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

January, 2026

Connor Burbridge

University of Saskatchewan

Outline

- ***Trichoderma* fungi.**
- Motivation.
- Research objectives.
- Workflow overview.
- Results highlights.
- Conclusions and future work.

What is *Trichoderma*?

- Filamentous fungi common in soil; key role in nutrient cycling.
- Used in biocontrol and as biofertilizers.
- Produce plant cell wall degrading enzymes and **secondary metabolites**.
- Future studies reveal their biology and potential applications.

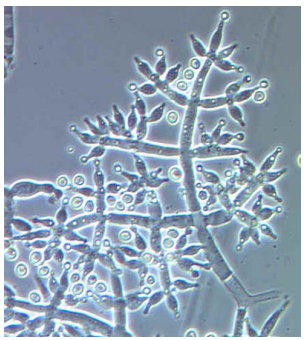


Figure: *T. harzianum* [1]

Figure: *Trichoderma* colony [2]

[1] Source: https://en.wikipedia.org/wiki/Trichoderma

[2] Source: https://biocontrol.entomology.cornell.edu/pathogens/trichoderma.php

Novel *Trichoderma* Genomes

- *Trichoderma* species are diverse, with many species not yet fully characterized.
- DC1 and Tsth20 both improve salt and drought tolerance.
- Tsth20 shows potential as a bioremediation agent.
- Recent advances in sequencing technology have made it possible to generate high-quality genomes for these species.
- Joint effort by Brendan Ashby, the Kaminskyj lab, and the Global Institute for Food Security.

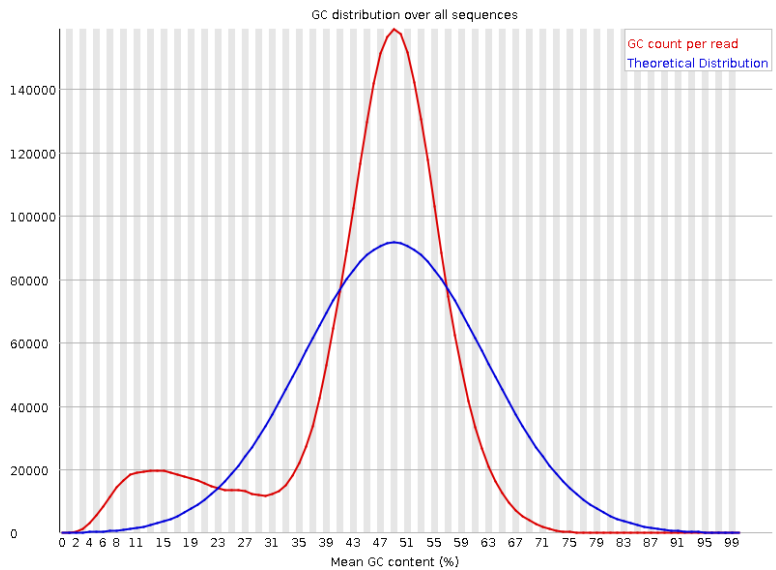
- *Trichoderma* fungi.
- **Motivation.**
- Research objectives.
- Workflow overview.
- Results highlights.
- Conclusions and future work.

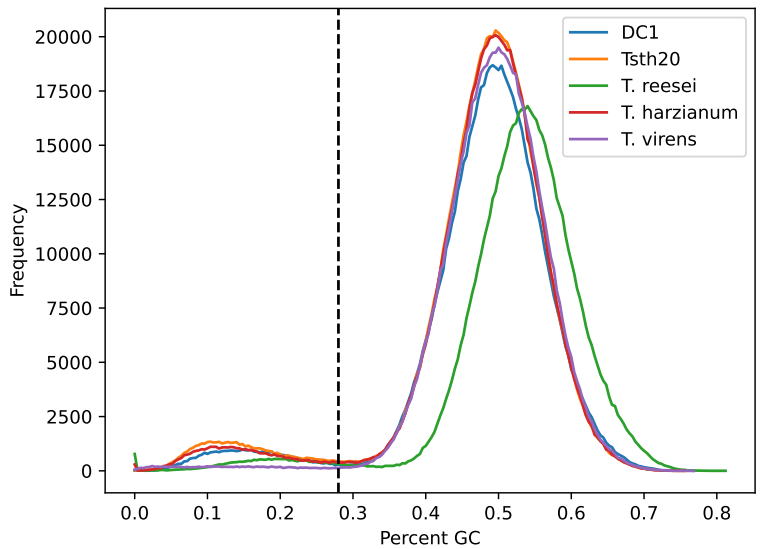
Motivation

- To understand *Trichoderma* biology, we must identify genes within their genomes.
- Gene prediction tools vary significantly in implementation and performance.
- Novel genomes with novel biology require gene annotation to connect phenotypes to genotypes.
- New high-quality *Trichoderma* genomes (DC1 and Tsth20) enable this comparative evaluation of gene finding tools.

Gene Finding is Still Challenging

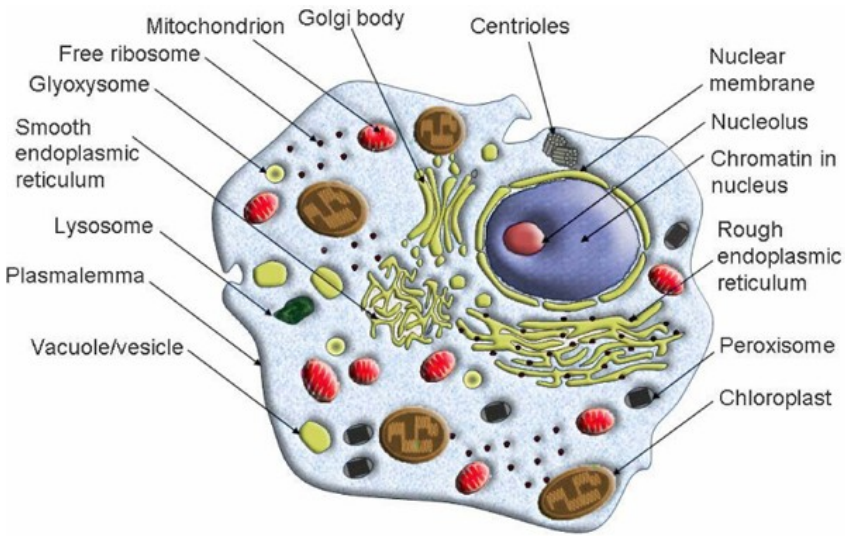
- We've been finding genes for decades. What's going on?
- Exons and introns as well as start and stop positions make gene finding tricky.
- Stop codons in AT-rich regions are highly prevalent.
- Inherent properties of genomes also complicate gene finding.



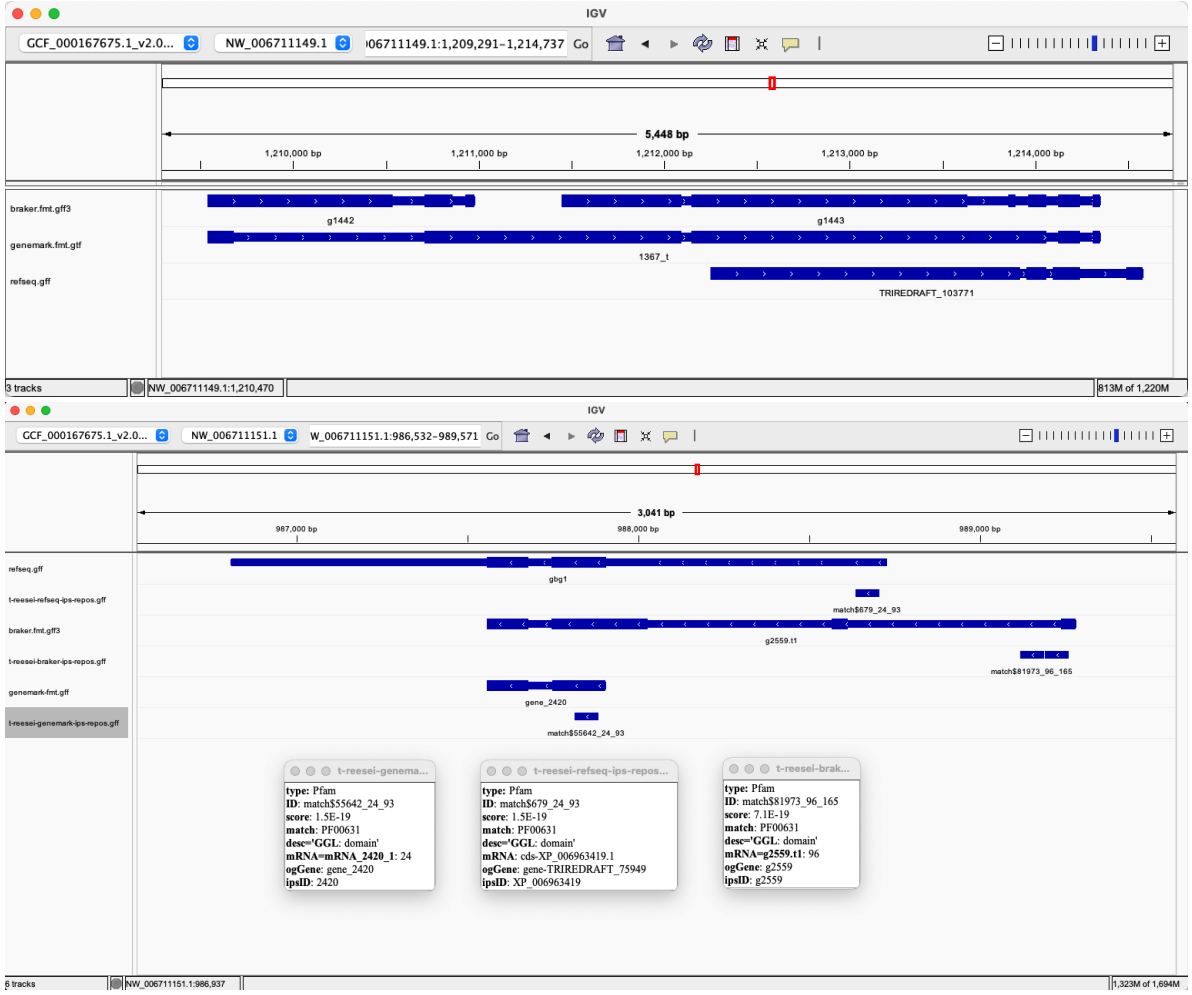


The Details Matter

- 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.
- Links back to cell wall degrading enzymes and secondary metabolites.

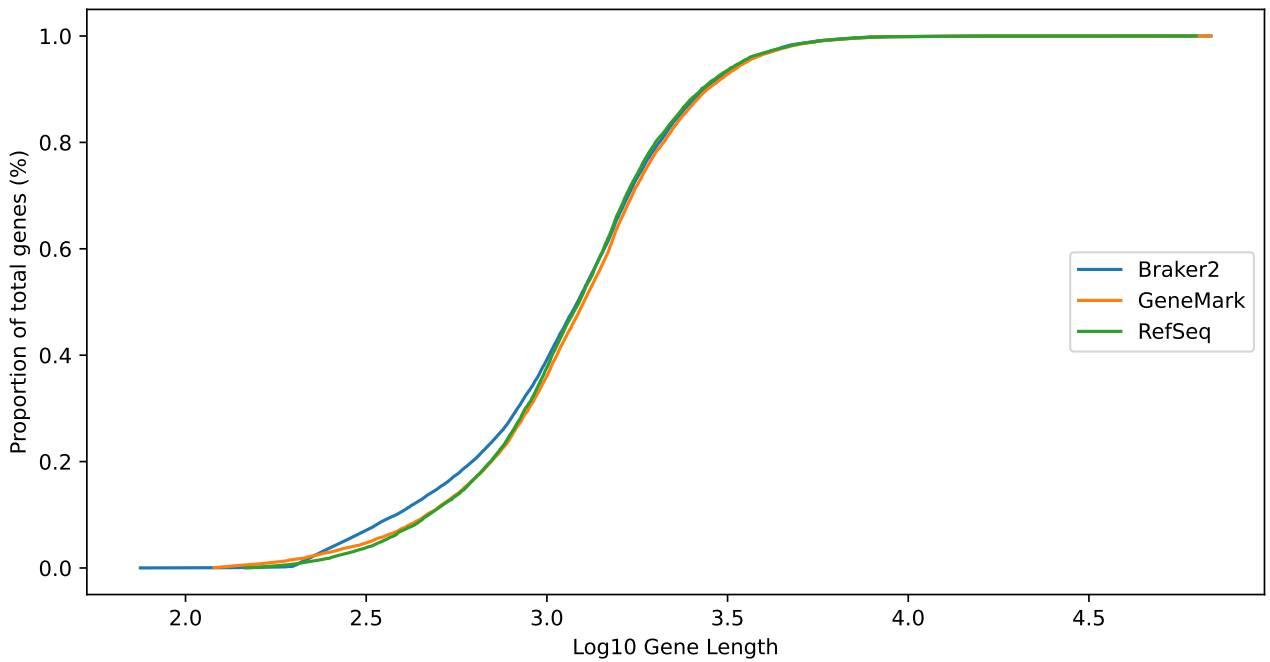


Variability in Gene Predictions



Small but Significant Details

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



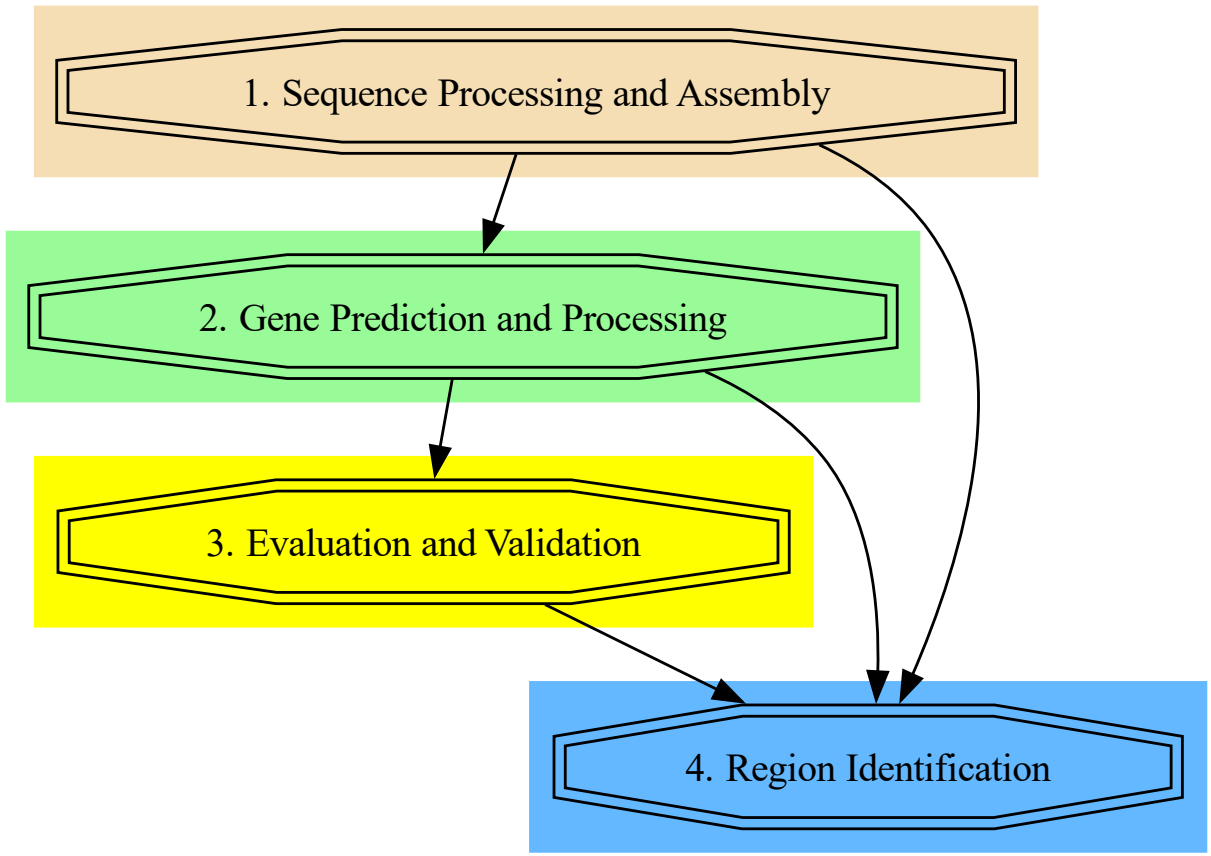
- *Trichoderma* fungi.
- Motivation.
- **Research objectives.**
- Workflow overview.
- Results highlights.
- Conclusions and future work.

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:
  - Number of genes and isoforms predicted.
  - Presence of conserved single-copy orthologs (BUSCOs).
  - 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.

- *Trichoderma* fungi.
- Motivation.
- Research objectives.
- **Workflow overview.**
- Results highlights.
- Conclusions and future work.

Workflow Overview



- *Trichoderma* fungi.
- Motivation.
- Research objectives.
- Workflow overview.
- **Results highlights.**
- Conclusions and future work.

Assemblies of DC1 and Tsth20

- Both assemblies are of high quality.
- Few contigs, which is good for gene finding.
- BUSCO analysis sees similar benefits.
- RefSeq assemblies are highly fragmented in comparison.

Strain	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

AT-rich Sequence Analysis

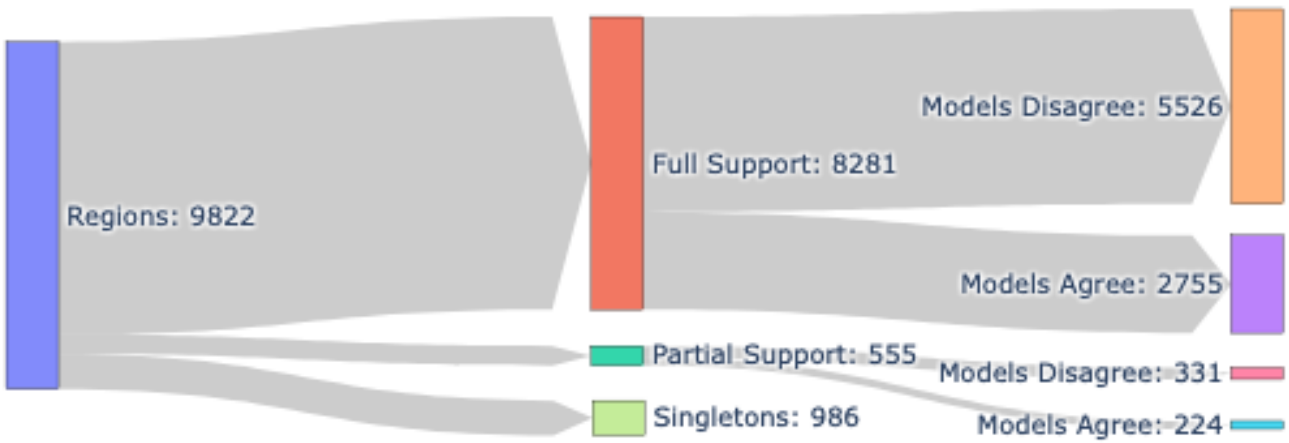
- GC content varies across genomes.
- Gene finders struggle in AT-rich regions.

Genome	Total Length (bp)	Low GC Segments	Low GC Length (bp)	Proportion (%)
DC1	38,616,239	610	2,064,202	5.35
Tsth20	41,588,851	245	891,216	2.14
T. reesei	33,395,713	689	1,269,347	3.80
T. harzianum	40,980,648	1,311	2,527,773	6.17
T. virens	39,022,666	96	162,356	0.42

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

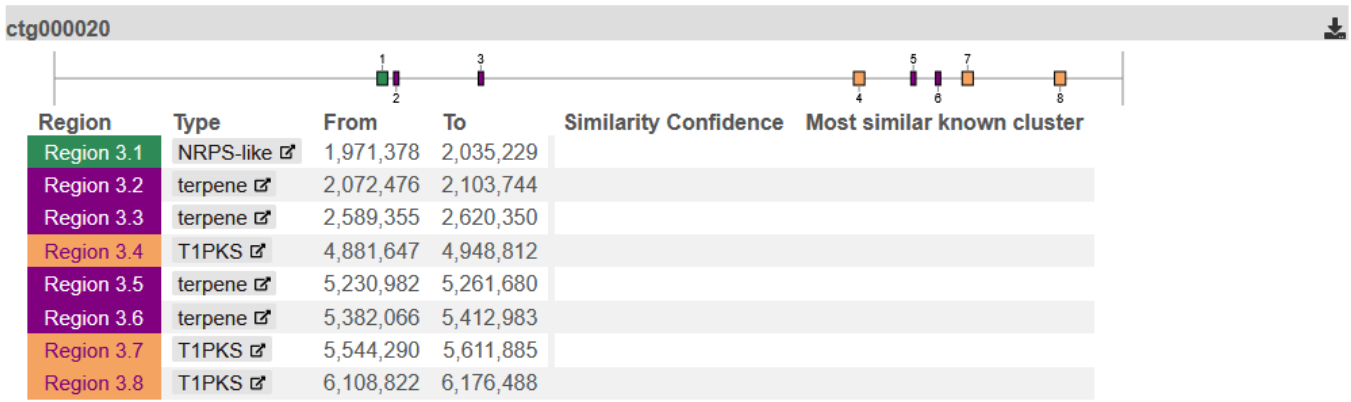
Regions of Gene Predictions

- Agreement of gene finders assessed by identifying regions of gene predictions.
- Regions classified based on agreement of start and stop positions.
- Transitive closure used to group overlapping predictions.



Secondary Metabolite Gene Clusters

- antiSMASH used to identify candidate secondary metabolite gene clusters.
- Tsth20 contains multiple candidate clusters.
- 35 candidates from Braker2 predictions and 58 from GeneMark.
- Limited exploration of predicted clusters performed.



Region	Type	From	To	Similarity Confidence	Most similar known cluster
Region 3.1	NRPS-like	1,971,378	2,035,229		
Region 3.2	terpene	2,072,476	2,103,744		
Region 3.3	terpene	2,589,355	2,620,350		
Region 3.4	T1PKS	4,881,647	4,948,812		
Region 3.5	terpene	5,230,982	5,261,680		
Region 3.6	terpene	5,382,066	5,412,983		
Region 3.7	T1PKS	5,544,290	5,611,885		
Region 3.8	T1PKS	6,108,822	6,176,488		

Ranking Gene Finders

Category	Braker2	GeneMark	RefSeq
Availability	2	2	N/A
Ease of install	1	2	N/A
Ease of use	2	2	N/A
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

- *Trichoderma* fungi.
- Motivation.
- Research objectives.
- Workflow overview.
- Results highlights.
- **Conclusions and future work.**

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.
- Explored a reference-agnostic approach to gene prediction comparisons.

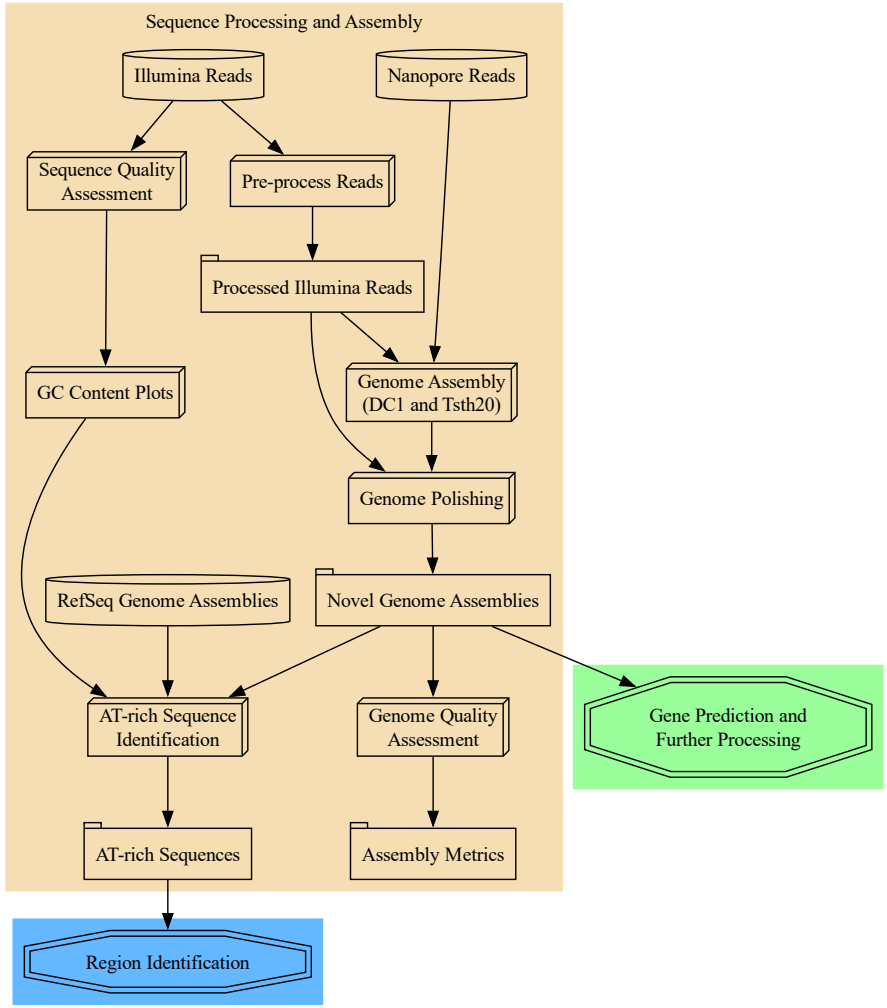
Future Work

- Explore additional gene finding tools.
- Experiment with different types of training data for Braker2.
- Extend analysis of identified regions of gene predictions.
- Explore secondary metabolite gene clusters in more detail.
- Expand the number of *Trichoderma* genomes analyzed

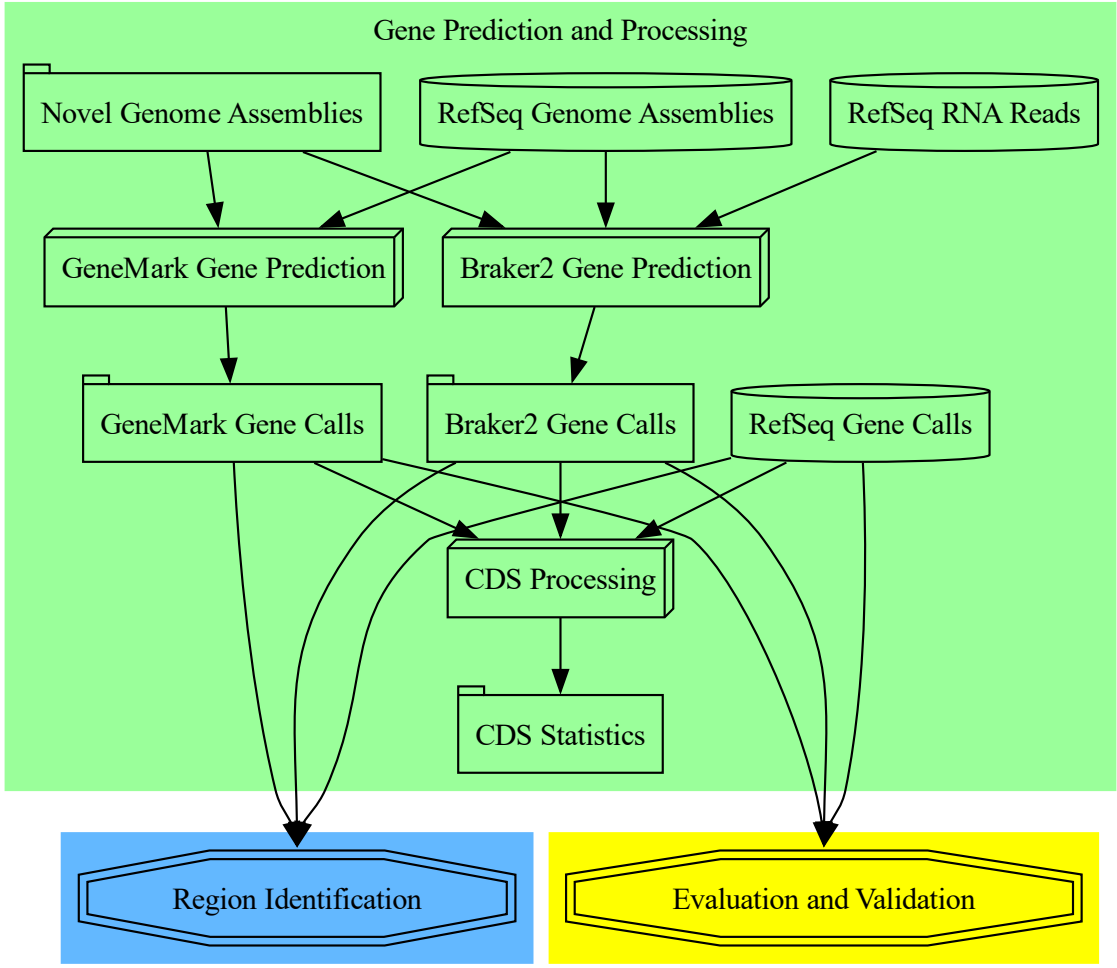
Acknowledgements

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

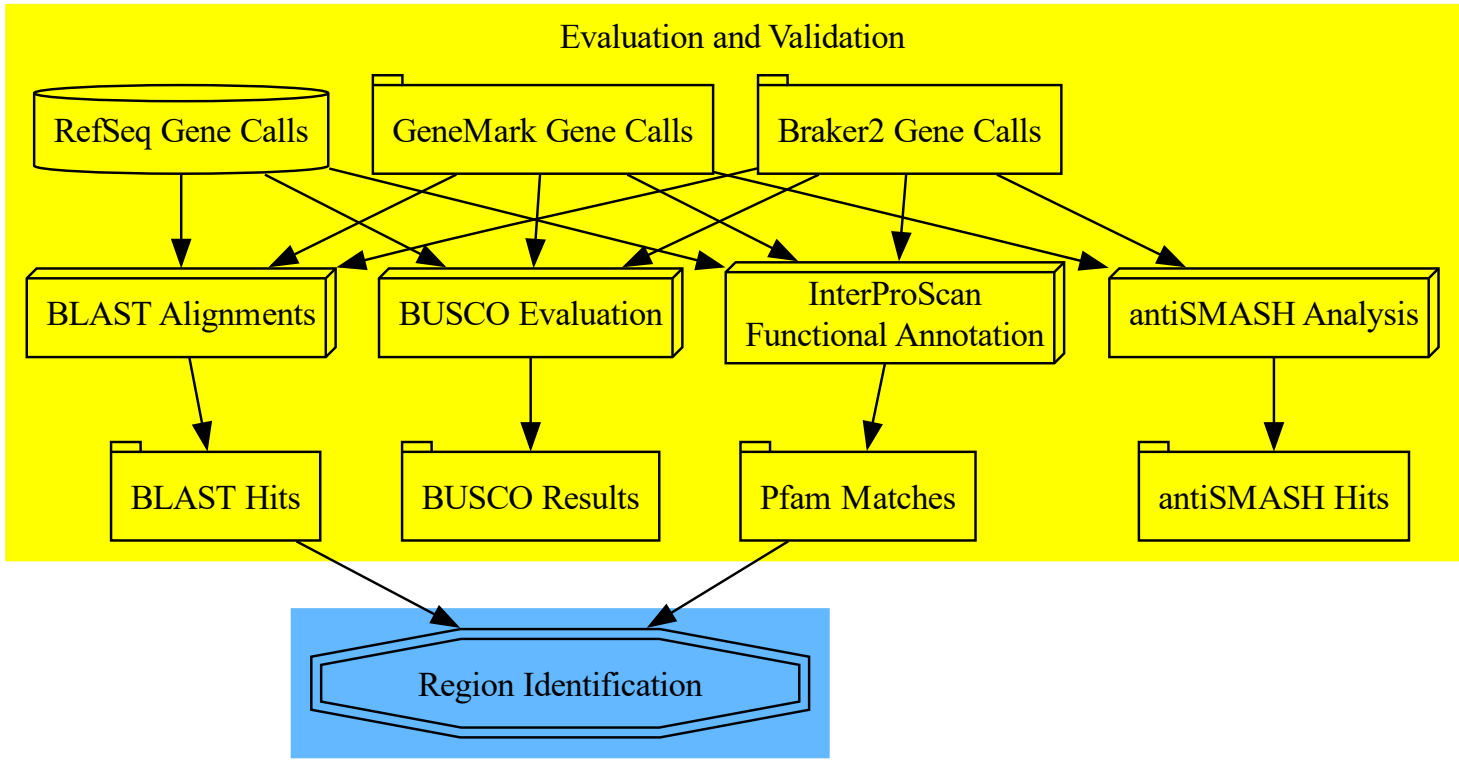
Workflow Details: Assembly



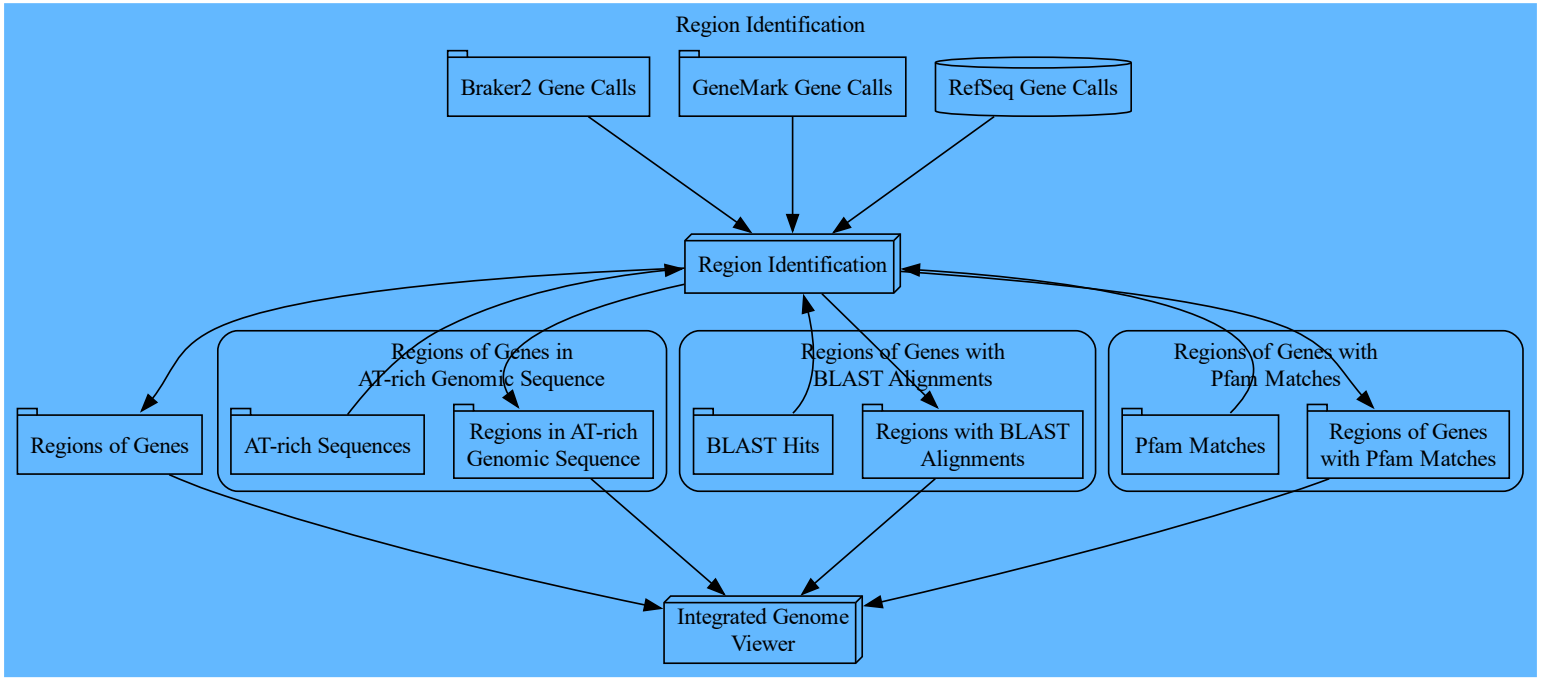
Workflow Details: Gene Finding



Workflow Details: Evaluation



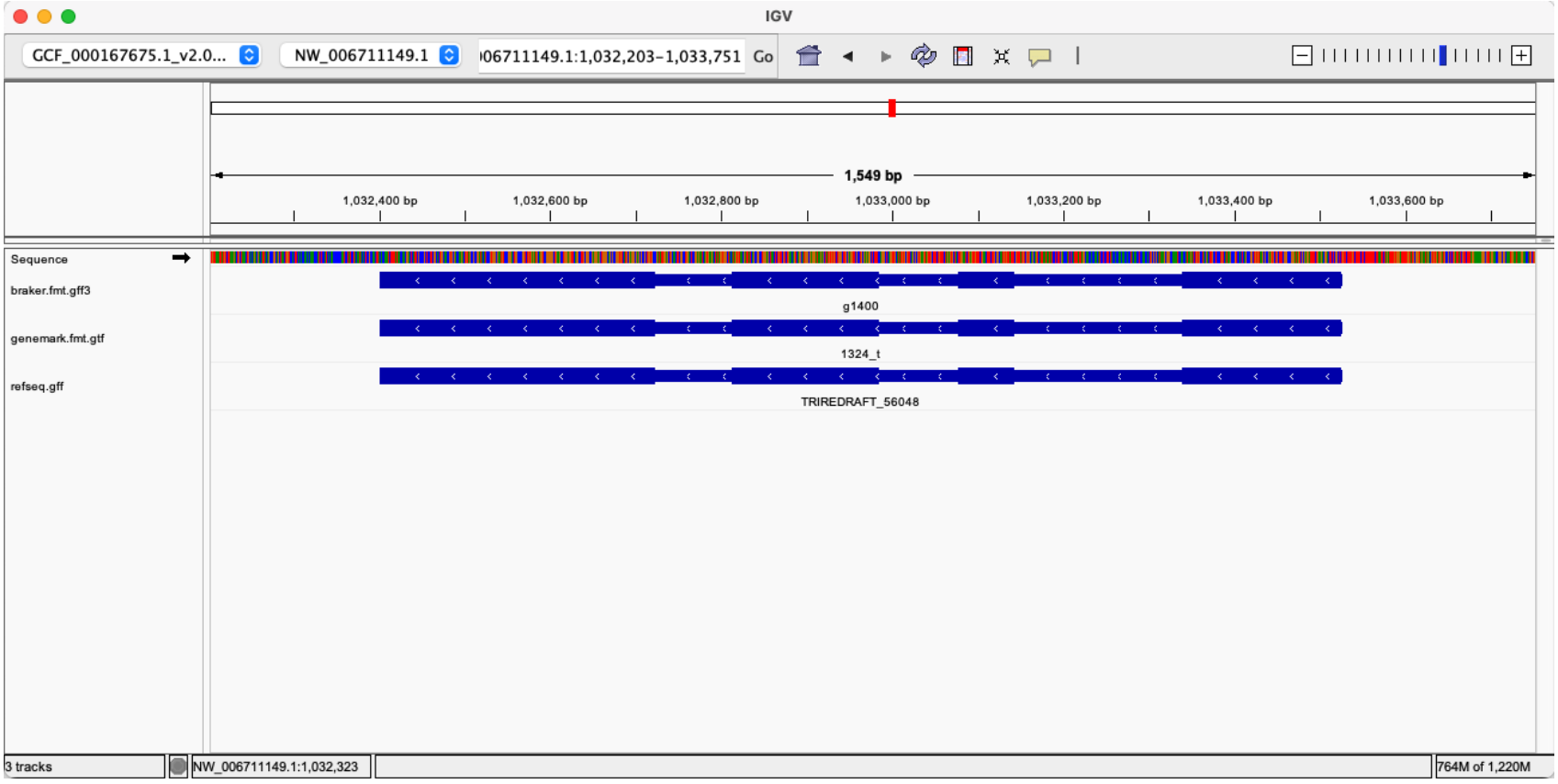
Workflow Details: Regions



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
06711149.1:1,032,203-1,033,751
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