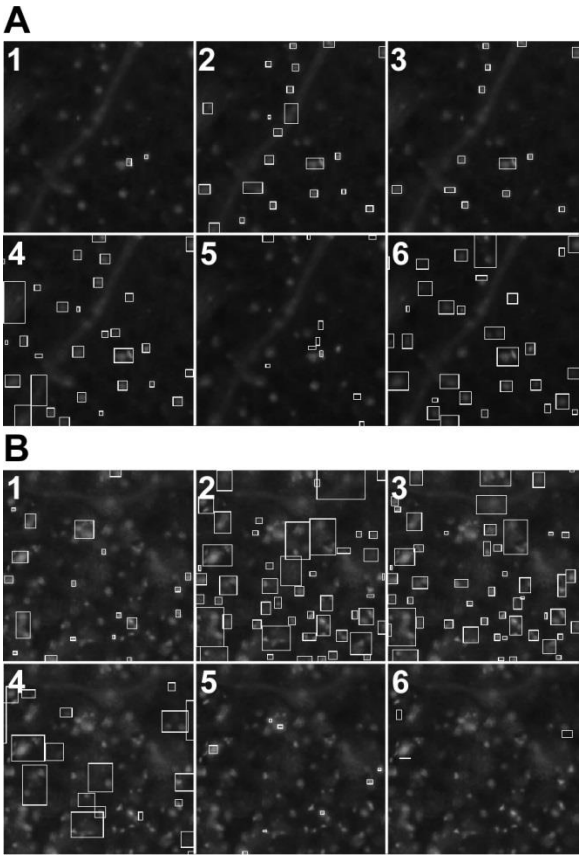
Appendix I, Figure 1. Screenshot of the Graphical User Interface of AutoSPOTs.

The AutoSPOTs GUI divides its various functions into four tabs. This screenshot of the Filter Settings tab illustrates the Preview Filters functions. Filters are defined and added in the top right section of the tab (settings for the Size filter are shown here). The desired filters are then selected from the Existing Filters menu (note that both color and size filters must be selected). The Use Grayscale option has been selected. Once a set of images has been loaded and an image selected in the lower left portion of the screen, the selected image will be displayed in the lower right display window. Callose deposits identified by the Preview Filters function will be indicated with a box and the total number of deposits identified will be displayed above the JPEG image.



Appendix I, Figure 2. Enumeration of callose deposits by AutoSPOTs using different color filter settings.

We infected leaves of Arabidopsis with four different strains of bacteria. We used AutoSPOTs to identify and quantify callose deposits with six different color filter settings. For filters 1 – 4, JPEG images were converted to grayscale. The RGB, Trip, and Drop values respectively, were 130, 40, and 100 for color filter 1; 100, 80, and 100 for filter 2; 90, 50, and 80 for filter 3; and 90, 100, and 100 for color filter 4. For color filters 5 and 6, JPEG images were analyzed as color images using the color ratio and RGB filters, respectively. Fifteen leaves were infected per treatment and ten images were taken per leaf. Standard errors are shown. For each color filter setting, we compared results of *Pto*DC300 versus the *hrcC* mutant and EtHAn + *hopM1* versus EtHAn. Significant differences are denoted; *p-value $\leq$ 0.05; **p-value $\leq$ 0.01.



Appendix I, Figure 3. Effects of different color filter settings on the accuracy of AutoSPOTs.

Each set of panels represents the same section of the same two JPEG images analyzed using the six different color filter settings (1-6) described previously. One JPEG image had few callose deposits (A) while the other image was dense with callose deposits (B). Callose deposits identified by the program are indicated with a box. The analyses for color filters 5 and 6 used color images, which we have converted to grayscale for publication purposes.

Appendix II: 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 (Fig. 1).

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.

Conflict of Interest

The authors declare no conflict of interest.

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.

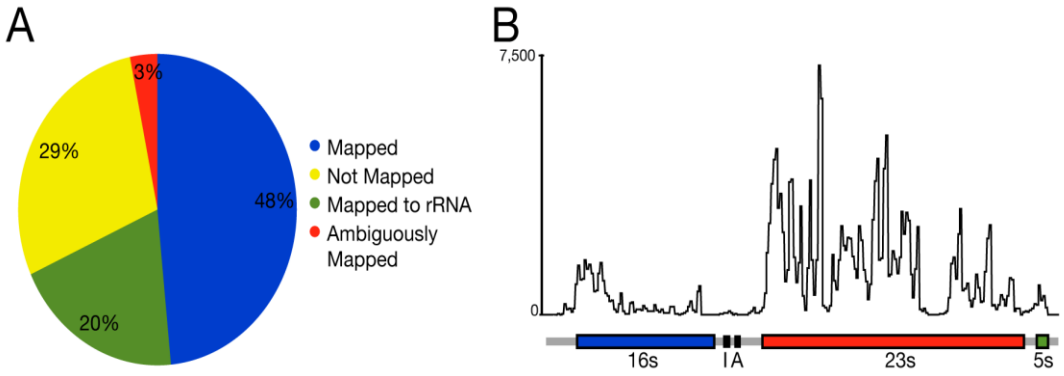
70. Poroyko, V.; White, J.R.; Wang, M.; Donovan, S.; Alverdy, J.; Liu, D.C.; Morowitz, M.J. Gut microbial gene expression in mother-fed and formula-fed piglets. *PLoS One* **2010**, *5*, doi:10.1371/journal.pone.0012459.

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.

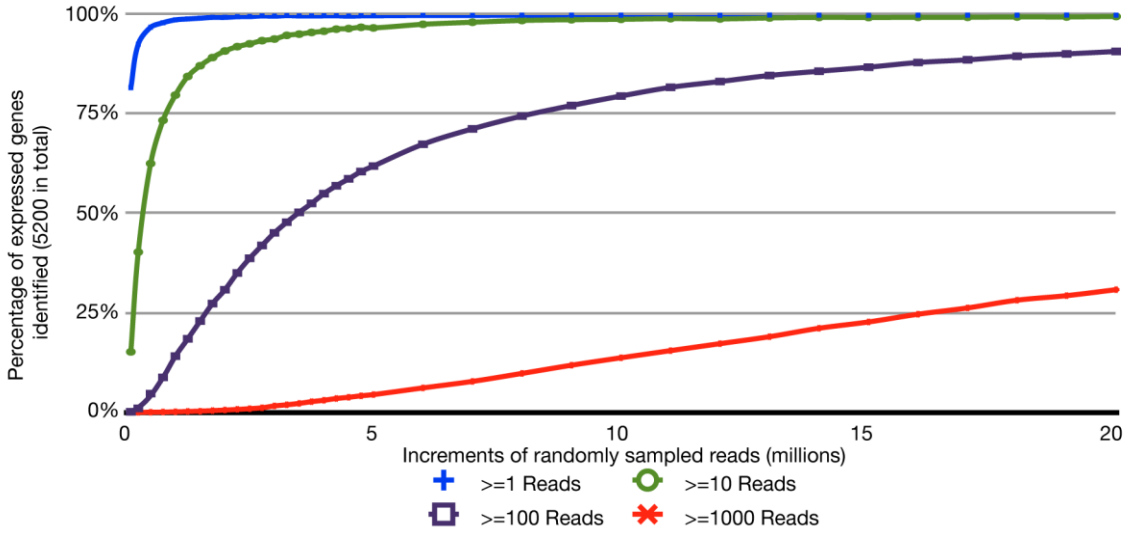
73. Kang, Y.; Norris, M.H.; Zarzycki-Siek, J.; Nierman, W.C.; Donachie, S.P.; Hoang, T.T. Transcript amplification from single bacterium for transcriptome analysis. *Genome Res*. **2011**, *21*, 925–935.

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.



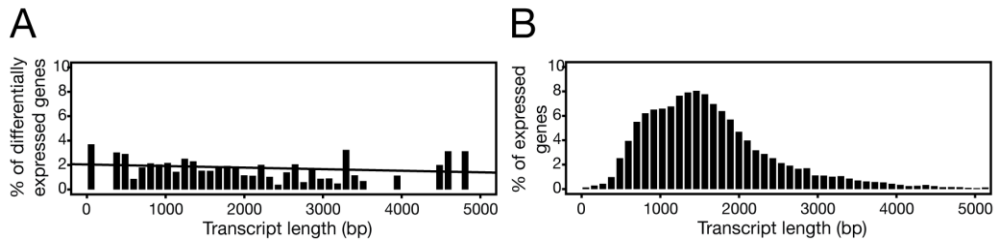
Appendix II, Figure 1. Categorization of RNA-Seq reads.

(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 II, 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 II, 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 II, Table 1. Potential effect of relative frequency on differential expression.

Gene name	Relative frequency					
	Sample1.1 *	**Sample1.2 ***	**Sample1.3 ***	**Sample2.1 †**	**Sample2.2 †**	**Sample2.3 †**
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.