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 17, 2022

Contents

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][3]. 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[7]. 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[7]. 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[7]. 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[6]. 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 Gene Prediction

In progress... Braker2

1.5 Repeat Identification and Masking

In progress...

1.6 Identification of AT-Rich Regions

In progress...

2 Research problem

Genome annotation, in the case of gene finding here, has been a popular computational problem for decades. The identification of possible genes provides other researchers with a valuable resource for future research avenues.

3 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] Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E., Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research. 39*, 339–346.

[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] 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.

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

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

[8] Sperschneider, J., et al. (2016) EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytologist. 2102*, 743-761.

[9] Teufel, F., et al. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology. 40*, 1023-1025.