Project Proposal for Analysis of Novel *Trichoderma* Genomes

Connor Burbridge[1]

[1]USask NSID: cbe453 , USask ID no. 11162928 , Supervisors: Dave Schneider & Tony Kusalik,

August 31, 2022

Contents

1 Background **3**

1.1 Trichoderma 3

1.2 Genome Assembly 3

1.3 Genome Annotation 4

1.4 Repeat Identificatio/Masking and Identification of AT-rich Genomic regions 4

1.5 Gene Prediction 5

2 Research problem **6**

3 Deliverables **6**

4 Timeline **7**

5 References **7**

1 Background

1.1 Trichoderma

Crop resistance to environmental stressors is a necessity for crop health and overall crop yields. Current popular methods for crop protection involve the use of pesticides and genetically modified organisms, which can be expensive and potentially politically dividing in the case of GMOs (citation needed). In addition, crops will suffer when soils are not sufficient for crop growth and health. These soil insufficiencies can include drought stress, nutrient stress and can also include solis that have been contaminated with hydrocarbons, making it difficult to grow crops in those regions and provides an opportunity for new bioremediation processes. Recently, two strains of *Trichoderma* have been identified in the prairie regions of Alberta and Saskatchewan. These two strains, named Tsth20 and DC1, have been found to have unique properties when incoluated in plants in the soils mentioned before.

Trichoderma is a type of fungi that can colonize the roots of plants in a non-toxic, non-lethal, opportunistic symbiotic relationship.[1] Many strains of *Trichoderma* have been shown to provide resistance to bacteria and other fungi in soils through the use of polyketides, non-ribosomal peptide synthetases and other antibiotic products[1][2]. In addition to these beneficial properties, the two strains mentioned above provide even further protection for plants in dry, salty soils and may also be considered as a bioremediation tool in soils contaminated with hydrocarbon content. However, little is known about how these mechanisms work in these new strains, so DC1 and Tsth20 were sequenced in an attempt to better understand the details of their genomes and secretomes.

1.2 Genome Assembly

Sequence assembly has been a long-standing issue in the field of computer science[8]. Determining the correct order and combination of smaller subsequences into an accurate sequence assembly is also computationally difficult in terms of compute resources such as memory, CPU count and storage required for input sequences[8]. In addition to these difficulties, there can be difficulties encountered during asssembly due to the nature of the data or genomes themselves. Insufficient data used in an assembly may result in short, fragmented assemblies, depending on the size of the genomes while sequence data that is not long enough can fail to fully capture repetitive regions in an assembly. To solve this problem, a wide range of assembly tools have been developed with their own unique approaches to genome assembly problems, so it is important to use an appropriate assembler for the task at hand, and important to evaluate the assembly thoroughly. One approach to aid in the issue of genome coverage during assembly, is to use a combination of long and short reads in what is known as a hybrid assembly. Combining both highly accurate short reads with deep coverage along with less accurate but much longer reads can produce high quality genome assemblies that capture long repetitive regions. Assemblies must also be evaluated with measures such as N50, L50, coverage, average contig length and total assembled length to ensure that the genomes assemble well at least based on those metrics[8]. Following appropriate assembly protocols is essential to the further success of a project as downstream processing such as annotation depends on a high-quality assembly.

1.3 Genome Annotation

With the explosion of sequence data and genomes assemblies made available in recent years, genome annotation has become a crucial part of the sequence analyiss pipeline. Genome annotation can involve the annotation of genes as well as the annotation of structures within a genome. These annotations can be performed using either evidence oriented homology-based annotation programs or *ab-initio* statistical methods which do not consider existing evidence[7]. It may also be possible to use a combination of both in some circumstances. Furthermore, the downstream products of these genes can then be annotated for functional properties to determine what functions these genes and gene products could potentially perform.

1.4 Repeat Identification/Masking and Identification of AT-rich Genomic regions

In addition to gene finding, repeat identification within assembled genomes is another problem that needs to be considered. Regions with long repeats can have a significant impact on genome assembly as well as gene finding due to the limitation of short reads used in some assemblies[9]. Short reads may be unable to bridge or cover entire repeat regions within a genome, so it is important to consider the use of long reads such as Nanopore or PacBio when pursuing a new genome assembly project. It is also possible for these repetitive regions to contain genes as well, making for an interesting investigation in regards to *Trichoderma*, as fungal genomes have been shown to contain many repeat regions with a high concentration of A and T nucleotides[10]. Once these repetitive regions have been identified (by any number of existing programs), the genome can be masked to exlude these regions in downstream processing if desired. This may affect the gene finding process described later and may be an interesting topic to look into considering the large number of available gene finding programs available.

1.5 Gene Prediction

Gene finding is another long standing computational problem in bioinformatics, which concerns itself with identifying potential genes within genomes based on patterns or evidence considered by the gene finding program. These two methods are called *ab initio* for programs that search for patterns and gene structure, while similarity or evidence-based searches use prior information such as RNAseq data, expressed sequence tags and protein sequences to identify gene within a new genome[11]. Complicating the process more are the introduction of introns and alternative splicing in Eukaryotes, making it possible for one 'gene' to have several differing versions at the same locus. Examples of *ab initio* methods include tools such as GeneMark-ES[4] and AUGUSTUS[5], while evidence based methods include tools such as Braker2[3], which can incorporate existing data such as RNAseq, protein sequences etc. in gene prediction models. As mentioned in the repeat finding section of this proposal, long repeats and transpoable elements can cause issues for gene finding programs, making this an interesting area of research, at least for fungal genomes such as *Trichoderma*. Fungal genomes are also interesting in the topic of gene finding as there are few programs targeted directly at gene finding in fungal genomes while organisms such as human, mouse, and *Arabidopsis* benefit from having many tools tested on them as they are model organisms in the field, at least according to table 1 in [11]. How these different methods and tools perform when applied to fungal genomes is an important consideration as fungi have features that can benefit plant growth as mentioned earlier.

2 Research problem

Genome annotation, in the case of gene finding in this work, has been a popular computational problem for decades. The identification of possible genes provides other researchers with a valuable resource for future research avenues. Due to the importance and popularity of this topic, there have been a variety of gene finding tools developed, each with their own implementation for certain applications or use cases. The abundance of gene finding tools raises questions such as which tool is better? Do different tools provide drastically different results? How much overlap exists between gene finding programs? Are there stark differences between *ab initio* methods and similarity-based methods? Comparative studies of gene finding tools have been performed before, however most of these studies are in reference to popular eukaryotic subjects of study such as humans, mice and *Arabidopsis*. Few if any comparative studies have been performed in fungal genomes, and even fewer in the case of *Trichoderma*.

This lack of comparative analysis in fungal genomes provides a research avenue to compare results from the application of different gene finding tools to new and existing *Trichoderma* assemblies. Results from using different annotation tools will allow us to identify a 'core' genome along with genes that are not predicted by all tools in all genomes, or outliers. This analysis will also allow us to compare results from both evidence-based gene finding methods as well as *ab-initio* gene finding methods. One interesting area to investigate between these methods would be in regions of low nuclotide diversity, or the AT-rich regions common to *Trichoderma* and other fungi. It is possible that gene finding programs may have difficulty with these regions of genomes, as they may contain repeat regions and as mentioned before, have low nucleotide diversity, making it unlikely but not impossible for genes to be found in these regions. Results from this work will provide insight into the use of different gene finding tools in the context of *Trichoderma*. This work will also provide a large set of called genes spanning at minimum two *Trichoderma* genomes, along with two new assemblies which have been analyzed for repeat and AT-rich regions.

3 Deliverables

- Lists of genes for each *Trichoderma* assembly considered

- A consenus or 'core' gene set shared by each genome considered based on different gene finding tools

- Repeat regions identified in assembled *Trichoderma* genomes.

- A potential list of true positive based on genes which have supporting RNAseq eveidence

4 Timeline

- Initial genome assemblies:

- Gene finding and annotation results:

- Analysis of gene finding results:

5 References

References

[1] Hermosa, R., Viterbo, A., Chet, I., Monte, E. (2012). Plant-beneficial effects of *Trichoderma* and of its genes. *Microbiology. 38*, 17-25.

[2] Ramirez-Valdespino, C., Casas-Flores, S., Olmedo-Monfil, V. (2019). *Trichoderma* as a Model to Study Effector-Like Molecules. *Frontiers in Microbiology 15.*

[3] Hoff, K. J., Lomsadze, A., Borodovsky, M., Stanke, M. (2019). Whole-Genome Annotation with BRAKER. *Methods Mol Biol.*, *1962*, 65-95.

[4] Borodovsky, M., Lomsadze, A. (2011). Eukaryotic Gene Prodeiction using GeneMark.hmm-E and GeneMark-ES. *Current Protocols in Bioinformatics. 4.*

[5] Stanke, M., Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research. 33*, 465-467.

[6] Jones, P., Binns, D., Chang, H. Y., Fraser, W., Li, W., ... Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics. 30*(9), 1236-1240.

[7] Yandell, M., Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics. 13*, 329-342.

[8] Sohn, J., Nam, J. (2016) The present and future of *de novo* whole-genome assembly. *Briefings in Bioinformatics. 19*1, 23-40.

[9] Lerat., E. (2009). Identifying repeats and transposable elements in sequences genomes: how to find your way through the dense forest of programs. *Nature Heredity. 104*, 520-533.

[10] Li, WC., Huang, CH., Chen, CL. et al. (2017). Trichoderma reesei complete genome sequence, repeat-induced point mutation, and partitioning of CAZyme gene clusters. *Biotechnol Biofuels. 10*, 170.

[11] Wang, Z., Chen, Yazhu., Li, Y. (2016). A Brief Review of Computational Gene Prediction Methods. *Genomics Proteomics Bioinformatics. 2*4, 216-221.