Comparative Transcriptomics of Grapevine Powdery Mildew

Christopher Wayne Snyder
Rochester Institute of Technology

Master’s Thesis

Completed as requisite for a Master’s of Science in Bioinformatics

Submitted May 8th, 2018
To: Head, Thomas H. Gosnell School of Life Sciences

The undersigned state that Christopher Snyder, a candidate for the Master of Science degree in Bioinformatics, has submitted his thesis and has satisfactorily defended it.

This completes the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology.

Thesis committee members:

<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael V. Osier, Ph.D.</td>
<td></td>
</tr>
<tr>
<td>Thesis Advisor</td>
<td></td>
</tr>
<tr>
<td>Lance Cadle-Davidson, Ph.D.</td>
<td></td>
</tr>
<tr>
<td>Gary R. Skuse, Ph.D.</td>
<td></td>
</tr>
<tr>
<td>Larry Buckely, Ph.D.</td>
<td></td>
</tr>
</tbody>
</table>
Table of Contents

Acknowledgements........................................................................................................................................ 3

Abstract...................................................................................................................................................... 4

INTRODUCTION........................................................................................................................................... 6

METHODS.................................................................................................................................................. 15

RESULTS................................................................................................................................................... 20

DISCUSSION............................................................................................................................................... 33

Conclusion................................................................................................................................................ 37

References................................................................................................................................................ 38

Works Cited .............................................................................................................................................. 38
Acknowledgements

I would like to thank first and foremost my advisor, Dr. Michael Osier for his patience throughout the completion of this project. I would also like to thank Dr. Lance Cadle Davidson for his support and guidance in regard to navigating the subject matter. I would also like to thank the remaining members of my Thesis committee, Dr. Gary Skuse and Dr. Larry Buckley. Last but not least, I want to thank my parents, whose continuous support have kindled and sustained my curiosity for all things scientific and otherwise.
Abstract

Grapevine Powdery Mildew is caused by the obligately biotrophic fungus *Erysiphe* (syn. *Uncinula*) necator. This plant disease leads to severe crop loss and subsequent economic burden for the wine and juice industry. While the life cycle of the fungus has been explained, the mechanisms underlying how the fungus surpasses the grape’s defenses are not fully understood. In order to identify suitable targets for gene silencing approaches, a better understanding of the Powdery Mildew transcriptome is necessary. *De novo* assembled contigs of *E. necator* G14 mRNA were compared to the published *E. necator* C-strain transcriptome to determine the areas of overlap between them. These were also aligned to the *E. necator* C-strain genome scaffold and annotated using both the NCBI non-redundant protein database as well as the Eggnog Orthologous group database. Evidence for novel transcription sites was found in the G14 transcriptome and these novel annotations were stored in a gff file for future use. The detection of previously unreported transcripts highlights the need for RNA sequencing approaches that can detect low copy number transcripts.
## Key terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcript</td>
<td>A sequence of mRNA which can be translated into a functioning protein</td>
</tr>
<tr>
<td>Contig</td>
<td>A single sequence of nucleotides returned from an assembly</td>
</tr>
<tr>
<td>Scaffold</td>
<td>A single string of nucleotides representing a portion of the Powdery mildew genome</td>
</tr>
</tbody>
</table>
INTRODUCTION

The purpose of this project is to explore and improve upon existing transcriptomic resources for *Erysiphe necator* (syn *Uncinula necator*). This was achieved by reviewing and comparing two Transcriptomes generated using high throughput RNA sequencing. *Erysiphe necator*, is an obligate biotrophic fungus that infects grape plants worldwide and presents a challenge to grape producers. Commonly referred to as Powdery Mildew, the fungus is the most prevalent grape pathogen in the world and has presented a challenge to European Grape growers since its introduction to the continent in 1847 (Campbel, 2006). *E. necator* grows primarily on the green tissue of the grapevines and has the potential to spread to the berries in the 2-3 weeks following bloom. The macroscopic signature of Powdery mildew can be seen in Figure 1 where the fungus is seen growing on grape berries. The fungus can decrease both the total yield of a grape producing vine as well as affect the taste of the consequent wines, juices, and fruit.

![Figure 1: Photograph of Grapevine Powdery Mildew (Erysiphe necator) growing on grapes (Vitis vinifera) Image source: (Nelson 2015)](image)

Grape growers historically control *E. necator* using fungicides. These fungicides typically function through single site action. This means that they target and inactivate a specific gene product that the mildew needs in order to live and grow. Single site action fungicides possess a weakness in that *E. necator* is capable of developing resistance to them over time (Ypema et al., 1997). The increased resistance of powdery mildew to commercial pesticides has garnered incentive for the development of alternative methods of powdery mildew control. English-Loeb et.al. discusses the possibility of using Tydeid mites as a natural strategy for combating Powdery
mildew. Tydeid mites consume the mycelium of the powdery mildew thus suppressing the rate of infection. While this strategy is effective in theory, it has been difficult to implement on a large scale due to a lack of pesticides that can remove harmful insects while preserving the beneficial mites (English-Loeb 1999).

In an effort to find a new and reliable method to combat Powdery mildew, researchers have turned their attention to the defense mechanisms employed by the grapevines themselves. Not all grape varieties are equally susceptible to *E. necator*. European (*Vitis vinifera*) varieties display far weaker resistance to the fungus than their North American and Asian cousins. In (Ramming et al, 2010) researchers discovered a locus on the genome of the host grapevine that allowed the plant to resist infection by *E. necator*. The authors named the locus “resistance to *Erysiphe necator* 4” (*Ren4*). They discovered the locus by crossing a PM resistant cultivar of *V. romanetii* with PM susceptible *V. vinifera* and performing segregation analysis on the hybrid progeny. The locus conferred resistance according to a dominant transmission pattern (Ramming et al. 2010). While *Ren4* and other resistance loci provide clues as to what strategies the grapevine employs as a resistance mechanism for powdery mildew, the exact molecular mechanism remains unknown. Answering this question involves the study of the underlying molecular biology of Powdery mildew infection from an evolutionary perspective.

**The Finger Lakes Region**

The impact of powdery mildew can be felt close to home, and one of the two transcriptomes described later is from isolate G14, collected from the Finger Lakes wine region of Upstate New York in the United States. This region is of particular interest due to its lower temperatures and higher humidity levels, which may affect pathogen biology. Further, there have been increased plantings in the Finger Lakes of *V. vinifera* European varieties, that are more susceptible to powdery mildew than the North American inter-specific hybrid varieties that have been prevalent in the region since the 1800s (Newman 1986). As of 2011, 22% of the grapevines in the Finger Lakes Wine region consist of *V. vinifera*. The introduction of these European varieties is chiefly by demand for dryer, European style wines (Yeh et al., 2014). Increased cultivation of more susceptible grapevine varieties into the region may have an impact on the epidemiology of
powdery mildew in the region. The threat of Powdery mildew in the Finger Lakes region is of particular economic concern due to the significant up-front costs of growing wine grapes.

**Evolutionary perspective**

The relationship between fungi and plants can take many forms. The premiere role of fungus in many environments is as a decomposer, consuming dead plant matter and allowing the nutrients to return to the soil. Some species of fungi form symbiotic relationships with plants by coating the plant’s roots and absorbing nutrients that the plant cannot process itself and passing them along to the plant in return for a steady supply of carbohydrates (Frank 1885). As discussed above, the relationship between grapevine and powdery mildew is that of a host and a parasite. Powdery Mildew is an obligate biotrophic fungus because it requires a living host in order to grow and reproduce. Due to the biotrophic nature of the fungus, it does not immediately kill the grapevines, instead growing on them long enough to reproduce and disperse. *E. necator* belongs to the division of fungi known as Ascomycota. This division is characterized by elongated spore bearing cells called asci. As a member of this monophyletic group, *E. necator* is capable of reproducing sexually as well as asexually (Gadoury 2012). Powdery mildew spreads through airborne spores. Spores produced via sexual reproduction are ascospores while spores produced asexually are conidia.

The asexual aspect of *E. necator* reproduction involves the formation of conidiophores as offshoots from the mycelium. These structures support the formation of conidia. Asexual conidia form through a process of mitosis and each conidium is genetically identical to the parent fungus. Alternatively, ascospores form through sexual reproduction. If two powdery mildew fungi of different mating types come in contact, they will form chasmothecia. These chasmothecia house specialized cells that may undergo meiosis to produce sexual ascospores. Asexual reproduction requires less energy and therefore is more frequent in favorable conditions. Chasmothecia form, if they form at all, near the end of the growing season or when disease pressures are high enough to increase the odds that colonies of opposite mating types converge. Unlike the other structures of *E. necator* that are hyaline or lacking color, the chasmothecia are pigmented. As they mature they turn from white, to yellow, to darkening shades of brown, finally arriving at black. While it
only takes four weeks for a chasmothecium to reach structural maturity, it takes months for it to reach physiological maturity (Gadoury 2012).

After germination of an ascospore or of an asexual conidium, powdery mildew forms a structure known as an appressorium. This appressorium is capable of penetrating the cell wall of the host plant via a penetration peg to form a haustorium within the host epidermal cell. The haustorial feeding cell supports secondary hyphal growth on top of the plant surface and formation of secondary appressoria and haustoria, as well as conidiophores and conidia for secondary spread of the disease. The haustorium is the only part of *E. necator* that is endophytic or inside the host plant. It is this portion of the fungus that must contend with the immune defenses of the host plant (Gadoury 2012).

The first line of defense in the host plant is its ability to recognize pathogen associated molecular profiles (PAMP). A common PAMP in fungal infections is chitin from the fungal cell wall. The plant’s reaction in this case is PAMP-triggered immunity (PTI). In response to this, fungi have evolved effector proteins that suppress PTI. The host plant responds to this through the evolution of resistance genes that recognize and suppress the effector proteins, as well as trigger immune responses known as effector triggered immunity (ETI). This dynamic leads to an evolutionary arms race between the host and the parasite based around recognition and suppression. This arms race generates selective pressure on both sides to generate the widest variety of effector/resistance proteins possible in order to suppress the proteins of the opponent. A symmetric pattern has emerged in many host parasite systems where each effector protein in the biotrophic fungus will have a counterpart resistance protein in the host. This counterpart is the effector protein’s cognate resistance protein. The haustorium of the powdery mildew secretes the effector proteins during infection.

In conjunction with the effector - resistance protein arms race, other classes of host genes also play a role in resistance to powdery mildew. The Mildew Resistance Locus O (MLO) has been the subject of multiple studies in a range of host plants. These studies have shown that a functioning allele of MLO must be present in order for an adapted powdery mildew species to be capable of successful infection. The consistency of MLO activity across multiple host species is
indicative that an MLO orthologue must be present in a functional form in *V. vinifera* in order for powdery mildew infection to be successful (Qiu 2015). While there exists a general understanding of the molecular biology behind PM infection, there is still a great deal to be discovered concerning the molecular mechanisms that are activated during infection.

The evolutionary arms race that characterizes the host parasite relationship between *V. vinifera* and *E. necator* indicates that the transcription patterns in *V. vinifera* and *E. necator* are interdependent. Fung et al. characterized the transcriptomic effects of powdery mildew infection on various strains of grapevine in order to gain insight into the various defense mechanisms at play. According to this experiment, a nonresistant *V. vinifera* strain exhibited stronger and more widespread shifts in gene expression as a response to powdery mildew inoculation than a resistant *Vitis aestivalis* accession (Fung et al, 2008). This experiment focused on the molecular mechanisms employed by the host plant in resisting the powdery mildew infection.

Understanding the parasite’s transcriptome is a key research area for controlling grape powdery mildew (Qiu et al., 2015). A transcriptome is the sum of the expressed genetic material in a sample at a given time. Unlike the genome, the transcriptome may vary depending on a variety of factors such as the part of the organism that the sample was taken from, the cell types included in the sample, the time at which the sample was taken, and the environmental conditions at the time the sample is taken.

One can build an idea of what properties *E. necator’s* genome and transcriptome may have by examining the genomes of its close evolutionary relatives. Spanu et. al. delves into the genomes of a close cousins of *E. necator*, *Blumeria graminis* (Barley Powdery Mildew), *Erysiphe pisi* (Pea Powdery Mildew), and *Golovinomyces orontii* (Thale Cress Powdery Mildew). Of these three species, *Erysiphe pisi* is in the same genus as *E. necator* while the other two fall in the broader family of Erysiphales. Genomic analysis reveals that Powdery genomes are up to four times larger than the genomes of other groups of ascomycetes. Despite the larger genome size, powdery mildew possessed fewer genes than other fungal groups within the division *Ascomycota*. According to the authors, the reason behind these abnormalities in the Powdery mildew genomes is the obligate biotrophic property of the fungus. Since a biotrophic, fungus
must contend with the immune system of a living host, it must be able to support a large degree of variability within the genome in order to keep pace with the adaptive defenses of the host plant. A majority (84%) of the effector proteins identified in barley powdery mildew were unique to that species out of the three species analyzed. The lack of conservation among effector genes is indicative of high variability within these genes due to co-speciation with the host plant (Spanu, 2010).

Jones et al. (2014) outlines the sequencing of the first *E. necator* genomes. They sequenced the genome of *E. necator* from five isolates to gain a better understanding of the observed resistance by *E. necator* to commercial fungicides. The genome scaffolds, or the contiguous sequences generated from the assembly process, were around 50 mb. This scaffold size is far off from the estimated genome size, which is around 130 million bp. This discrepancy is due to highly repetitive elements within the *E. necator* genome. Repetitive sections of the genome are challenging for shotgun-sequencing approaches because it is unclear based on short reads how many repeats were present in the original sequence and where they properly map. The inferred size of the genome is based on the kmer distribution. As discussed, the estimated genome size for *E. necator* places it in a special position in that its genome is larger than most plant pathogens (Jones et al., 2014). The results of Jones et al. 2014 are compared with other omics analyses of powdery mildews in Bindschedler et al. 2016. According to the results from genomic sequencing of *Erysiphe necator*, *Blumeria graminis f. sp. hordei* and *B. graminis f. sp. tritici*, powdery mildews exhibit significantly lower gene densities than other members of Ascomycota (Figure 2).
This research will base itself upon two cDNA libraries. The first transcriptome was produced using 454 sequencing in 2012 (Myers 2012). The second transcriptome was generated based on Illumina sequencing in 2014. The Transcriptome generated in (Myers 2012) is referred to henceforth as the G14 Transcriptome and the transcriptome generated in (Jones 2014) will be referred to as the C-strain Transcriptome.

The guiding question in this thesis is to determine the number of transcripts that can be identified in the G14 transcriptome that are not included in the C-strain transcriptome. Online databases are continuously receiving new entries and refining curation. Since the G14 and C-strain transcriptomes are 4 and 6 years old respectively at the time of the completion of this research, the possibility is open that improved annotations could be determined through repeated database
searches. Novel annotations were developed using two methods, The online eggNOG-mapper (Huerta-Cepas 2017) as well as the NCBI non-redundant protein database (Altschul 1990).

Functional annotation through homology is the de-facto standard for non-model organisms, and Basic Local Alignment Search Tool (Blast) is the primary tool for testing homology. While the best Blast hit might not represent the closest evolutionary neighbor (Koski 2001), similarity in sequence can be considered strong evidence for preservation of function. Its instrumentality in finding homology results from its speed as well as its predisposition to find local alignments. Blast functions by locating short regions of perfect alignment and extending them in both the 5’ and the 3’ direction until a given threshold is reached. Blast is employed as the flagship search algorithm by the National Center for Biotechnology Information (NCBI) where it is referred to as NCBI-Blast. Blast is implemented in several different modalities depending on the type of sequence used as a query and as a database. This project employs two of these modalities Blastn and Blastx. Blastn refers to a blast search where the query set is comprised of nucleotide sequences as well as the database. Blastx refers to a blast search where the query set is comprised of nucleotide sequences but the database consists of Amino acid sequences.

The identification of orthologous groups represents a more recent approach to functional annotation. Instead of annotating based on a one to one alignment between a query sequence and a target sequence, orthology based annotation uses a database of orthologous groups. Orthologous group represent a collection of genes taken from multiple sequences that have similar sequence. The similarity in sequence in these groups is used as evidence that they share the same function (Tatsuov 1997). The task for the algorithm is to determine to which orthologous group a given query sequence belongs. The eggNOG-mapper applies a combination of an alignment algorithm called Diamond to match query sequences held within the Eggnog database. This method of annotation benefits from the corroboration of multiple homologs (Huerta-Cepas 2017).

In addition to functional annotation through homology, a novel method was developed in this project for predicting reading frame. In this approach, transcripts of known reading frame were used to train a Markov model based on codon transition frequencies. The assumption behind this
method was that sequences in the same reading frame will have more similar codon transition patterns than sequences that are in different reading frames.

This project was able to produce a set of 3518 novel functional annotations for the Grapevine Powdery Mildew fungus. These annotations were found in previously unannotated areas of the Genome. The novel Hidden Markov algorithm was also found to correctly predict the reading frame in raw transcriptomic sequence.
METHODS

Data Sources
A general overview of work completed in this project is featured in Figure 3.

![Chord Diagram Depicting all major alignments made in the project. The arrow originates at the query and points toward the target database. Bi directional arrows indicate a reciprocal alignment.](image)

The C-strain Transcriptome was generated using a combination of RNA sequencing data along with gene models based on a genome scaffold, both using Illumina sequencing by synthesis (Jones et al 2014). This transcriptome consists of 6,533 genes. These were identified using a genome-guided de novo assembly. The resulting gene models were then used for additional ab initio gene discovery using Augustus on repeat-masked scaffolds and combining these transcripts with homology-based gene prediction from peptide sequences of *Blumeria graminis* f. sp. *hordei*, *B. graminis* f. sp. *tritici*, and *Botrytis cinerea*. This set of peptides excludes any peptide that aligned via BLASTP to any protein associated with transposable elements (e < 10^{-10}).

Candidate transcripts that did not match any entries using BLASTX in the NCBI non-redundant protein database (e < 10^{-3}) were also removed unless evidence of transcription was present in the RNA seq data.
The G14 Transcriptome is based on reads generated from a single reference transcriptome. These reads were then assembled into 39,686 contigs, including splice variants. The contigs in the G14 transcriptome are divided into three partitions. The first 12,148 contigs were ordered based on homology to *Blumeria graminis* f. sp. *hordei* (barley powdery mildew) proteins. The second partition consists of 15,728 contigs which were ordered based on similarity to the NCBI non-redundant protein database. The final partition consisted of 11,810 contigs which were ordered by descending length.

G14 and C-strain’s transcriptomes were both used as queries against five different *E. necator* genome scaffolds, representing different strains of fungus: Lodi, C, Ranch-9, e1-101, and Branching. The published annotations including those used for the transcriptome were based on the C-strain scaffold. Thus, the C-strain scaffold was used as a database for aligning the G14 Contigs. Since the C-strain scaffold represents a recent and annotated version of the *E. necator* genome, any contigs that align to loci not previously annotated could be evidence for a novel transcription site. The reasoning for any difference in expression content could have several causes rooted in the process by which the two transcriptomes were developed. The methodology used to prepare these transcriptomes differed at several points (Table 1). Differences emerged in the fungal tissues extracted, normalization of cDNAs for library preparation, high throughput sequencing approach, and the method by which the sequencing reads were assembled.

*Table 1: Comparison of methods*

<table>
<thead>
<tr>
<th>Stage of preparation</th>
<th>G14</th>
<th>C-strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tissue harvest</td>
<td>Epiphytic peels</td>
<td>Whole infected leaf</td>
</tr>
<tr>
<td></td>
<td>(Cadle-Davidson 2010)</td>
<td></td>
</tr>
<tr>
<td>Library preparation</td>
<td>Normalized cDNA</td>
<td>Non-normalized cDNA</td>
</tr>
<tr>
<td>Sequencing</td>
<td>Pyrosequencing (454)</td>
<td>Sequencing by synthesis (Illumina)</td>
</tr>
<tr>
<td>Assembly</td>
<td>De-novo</td>
<td>Genome-guided</td>
</tr>
<tr>
<td># Contigs</td>
<td>39,686</td>
<td>6,533</td>
</tr>
</tbody>
</table>
**Quality Control**

Blast hits were evaluated based on five criteria. These included percent identity, query coverage, alignment length, bit score, and e-value. Percent identity refers to the percentage of bases in the region of overlap that were identical across query and subject. Query coverage refers to the percentage of the total query length that is contained within the area of overlap. Alignment length refers to the number of base pairs that comprises the overlap between the query and the subject. The bit score is a single value that reflects the quality of the alignment based on the number of matches, mismatches, insertions, and deletions. The bit score is dependent on the length of the alignment, meaning a longer alignment with a large number of gaps and mismatches could have a similar bit score to a shorter alignment with fewer gaps and mismatches. An e-value is a statistical measure produced by the BLAST algorithm that reflects the confidence level of the query subject pairing. The e-value is the probability that the given query could be a false positive. While the e-value is not necessarily as dependent on the length of the alignment as is the case with a bit score, it is dependent on the database used when executing the blast search. An e-value is the probability that the given query could find a hit of similar strength somewhere else in the database. Therefore, e-value will be dependent on the size of the database used in the blast search.

**Reciprocal blast**

Reciprocal blast was carried out using the terminal version of Blastn. Quality control was levied upon blast results by assigning thresholds to the e-value, percent identity, alignment length, and percent query coverage. Only those Blast hits that returned an e-value lower than 1e-10 were included in the analysis.

**Genome Scaffold Alignment**

Blastn was also used to align all contigs from both the G14 and the C-strain transcriptomes. Of the five Genome scaffolds described in Jones et al. 2014, only the C-strain scaffold was included in the analysis. Only the best hit for each contig was included in the final analysis. To visualize the alignment locations, an R script was generated which plots a given scaffold along with all of the aligned transcripts and predicted gene models.
Total scaffold coverage was determined by enumerating the number of bases that aligned to only a C-strain transcript, only a G14 transcript, both a C-strain and a G14 Transcript, as well as which bases aligned to neither transcriptome. Only the best hits for each transcriptome were include. Percent identity was not taken into account for these enumerations. Novel transcripts were designated by comparing the start and stop loci on the subject strand as returned from Blast. An alignment was considered novel if it aligned to a genomic locus where there were no existing gene models, contigs that aligned to C-strain transcripts were also removed from the novel transcript set. The locations of existing gene models were taken from a gff file created in (Jones et al. 2014).

**NR Protein Blast Annotation**

Novel annotations for the G14 contigs were taken from a Blastx against the NCBI non-redundant Protein database. Only those Blast hits that returned an e-value lower than $10^{-10}$ were included in the analysis. Annotations generated through NR-blast were stored in gff3 file format. Transcripts that did not align to known *E. necator* proteins but did align to other ascomycetes were considered novel. Non-redundant database annotations can be found at kimchi.rit.edu/~mosier/lab/.

**EggNOG-mapper Annotation**

Novel annotations for *E. necator* contigs were derived from the NCBI non-redundant protein database as well as the EggNOG-mapper tool. The online version of the annotation tool was used [http://eggnogdb.embl.de/-/app/emapper]. The EggNOG-mapper provides two different algorithms for alignment. Annotations for the C-strain and G14 transcripts were detected using the Diamond algorithm. The taxonomic scope was constrained to Ascomycota. Annotations generated through the EggNOG-mapper were stored in gff3 file format. These annotations can be found at kimchi.rit.edu/~mosier/lab/.

**Gene Ontology Term analysis**

GO Slim mappings were conducted using the online GO Slim Mapper from the Saccharomyces Genome Database (SGD) [https://www.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl] Yeast GO slims were chosen as the background for the mapping.
**Reading frame determination**

A novel Markov model approach was implemented. This approach involved decomposing all contigs into codons. Reading frames are distinguished by starting position as well as strand. There are three possible stating positions on each stand of cDNA resulting in 6 total reading frames for each contig. The reading frames are labelled as follows: 1plus, 2plus, 3plus, 1minus, 2minus, and 3minus. The C-strain transcriptome was formed in a manner so that all of the contigs were presented in the 1 plus reading frame. This means that the first three nucleotides of each contig represented the first codon. This can be verified by taking the published amino acid sequences and aligning them to the translated contigs. This Markov model was applied to the G14 contigs, the reading frame with the greatest normalized log odds score was recorded for each. These reading frame predictions are stored in the G14_nr_novel.gff file.

**Sequence alignment visualization**

Colored sequence alignment visualizations were generated using the online muscle multiple alignment tool. Amino acid sequences were input in fasta format and the ClustalW algorithm was employed to perform the alignment (Edgar 2004).
RESULTS

Contig Length
C-strain’s contigs were longer on average than the G14 contigs (Fig. 4). The mean contig length in the G14 transcriptome was 647 base pairs while the mean contig length in the C-strain transcriptome was 1,416 base pairs.

Figure 4: Violin plot comparing contig length. Thickness reflects the distribution of the data. Note that the x axis is on a logarithmic scale. Boxplots denoting the median and interquartile range are also included.

Reciprocal Blast
To determine the level of overlap between the G14 and the C-strain transcriptome, a nucleotide blast was carried out in a reciprocal manner. With an e-value cut of $10^{-10}$, 97.6% of C-strain’s transcripts aligned to the G14 Transcriptome while 38.6% of G14’s contigs aligned to the C-strain transcriptome (Figure 5). Multiple alignment was observed in both directions, with each C-strain transcript aligning on average to 3.4 contig and G14 contigs aligning on average to 1.4 C-strain transcripts. G14 contigs that aligned were on average longer than those that did not find hits ($p < 1^{-16}$). Of the 15,317 of G14 contigs which aligned to the C-strain transcriptome, 15,260 (99.6%) had matches in the reciprocal blast when the C-strain transcriptome was used as the query set. Of the 6,376 C-strain transcripts that aligned to the G14 transcriptome, all but 2 were also found in the target set when the roles were reversed.
**Aligning Contigs to E. necator C-strain Genome Scaffold**

All 6,533 of C-strain’s C-strain transcripts aligned with the C-strain genome, and 38,838 (97.9%) of the G14 contigs aligned to the C-strain genome scaffold (e < 10^{-10}). Only 2 of the G14 contigs that aligned to the C-strain transcriptome did not also align to the C-strain genome scaffold. However, 23,523 (60.6%) of the contigs that aligned to the genome scaffold did not find hits in the C-strain transcriptome. Only the highest scoring hit for each query sequence are included in the following analysis. Figure 6 depicts the distribution of blast scores for each transcriptome against the C-strain scaffolds. The average blast score for the C-strain Alignments was larger than that for the G14 alignments. The percent identity distribution for the C-strain hits clustered tightly around 100% while the G14 alignments fell along a broader distribution with most of the hits exceeding 95% identity. On average, C-strain’s transcripts displayed significantly greater alignment length (p < 10^{-16}) and percentage identity (p < 10^{-16}). However, the alignment of the G14 contigs increased the percentage of the C-strain scaffold with evidence of expression from 13% to 38% (Figure 8).
Figure 6: Alignment quality distributions of G14 and C-strain Transcriptomes against the C-strain scaffold. Only the best hits are included in this figure.

Figure 7: Comparison of % identity and alignment length for alignments to the C-strain genome scaffold.
Most of the 38,838 G14 contigs that aligned to the C-strain scaffold (72.0%) matched previously unannotated loci. For those that co-located with C-strain transcripts, relative mapping to the scaffolds indicated that 6820 G14 contigs extended the 5' and/or 3' expression evidence of C-strain transcripts; however, the converse was also frequently true, with 4845 C-strain transcripts extending beyond G14 contigs (Figure 11). Specific examples of C-strain and G14 alignments to the C-strain scaffold illustrate this concept of relative mapping (Figures 9 and 10). In some cases, G14 contigs extend a region of C-strain alignment in the 3' direction (Figure 9). In other cases, the G14 contig alignments extended the C-strain alignments in both the 3' and the 5' direction (Figure 10).
Figure 9: Example of a G14 extension of a C-strain Transcript on the 3’ end. Green lines represent C-strain transcripts and blue lines represent aligned G14 Transcripts.

Figure 10: Example of a C strain scaffold where G14 contigs extend a C-strain in both the 5’ and the 3’ direction.
Figure 11: Quantification of the relative positions of C-strain and G14 alignments. Perfect overlap refers to instances where contigs from either transcriptome fell on the same location. Upstream and downstream refers to overlapping alignment with 5’ and 3’ overhang respectively. G14 only refers to cases where a G14 contig aligned with a completely new location on the scaffold. Most of G14 contigs found their best alignment at novel locations.

The set of G14 contigs that aligned to novel loci on the C-strain scaffold were identified and characterized for alignment length and bit score (Figure 12). The mean bit score for these novel loci was 389 and the mean alignment length was 410 base pairs while the mean identity was 98.2%. These novel loci were distributed over 3,303 of the C-strain scaffold contigs.

**Annotations**

Functional annotation was carried out on both the G14 and C-strain Transcriptomes. The G14 transcriptome was annotated using the non-redundant protein database. Figure 13 displays the number of contigs by the species of the best homolog returned for those contigs. The largest source of annotations came from Barley Powdery Mildew (*Blumeria Graminis f. sp. hordei*) with 411 transcripts. All species included in this figure belong to Ascomycota.
Figure 12: Alignment quality distribution for G14 contigs that aligned at novel loci on the C-strain genome scaffold. Black vertical and horizontal lines represent the mean % identity and alignment length for the contigs that aligned to previously annotated loci respectively.

Figure 13: Non-redundant blast annotations divided by subject species.
Both the G14 and C-strain Transcriptomes were submitted to the online annotation tool called the EggNOG-mapper. The G14 transcriptome returned ~2,800 transcripts in each of the six reading frames while the C-strain transcriptome returned 5424 annotations all in one reading frame. The EggNOG-mapper returned annotations for 11,928 of the G14 contigs ($e < 10^{-10}$).

**Novel Transcription sites**

The location of the best hit for each of G14’s and C-strain’s contigs were superimposed and the number of unique and shared loci were recorded. Of 38,838 G14 contigs that aligned to scaffolds, 27,935 aligned to novel loci while the remaining 10,903 contigs aligned to loci which had been annotated previously. After the contigs which aligned to C-strain transcripts and aligned to existing *E. necator* peptides in the NR database were removed. A final set of 18,571 novel contigs were left over. The average blast score was greater for those contigs that aligned to pre-existing gene model loci or to C-strain transcripts than for contigs that aligned to novel loci ($p < 10^{-16}$).

The annotation data was superimposed onto the potential transcription sites to yield the following results. Of the total 18,571 novel contigs that aligned to novel loci on the C-strain scaffold, 3,518 were annotated by either NR Blast or the EggNOG-mapper while 15,453 remain unannotated. The EggNOG-mapper mapped Gene Ontology (GO) terms to 532 of the contigs that mapped to novel loci.

**Novel Annotations**

The EggNOG-mapper included predicted gene names for 9,070 of the G14 Contigs and 3960 of the C-strain contigs. Of the 9070 contigs that received predicted gene names, 902 (~10%) were annotated with gene names that were not found in the C-strain gene names. These predicted gene names were used as input for the GO Slim Mapper and the results are included in figures 14, 15, and 16 which display mappings to the Yeast GO Slim terms by molecular function, cellular component and biological process respectively. For Molecular Function, the G14 Transcriptome returned the highest number of novel annotations associated with transferase activity with 59
predicted gene names associated with that term. The cellular component GO Slim term with the most novel G14 annotations was the cytoplasm with 155 predicted gene names. Finally, the top Biological process GO Slim term was “transcription from RNA Polymerase II promoter” with 41 predicted gene names.

An example of a high-quality annotation is found in the G14 contig with the identifier EN8380. This contig aligned with 93% identity and 85% query coverage to the Flavoprotein subunit of succinate dehydrogenase enzyme of *Blumeria graminis sp tritici*. This contig aligned with 100% identity and 99% query coverage to the *E. necator* C-strain scaffold. A visualization of the full amino acid level alignment of this contig to the C-strain scaffold and the *B. graminis* succinate dehydrogenase can be seen in figure 17.
Figure 14: Novel go slim mappings: Molecular Function. All GO Slim Molecular Function terms are included. The bars are ordered by the number of novel G14 Genes.
Figure 15: Novel GO Slim Mappings: Cellular Component. All Cellular Component Yeast GO Slim Terms are included. The bars are ordered by the number of novel G14 Genes.
Figure 16: Novel GO Slim Mappings: Biological Process. The red portion of each bar represents the number of predicted genes that were not already included in the C-strain annotation set. The bars are ordered by the number of novel G14 Genes.
Figure 17: Alignment of G14 contig to B. graminis (B.) succinate dehydrogenase, flavoprotein subunit.


**DISCUSSION**

Grapevine Powdery Mildew continues to present a challenge to grape growers all over the world. Knowledge of the transcriptome of the parasite could provide valuable downstream information for researchers aiming to develop host resistance or safer and more effective fungicides. The purpose of this project was to compare two existing transcriptomic Powdery Mildew resources. At the time of this project, there were 6,515 protein sequences and 2,597 nucleotide sequences for *E. necator* stored on the National Center for Biotechnology Information. Given the differences in the preparation procedures for these data sets, we stand to learn more from where they agree rather than where they disagree. The behavior of the transcriptomes is as we would expect given the process by which each was generated. It is unsurprising that the C-strain transcriptome exhibited strong alignment given that those contigs were based on the C-strain scaffold. The veracity of these transcripts is strengthened by the alignment of G14 contigs at the same loci.

The fact that almost all of the transcripts from the C-strain genome aligned to the G14 transcriptome is further evidence for the veracity of those transcripts. The 2.4% of transcripts that did not align to the G14 strain are likely a result of divergent evolution. As an obligate parasite, Powdery mildew displays a rapid mutation rate to adapt to host defense. Multiple between the C-strain and G14 transcriptomes is likely to due to a combination of two factors. The first is that the G14 transcriptome contains multiple contigs reflecting the same transcribed loci. And the second reason is the fact that the C-strain transcriptome was filtered to remove transposable elements while the G14 transcriptome was not. There is more information to be gathered in the set of G14 contigs that aligned to the C-strain scaffold but did not align to an existing gene model loci or any member of the C-strain transcriptome. These alignments could be evidence for potential novel transcription sites or be the result of spurious alignments. If the true number of genes in the *E. necator* genome is close to the number reported by Jones et al 2014, this would mean that the gene density is among the lowest in Ascomycota (Bindschedler 2016).

The exact cutoff values to use for identifying true alignments is dependent upon the size of the database being queried. The expected (e) value provides a general reflection of the quality of the alignment but is imprecise and does not provide meaningful information about the alignment itself. The alignment score represents a better summary of a given Blast alignment because it
correlates more strongly with other alignment quality metrics such as percent identity and query coverage. A combination of percent identity and alignment length provides the most intuitive reflection of alignment quality. For example, an alignment with 99% identity that is more than 500 base pairs long is almost certainly a reflection of a true biological similarity. The most encouraging aspect of the alignment results between the G14 transcripts and the C-strain scaffold was the strong percent identity values.

The dominant pattern observed in this data was repetitive sequences, even among expressed regions. This was made clear through both the self-alignment of the transcriptomes as well as the multiple alignment discerned through the alignment to the genome scaffold. This observation is supported by (Jones 2014) which makes the claim that this repetition comes as a result of transposon activity. Analysis of multiple transcriptomes has confirmed that several hundred genes within the *E. necator* genome represent repetitive elements. This pattern was mirrored by the results of aligning the G14 transcriptome to the C-strain scaffold where these contigs found 6.1 hits on average.

Multiple streams of evidence were taken into consideration over the course of this project. One question that can be raised in lieu of these results is the extent to which the type of technology used could influence the quality of the sequence. The G14 transcriptome was generated using the pyrosequencing method developed by 454 Life Science. The C-strain Transcriptome on the other hand was generated using Illumina sequencing. Given the current state of the sequencing market, one is led to believe that Illumina sequencing is the cheaper and more effective method. Illumina has outlived pyrosequencing, but that does not imply that the data generated using pyrosequencing does not hold value. Another explanation for the low quality of overlap could be the biological state that the fungal samples were in when they underwent sequencing.

A major limitation in this project is the dynamic nature of sequence databases. The results of the homology annotations reported here represent a snapshot of the databases at this present moment in time. The results are markedly different from that which was found in G14 2012 and it is safe to assume that a query of the same sequences will return a markedly different set of annotations.
This can be seen as good or bad depending on the perspective taken. The problem that arises from this is that no set of homology annotations can be considered final, only the best to date. However, the continued expansion and curation of sequence databases also serves as a source of optimism that our ability to determine the function of novel transcripts will become increasingly likely over time.

Jones et al 2014 discovered that a large portion of the proteins expressed by *E. necator* are unique to the species. This pattern was reflected in the results of this project. Where a non-trivial portion of the contigs found annotations with middling significant and alignment length. It is likely given the strong evolutionary pressure that is exerted upon *E. necator* which lives in constant conflict with the immune system of the host *Vitis vinifera*. The effector proteins present in biotrophic fungi are species specific almost by definition. The lack of conservation among these proteins means that the criteria for identifying an effector protein are broad, including short sequence length, the presence of a signal peptide and the lack of any transmembrane domains (Bindschedler 2016).

One of the key concerns made in the G14 2012 was the probability of a given contig being a chimera or a partial transcript. The presence of a genome scaffold helped identify these cases by tracing the local regions of alignment stemming from a single contig. The fact that such an overwhelming majority of the G14 contigs aligned to the C-strain scaffold is a reflection of the quality of the assembly.

Cases where a contig from the G14 transcriptome align to the same loci as a C-strain Transcript or Gene model on the C strain scaffold not only strengthen the case for that loci holding a gene. The presence of 5’ and 3’ overhangs also indicate that additional exons are present at loci. This serves as evidence for alternative splicing at these loci. Alternative splicing has been reported in several species of fungi and likely serves as an asset in the ongoing molecular battle that the Powdery mildew organism is engaged in with the host (Zhao 2013).

Countless hours could be spent evaluating the function of each of the novel transcribed regions discovered in this project. However, given the growing field of advanced high throughput sequencing technologies, my prescription would be for future researchers to levy those tools to
build a more complete picture of the inner mechanisms of this extraordinary parasite. Further in-depth annotation of each gene model and transcript would help researchers select targets for gene silencing. Annotation through homology can only take us so far however, confidence in the role that each of these transcripts plays in the organism must be grounded in experimental results. The low biomass availability and the inability of the fungus to grow apart from the host represent obstacles in the pursuit of lab based molecular characterization. New methods for gene knockout that do not rely on culturing and transformation might provide a solution to this however (Nowara 2010). While we wait for these new laboratory methods to bear fruit, we can focus on divulging as much value as we can from the available bioinformatics tools and databases.

A valuable next step to take with regard to this dataset would be to develop a means to identify candidate effector genes from the G14 transcriptome. Knowledge of which transcripts coded for effector proteins would benefit downstream analysis and development of novel fungicides. This is not a simple task, the high diversity and mutation rate of known effector proteins makes the identification of novel effectors a complex problem prone to error. An accurate effector protein prediction tool would be especially advantageous in this area.

The transcriptomic sequences themselves present a valuable step towards gene silencing as a means for Powdery Mildew control (Yin 2010; Liu 2002). These sequences can be used to develop small RNA expression vectors that might silence specific target genes within the fungus. Annotation data can then be leveraged to inform target selection. The absence of an annotation would not necessarily exclude a transcript from candidacy as a target. If it can be demonstrated in vivo that these transcripts are both endemic to E. necator and are necessary for fungal growth and/or reproduction. That would be sufficient evidence to justify the formation of a microRNA target.
Conclusion

The G14 2012 *E. necator* transcriptome displays strong alignment to the C-strain Genome scaffold. This alignment suggests the presence of more protein coding genes than predicted by Jones et al. 2014. This project also resulted in a novel set of annotations describing a subset of the G14 Transcriptome. It was also found that a Markov model can be used to determine reading frame in de novo assembled transcripts.
References

Works Cited


Campbell, Christopher. *The botanist and the vintner: how wine was saved for the world.* Algonquin Books, 2006.


GO Slim Mapper, (Saccharomyces Genome Database) Available from: https://www.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl


