# Project Proposal: Comparative analysis of Gene Finding tools when applied to *Trichoderma* genomes

Connor Burbridge[1]

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

October 19, 2022

# Contents

**1 Background** **3**
- 1.1 Trichoderma . . . 3
- 1.2 Genome Assembly . . . 3
- 1.3 Repeat Identification/Masking and Identification of AT-rich Genomic regions . . . 4
- 1.4 Gene Finding . . . 5

**2 Research problem** **6**

**3 Deliverables** **7**

**4 Timeline** **8**

**5 References** **8**

# 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[1]. In addition, crops suffer when soils are not sufficient for crop growth and health. Soil insufficiencies can result in drought stress as well as nutrient stress, leading to poor overall yields.

*Trichoderma* is a type of fungi that can colonize the roots of plants in a non-toxic, non-lethal, opportunistic symbiotic relationship[2]. 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[2][3]. 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 beneficial properties when used as an inoculant for plants in the soils mentioned before. In addition to these beneficial properties, the two strains mentioned previously provide even further protection for plants in dry, salty soils and one strain also has potential for use as a bioremediation tool in soils contaminated with hydrocarbon content. However, little is known about the mechanisms at work in these strains, so DC1 and Tsth20 were sequenced by the Global Institute for Food Security (no publication yet so no reference) in an initial attempt to better understand the details of these genomes and secretomes.

## 1.2 Genome Assembly

Sequence assembly has been a long-standing application problem in the field of bioinformatics[10]. 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 cycles and storage required for input sequences[10]. In addition to these difficulties, there can be other issues encountered during asssembly due to the nature of the data or genomes themselves, such as low quality base calls for long read data or the inherent content of genomes themselves with repetitive regions.

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 previously mentioned issue of assembly correctness 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 are well assembled, at least based on these metrics[10]. 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 Repeat Identification/Masking and Identification of AT-rich Genomic regions

Repeat identification within assembled genomes is a problem that needs to be considered during the genome annotation process. 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[11]. 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 from technologies such as Nanopore or PacBio to provide a complete picture of these regions when pursuing a new genome assembly project. It is also possible for 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[12]. Once these repetitive regions have been identified, the genome could be masked to exlude these regions in downstream processing if desired, as these regions may be poorly assembled and may result in found genes that do not truly exist in those regions. However, this may not be as common today, as repetetive regions have been shown to contain genes as well[13]. 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.

## 1.4 Gene Finding

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. There are two common methods for gene finding, those methods being *ab initio* methods, where programs search for patterns and gene structures, and similarity or evidence-based searches, which use prior information such as RNAseq data, expressed sequence tags and protein sequences to identify genes within a new genome[14]. Complicating the process more is the introduction of introns and alternative splicing in eukaryotes, making it possible for one gene to have several possible transcripts at the same locus. Examples of *ab initio* methods include tools such as GeneMark-ES[5] and GlimmerHMM[6], while evidence based methods include tools such as Braker2[4], 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 transposable 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 of Wang, Z *et al*[14]. 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.

There are also other aspects of gene finding tools that are important to consider. These include features such as whether or not the gene finders find non-coding RNAs, annotation of 5' and 3' UTR regions, and in the case of ab-initio methods, the assumptions made by the underlying models used for gene finding. These features and others can influence a user's decision on which gene finding tool to consider and will complicate comparative analysis of multiple gene finding tools.

# 2 Research problem

Genome annotation, and more specifically gene finding, 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 its own target applications or use cases. The abundance of gene finding tools raises questions such as which tool to use? Do different tools provide drastically different results? Are there differences in the number and/or of accuracy predicted genes in repetitive genomic regions 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.

As mentioned previously, it is important to understand the features of these gene finders prior to developing an approach for comparing their outputs. With this in mind, the tools selected for this research should cover both *ab initio* and evidence-based methods. The selected tools should also cover the identification of other features such as small RNAs and untranslated regions as it is possible that some gene finders may identify features such as small RNAs without explicitly stating that they do. Said features may also be useful for GIFS researchers working with this data in the future. To cover several of these areas of interest, the following tools are proposed

for this project. Note that other tools are also available for consideration and this set does not have to be final.

- RepeatMasker[15]: a tool used to identify repetititve regions within a genomic sequence.
- Infernal/Rfam[16]: a tool used in combination with a database to analyze genomic sequences for non-coding RNAs. Useful to compare against gene calls from the tools proposed below.
- GeneMark-ES[5]: an *ab initio* HMM-based gene finding tool with a specific option for fungal genomes.
- GenomeThreader[17]: an evidence-based gene finding tool that uses dynamic programming to efficiently identify splice candidates for genes using splice-aware RNASeq alignments as input.
- Braker2: a gene finding tool that uses a hybrid approach combining information from *ab initio* and evidence-based methods. In addition, Braker2 also has the ability to identify UTR regions as well.

A comparative approach or pipeline will then be developed to compare outputs and performance of these gene finding tools. Points of comparison will include gene finding features (proficiencies and deficiencies), efficiency of selected tools (runtimes, memory requirements), requirements of select tools (operating system, additional databases, required librairies etc.), ease of installation, called genes as compared to aligned RNA sequence data, identification of genes in repetitive regions, identification of small RNAs, distributions of gene lengths along with maximum/minimum gene lengths and performance of gene finders in AT-rich 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 newly assembled *Trichoderma* genomes, those being DC1 and Tsth20, as well as other previously assembled *Trichoderma* genomes from NCBI or other sources.

# 3 Deliverables

- High-quality hybrid assemblies of both Tsht20 and DC1.

- Lists of genes for each *Trichoderma* assembly and gene finding tool combination considered
- A consenus or core gene set shared by all genomes considered based on different gene finding tools
- Repeat regions identified in assembled *Trichoderma* genomes.
- A potential list of true positives based on genes which have supporting RNAseq eveidence
- Final comparative tables including the criteria mentioned above along with performance for each gene finder relative to the genome considered.

# 4 Timeline

- Initial genome assemblies: Collection of existing assemblies (1 week), Assembly of DC1/Tsth20 (2-4 weeks)
- Gene finding and annotation results: Identification of gene finding tools (1-2 weeks), Gene finding analysis on selected genomes (1-2 months)
- Downstream analysis of gene finding results (1-2 months)

# 5 References

# References

[1] Zhang, C., Wohlhueter, R., Zhang, H. (2016). Genetically modified foods: A critical review of their promise and problems. *Food Science and Human Wellness*, *5*(3), 116-123.

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

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

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

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

[6] Majoros, W.H., Pertea, M., Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open-source *ab initio* eukaryotic gene-finders. *Bioinformatics. 20*(9), 2878-2879.

[7] 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.

[8] 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.

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

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

[11] 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.

[12] 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.

[13] Slotkin, R.K. (2018). The case for not masking away repetitive DNA. *Mobile DNA. 9*(1), 15.

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

[15] Smit, A., Hubley, R., Green, P. (2013-2015) RepeatMasker Open-4.0. https://www.repeeatmasker.org

[16] Nawrocki, E.P., Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics. 29*, 2933-2935.

[17] Gremme, G., Brendel, V., Sparks, M.E., Kurtz, S. (2005) Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology. 47*, 965-978.