



Evaluation of Gene Finding Tools When Applied to *Trichoderma* Genomes - Update

July, 2025



Connor Burbridge

Outline

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



What is *Trichoderma*?

- *Trichoderma* is a genus of filamentous that is ubiquitous in soil and plays 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.

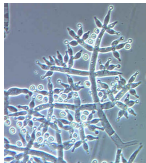


Figure: *T. harzianum*



Figure: *Trichoderma* colony

Novel *Trichoderma* Genomes

- *Trichoderma* species are diverse, with many species not yet fully characterized.
- DC1 and Tsth20, shown to improve drought and salt tolerance when applied to crops, and have been shown to breakdown hydrocarbons in soils.
- Recent advances in sequencing technology have made it possible to generate high-quality genomes for these species.



Motivation

- To better understand the biological mechanisms at work in *Trichoderma* spp., we first need to understand their genomes by identifying genes within them.
- 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, increased accessibility to high-quality sequencing services has led to the generation of many new *Trichoderma* genomes, including DC1 and Tsth20.
- These genomes provide an opportunity to evaluate gene finding tools in a comparative context, and contribute to the understanding of *Trichoderma* biology.

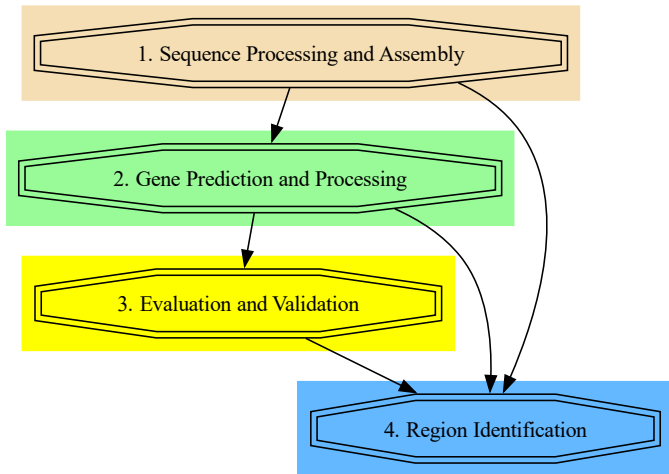


Research Objectives

- Assemble and evaluate novel assemblies of DC1 and Tsth20.
- Apply gene finders Braker2 and GeneMark in DC1, Tsth20, and three other RefSeq assemblies - *T. reesei*, *T. harzianum* and *T. virens*.
- Compare gene finding tools based on relevant criteria, including:
 - Proportions of gene lengths predicted.
 - Presence of functional domains and closely related protein sequences.
 - Presence of genes in AT-rich sequence.
 - Agreement of gene finders on start and stop positions of a gene.



Workflow Overview



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
<i>T. harzianum</i>	532	40.98 Mb	4.08 Mb	47.61	2.41 Mb	7
<i>T. virens</i>	93	39.02 Mb	3.45 Mb	49.25	1.83 Mb	8
<i>T. reesei</i>	77	33.39 Mb	3.75 Mb	52.82	1.21 Mb	9

- DC1 and Tsth20 assemblies are of 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

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

- Braker2 predicts more genes and coding sequences than GeneMark and RefSeq in *T. reesei*, 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 Coding Sequence Lengths

Genome	Tool #1	Tool #2	<i>P</i> -value
DC1	Braker2	GeneMark	0.999
Tsth20	Braker2	GeneMark	0.965
<i>T. reesei</i>	Braker2	GeneMark	$P < 0.005$
<i>T. reesei</i>	GeneMark	RefSeq	0.002
<i>T. reesei</i>	Braker2	RefSeq	$P < 0.005$
<i>T. harzianum</i>	Braker2	GeneMark	0.863
<i>T. harzianum</i>	GeneMark	RefSeq	$P < 0.005$
<i>T. harzianum</i>	Braker2	RefSeq	$P < 0.005$
<i>T. virens</i>	Braker2	GeneMark	0.635
<i>T. virens</i>	GeneMark	RefSeq	$P < 0.005$
<i>T. virens</i>	Braker2	RefSeq	$P < 0.005$



BUSCO Results

Strain	Complete	Single	Duplicated	Fragmented	Missing
DC1	99.5	80.2	19.3	0.1	0.4
Tsth20	99.9	81.7	18.2	0.0	0.1
<i>T. harzianum</i>	99.7	80.2	19.5	0.0	0.3
<i>T. virens</i>	99.8	79.0	20.8	0.1	0.1
<i>T. reesei</i>	99.9	85.5	14.4	0.1	0.0

(a) Braker2

Strain	Complete	Single	Duplicated	Fragmented	Missing
DC1	99.2	98.8	0.4	0.3	0.5
Tsth20	99.8	99.1	0.7	0.0	0.2
<i>T. harzianum</i>	99.6	98.9	0.7	0.0	0.4
<i>T. virens</i>	99.7	99.2	0.5	0.1	0.2
<i>T. reesei</i>	99.6	99.5	0.1	0.0	0.4

(b) GeneMark

Strain	Complete	Single	Duplicated	Fragmented	Missing
<i>T. harzianum</i>	99.9	99.2	0.7	0.0	0.1
<i>T. virens</i>	99.5	98.8	0.7	0.3	0.2
<i>T. reesei</i>	99.8	99.5	0.3	0.0	0.2

(c) RefSeq

Agreement of Gene Predictions

- In DC1 and Tsth20, Braker2 and GeneMark tend to agree on the start and stop positions of genes when they predict the same gene. Both gene finders have a large portion of singleton predictions.

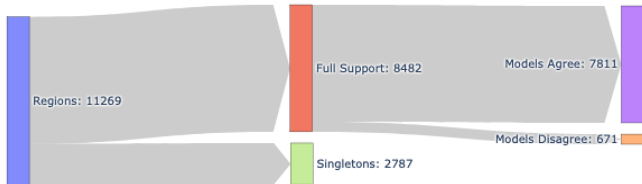


Figure: DC1

- In *T. reesei*, Braker2, GeneMark and RefSeq tend to disagree more on the start and stop positions of genes when they predict the same gene.
- Disagreement is more pronounced in *T. harzianum* and *T. virens*, where there are fewer genes with supporting predictions from each gene finder.

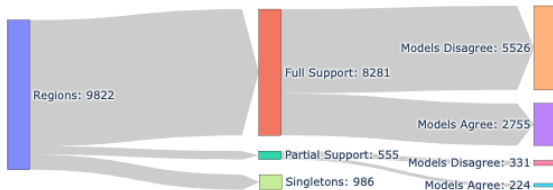


Figure: *T. reesei*

InterProScan Results

Assembly	Braker2	GeneMark	RefSeq
DC1	73%	74.1%	N/A
Tsth20	73.3%	74.1%	N/A
<i>T. reesei</i>	72.4%	76%	76.4%
<i>T. harzianum</i>	73.8%	74.5%	66.11%
<i>T. virens</i>	74.7%	74.8%	73.2%

- All three gene finders predict a similar proportion of genes with functional domains, except in the case of *T. harzianum*, where RefSeq predicts a significantly lower proportion of genes with functional domains.



tblastn Results

Subject	Query	Braker2	GeneMark	RefSeq
DC1	<i>T. atroviride</i>	5902	4679	N/A
DC1	<i>F. graminearum</i>	4955	4114	N/A
DC1	<i>S. cerevisiae</i>	2105	1850	N/A
<i>T. reesei</i>	<i>T. atroviride</i>	5072	5174	4989
<i>T. reesei</i>	<i>F. graminearum</i>	4577	4685	4529
<i>T. reesei</i>	<i>S. cerevisiae</i>	2055	2114	2022
<i>T. harzianum</i>	<i>T. atroviride</i>	6363	4611	6835
<i>T. harzianum</i>	<i>F. graminearum</i>	5659	4198	5982
<i>T. harzianum</i>	<i>S. cerevisiae</i>	2424	1963	2560

- Regions with Braker2 and RefSeq gene predictions consistently have more tblastn hits than GeneMark, with the exception of *T. reesei*.



Gene Predictions in AT-rich Sequence

Assembly	Braker2	GeneMark	RefSeq
DC1	31	11	N/A
Tsth20	11	2	N/A
<i>T. reesei</i>	39	48	107
<i>T. harzianum</i>	81	30	154
<i>T. virens</i>	21	8	20

- Very few genes are predicted in AT-rich regions of DC1, Tsth20 and *T. virens* assemblies in comparison to the *T. reesei* and *T. harzianum* assemblies.
- Possibly due to higher quality assemblies? Although *T. virens* is still fragmented.
- A two-sided binomial test confirms that gene finders do not predict the same proportion of genes in AT-rich sequence as they do in normal genomic sequence.

Conclusions

- In terms of gene finding performance based on several criteria, RefSeq predictions are typically the best, while GeneMark performs the worst.
- Braker2 performs well in *T. reesei*, but not as well in the other assemblies. Users should be careful with the training data selected for Braker2, as it can have a significant impact on the results.
- RefSeq predictions are not always available, so Braker2 can be used if appropriate training is available, but GeneMark can be used as a fallback.



Future Work

- **Explore the landscape of secondary metabolite gene clusters in *Trichoderma* genomes further.**
- Extend the analysis to include more *Trichoderma* species, more gene finding tools, and more datasets.
- Investigate further methods for validation of gene predictions.
- Continue to analyze the newly assembled genomes of DC1 and Tsth20.



Timeline for Completion

- Initial draft of thesis complete - end of August 2025
- Revisions and changes - Sept./Oct. 2025
- Submit thesis to committee - Early November 2025
- Schedule defence for December 2025 or January 2026
- Time for extra revisions post-defence - January/early February 2026
- Completion - End of February 2026



Current Status

- Introduction - 70%
- Background - 85%
- Methods - 90%
- Research Questions - 85%
- Results and Discussion - 80(?)%
- Conclusions and Future Work - 75%
- What to focus on next?



Acknowledgements

- My supervisors for their committed support through COVID and other challenges.
- The Global Institute for Food Security for providing the data for this project as well as a portion of my funding.
- My committee members for their feedback and support.



Datasets and Tools

- **Datasets:**

- Novel *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.





Category	Braker2	GeneMark	RefSeq
Availability	3	3	0
Ease of install	1	2	0
Ease of use	3	3	0
<hr/>			
# of genes predicted	0	3	3
# of transcripts predicted	3	0	2
Predicts shortest genes	2	1	0
Predicts more shorter genes	1	0	3
BUSCO Performance	2	1	3
Performance in AT-rich sequence	2	1	3
Predictions with InterProScan support	3	3	3
Final Score (Publicly Available)	20	17	N/A
Final Score (Ignoring Availability)	13	9	17



Image Credits

- *T. harzianum* image:
<https://en.wikipedia.org/wiki/Trichoderma>
- *Trichoderma colony* image:<https://biocontrol.entomology.cornell.edu/pathogens/trichoderma.php>

