Evaluation of Gene Finding Tools on *Trichoderma* Genomes - Update

July, 2025

Connor Burbridge

UNIVERSITY OF SASKATCHEWAN

Outline

- Progress update with results
- Progress assessment
- Timeline for completion
- Discussion of next steps

Background

- Gene finding is a critical step in genome annotation.
- *Trichoderma* species are important for agriculture and biotechnology.
- Various tools exist for gene prediction, but their implementations and performance can vary significantly.
- Few studies have compared these tools in fungi, and even fewer in *Trichoderma*.
- In addition, high-quality reference genomes are not available for all species, which makes comparisons difficult.

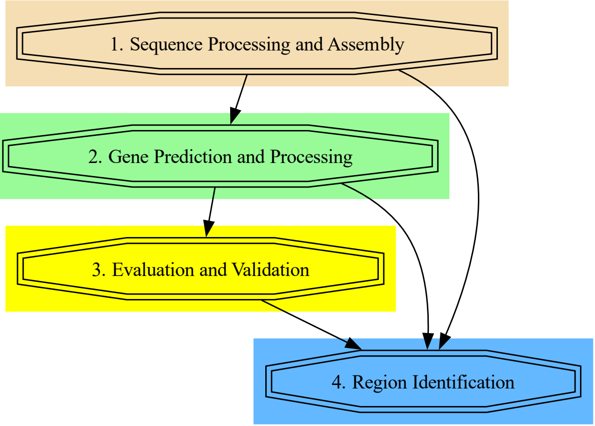
Why *Trichoderma*?

- *Trichoderma* species are ubiquitous in soil and play a significant role in nutrient cycling.
- They are also used in biocontrol and as biofertilizers.
- Known for their production of plant cell wall degrading enzymes, and secondary metabolites.
- Further genomic studies can help in understanding their biology and potential applications.

Novel *Trichoderma* Genomes

- *Trichoderma* species are diverse, with many species not yet fully characterized.
- Recent advances in sequencing technology have made it possible to generate high-quality genomes for these species.
- Two novel genomes of *Trichoderma* species have been sequenced at the Global Institute for food Security
- DC1 and Tsth20, shown to improve drought and salt tolerance when applied to crops.
- These genomes provide an opportunity to evaluate gene finding tools in a comparative context, and contribute to the understanding of *Trichoderma* biology.

Workflow Overview



1. Sequence Processing and Assembly

2. Gene Prediction and Processing

3. Evaluation and Validation

4. Region Identification

Datasets and Tools

- **Datasets:**
  - *Trichoderma* genomes: DC1 and Tsth20.
  - Reference genomes and annotations: *T. reesei, T. harzianum, T.virens*.
  - RNAseq training data from *T. reesei*.
  - Benchmarking Universal Single-Copy Orthologs (BUSCO) fungal database.
  - Protein sequence queries for tblastn from *T. atroviride, Fusarium graminearum, and Saccharomyces cerevisiae*.
- **Tools:**
  - Sequence processing: FastQC, Trimmomatic, Hisat2.
  - Genome assembly: NextDenovo and NextPolish.
  - Gene finding tools: Braker2 and GeneMark-ES.
  - Evaluation tools: BUSCO, tblastn, InterProScan, and custom scripts.

Assembly Results

Name	Total Contigs	Total Length	Largest Contig	GC%	N50	L50
DC1	8	38.6 Mb	11.49 Mb	47.97	5.69 Mb	3
Tsth20	7	41.58 Mb	8.02 Mb	47.33	6.52 Mb	3
T. harzianum	532	40.98 Mb	4.08 Mb	47.61	2.41 Mb	7
T. virens	93	39.02 Mb	3.45 Mb	49.25	1.83 Mb	8
T. reesei	77	33.39 Mb	3.75 Mb	52.82	1.21 Mb	9

- DC1 and Tsth20 assemblies are high quality with few contigs in comparison to other *Trichoderma* assemblies.
- Input sequences and assemblies show a bimodal distribution of GC content.
- AT-rich sequence content may be related to transposable elements and repeat-induced mutations, which may be of interest in secondary metabolite production.

Gene Finding Results

Assembly	Braker2		GeneMark		RefSeq	
	Genes	CDS	Genes	CDS	Genes	CDS
DC1	8546	8637	11353	11353	N/A	N/A
Tsth20	8784	8858	12362	12362	N/A	N/A
T. reesei	9659	10175	9196	9196	9109	9118
T. harzianum	8314	8385	12164	12164	14269	14090
T. virens	7801	7863	11866	11866	12405	12406

- Braker2 predicts more genes and coding sequences than GeneMark and RefSeq in *T. ressei*, but fewer in other assemblies.
- GeneMark and RefSeq predictions are similar, except in *T. harzianum*, where RefSeq predicts 17% more genes.
- GeneMark only predicts one coding sequence per gene, while Braker2 and RefSeq predict multiple coding sequences.

Examining C