Thesis Defence: Evaluation of Gene Finding Tools When Applied to *Trichoderma* Genomes

January, 2026

Connor Burbridge

UNIVERSITY OF SASKATCHEWAN

Outline

- *Trichoderma* fungi.
- Novel *Trichoderma* genomes.
- Motivation.
- Research objectives.
- Workflow overview.
- Results.
- Conclusions and recommendations.
- Questions.

What is *Trichoderma*?

- *Trichoderma* is a genus of filamentous fungi 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.

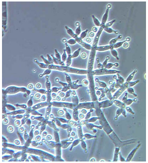


Figure: *T. harzianum*

Figure: *Trichoderma* colony

Novel *Trichoderma* Genomes

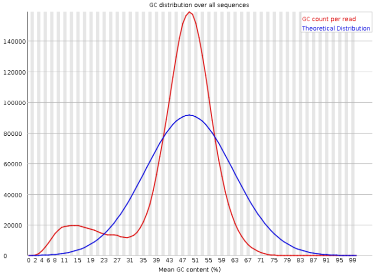
- *Trichoderma* species are diverse, with many species not yet fully characterized.
- DC1 and Tsth20, shown to improve drought and salt tolerance when applied to crops, and have been shown to breakdown hydrocarbons in soils.
- Recent advances in sequencing technology have made it possible to generate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

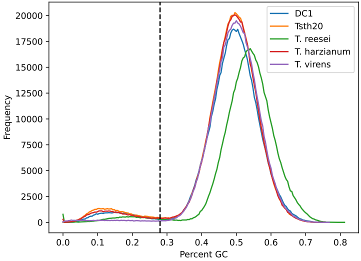
Motivation

- To understand *Trichoderma* biology, we must identify genes within their genomes.
- Gene prediction tools vary significantly in implementation and performance.
- **Few comparative studies exist for these tools in fungi, particularly in *Trichoderma*.**
- New high-quality *Trichoderma* genomes (DC1 and Tsth20) enable comparative evaluation of gene finding tools.

Gene Finding is Still Challenging

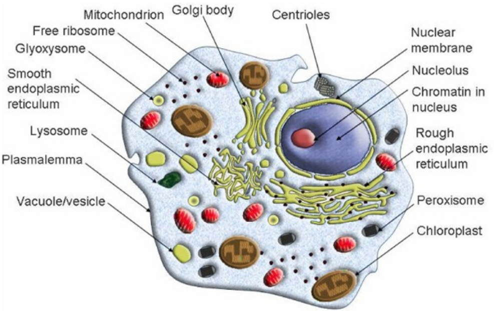
- We've been finding genes for decades. What's going on?
- Exons and introns as well as start and stop positions make gene finding tricky
- Inherent properties of genomes also complicate gene finding.



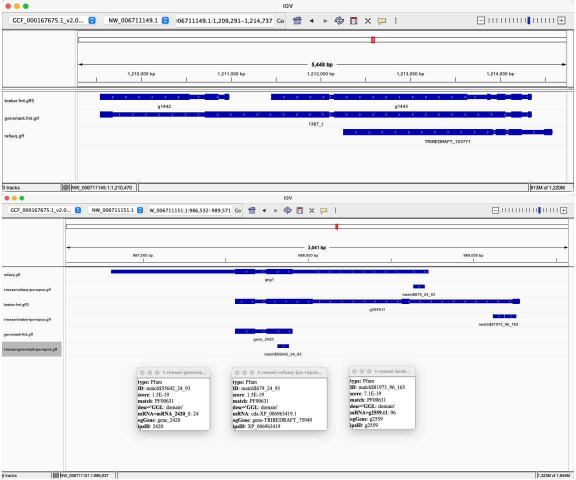


Genes are Sensitive

- Minor changes in start/stop positions can have major impacts on predicted protein sequences.
- Ex. The subcellular localization process relies heavily on N-terminal sequences.



Uncertainty in Gene Predictions



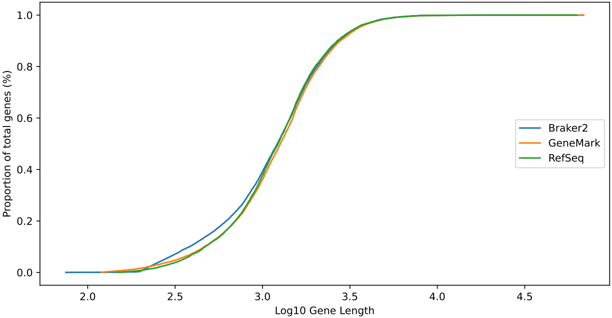
type: Pfam
ID: match$55642_24_93
score: 1.5E-19
match: PF00631
desc='GGL: domain'
mRNA=mRNA_2420_1: 24
ogGene: gene_2420
ipsID: 2420

type: Pfam
ID: match$679_24_93
score: 1.5E-19
match: PF00631
desc='GGL: domain'
mRNA: cds-XP_006963419.1
ogGene: gene-TRIREDRAFT_75949
ipsID: XP_006963419

type: Pfam
ID: match$81973_96_165
score: 7.1E-19
match: PF00631
desc='GGL: domain'
mRNA=g2559.t1: 96
ogGene: g2559
ipsID: g2559

Small but Significant Details

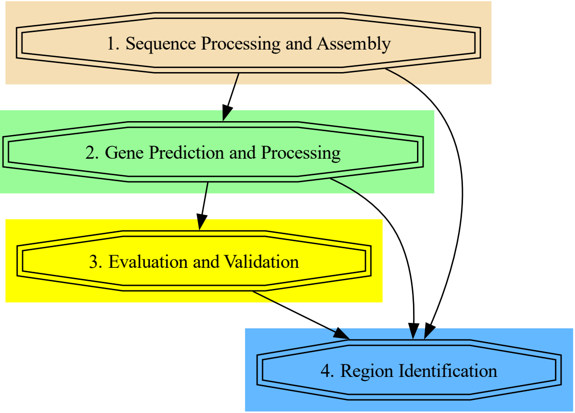
- Before diving too deep, how can we objectively confirm that gene finding tools predict different genes?
- Kolmogorov-Smirnov (KS) tests were performed on gene length distributions from preliminary gene predictions.



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.
- 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



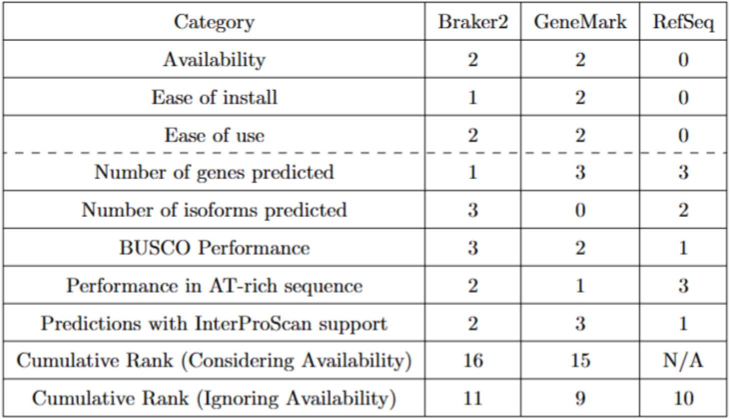
1. Sequence Processing and Assembly

2. Gene Prediction and Processing

3. Evaluation and Validation

4. Region Identification

Overview of Results



Category	Braker2	GeneMark	RefSeq
Availability	2	2	0
Ease of install	1	2	0
Ease of use	2	2	0
Number of genes predicted	1	3	3
Number of isoforms predicted	3	0	2
BUSCO Performance	3	2	1
Performance in AT-rich sequence	2	1	3
Predictions with InterProScan support	2	3	1
Cumulative Rank (Considering Availability)	16	15	N/A
Cumulative Rank (Ignoring Availability)	11	9	10

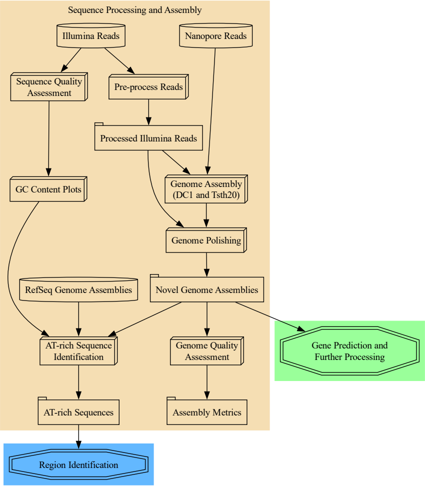
Key Findings

- Overall performance of gene finders are similar, but vary by specific criteria.
- Braker2 generally performs best, particularly when relevant training data is available.
- Candidate secondary metabolite gene clusters were found in both DC1 and Tsth20 assemblies.
- DC1 and Tsth20 assemblies are of high quality, providing a solid foundation for future genomic studies.

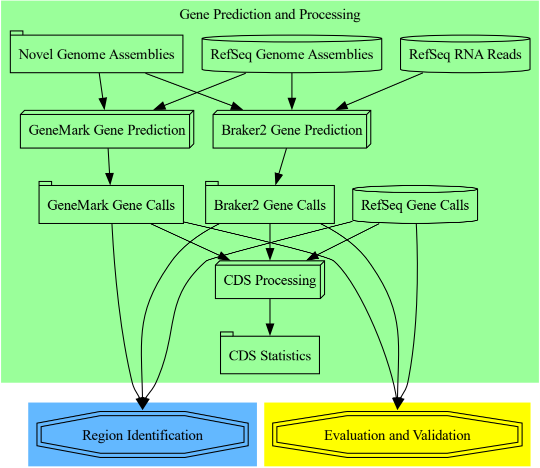
Acknowledgements

- 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.

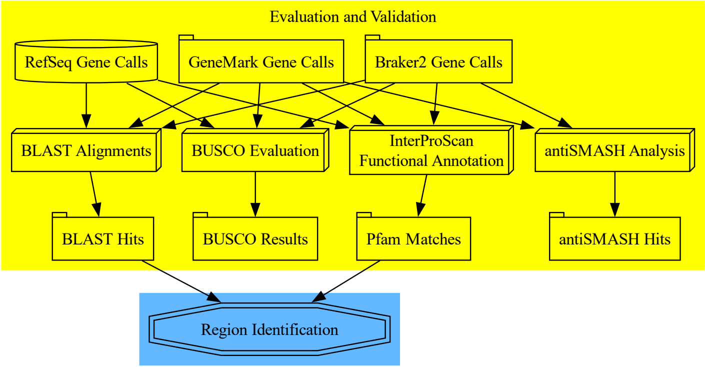
Workflow Details: Assembly



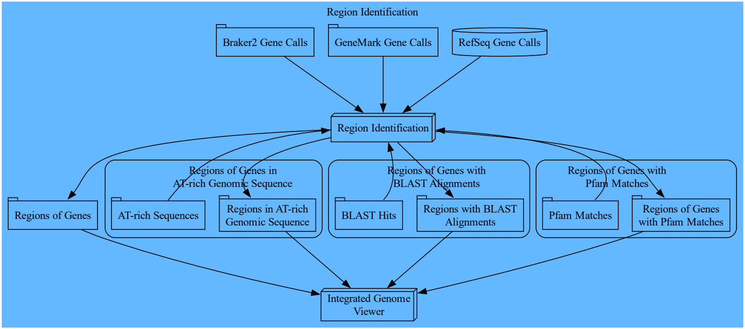
Workflow Details: Gene Finding



Workflow Details: Evaluation



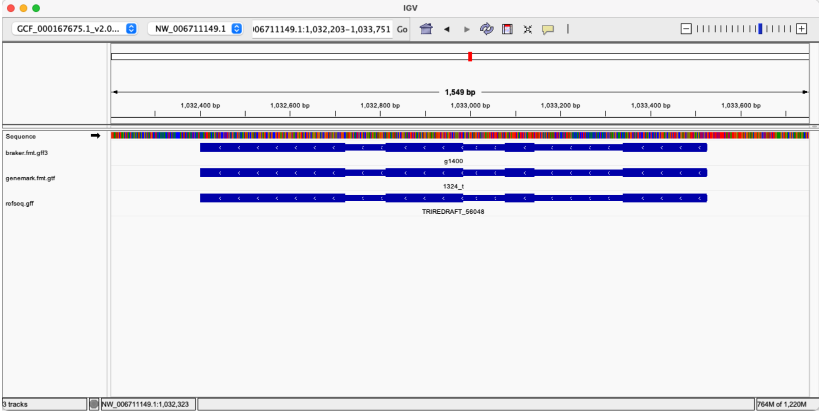
Workflow Details: Regions


Region Identification
Braker2 Gene Calls
GeneMark Gene Calls
RefSeq Gene Calls
Region Identification
Regions of Genes in AT-rich Genomic Sequence
Regions of Genes with BLAST Alignments
Regions of Genes with Pfam Matches
Regions of Genes
AT-rich Sequences
Regions in AT-rich Genomic Sequence
BLAST Hits
Regions with BLAST Alignments
Pfam Matches
Regions of Genes with Pfam Matches
Integrated Genome Viewer

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.

Regions



IGV
GCF_000167675.1_v2.0...
NW_006711149.1
006711149.1:1,032,203-1,033,751
Go
1,549 bp
1,032,400 bp
1,032,600 bp
1,032,800 bp
1,033,000 bp
1,033,200 bp
1,033,400 bp
1,033,600 bp
Sequence
braker.fmt.gff3
g1400
genemark.fmt.gtf
1324_t
refseq.gff
TRIREDRAFT_56048
3 tracks
NW_006711149.1:1,032,323
764M of 1,220M

Image Credits

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