Comparison of Gene Finding Tools in the Context of *Trichoderma* Genomes

Committee members: Dave Schneider, Tony Kusalik, Matthew Links, Leon Kochian

Connor Burbridge

May 31, 2023

Background: *Trichoderma*

What is *Trichoderma*?

- *Trichoderma* is an opportunistic symbiotic fungi, which can colonize the roots of plants
- *Trichoderma* strains have been shown to provide several benefits to the host plant it colonizes, those generally being:
  - Increased resistance to abiotic and biotic stressors
  - Facilitating nutrient uptake
  - Increased germination rates
- These benefits have resulted in *Trichoderma* being used in manufacturing processes for antibiotics and other materials

Background: Previous GIFS Work

Two strains have been sequenced in previous work within GIFS:

- These strains have been named DC1 and Tsth20
- Strains from the prairie regions of Canada, including Alberta and Saskatchewan
- One of these strains has been shown to improve crop tolerance to soils with high salt content. The other shows potential as a bioremediation agent for soils contaminated with hydrocarbons
- How exactly do these processes work? Which genes are included in these processes?
- To answer these questions, both strains were sequenced with Illumina and Nanopore technologies

Research Problem

These sequenced strains offer an opportunity to assemble and annotate them:

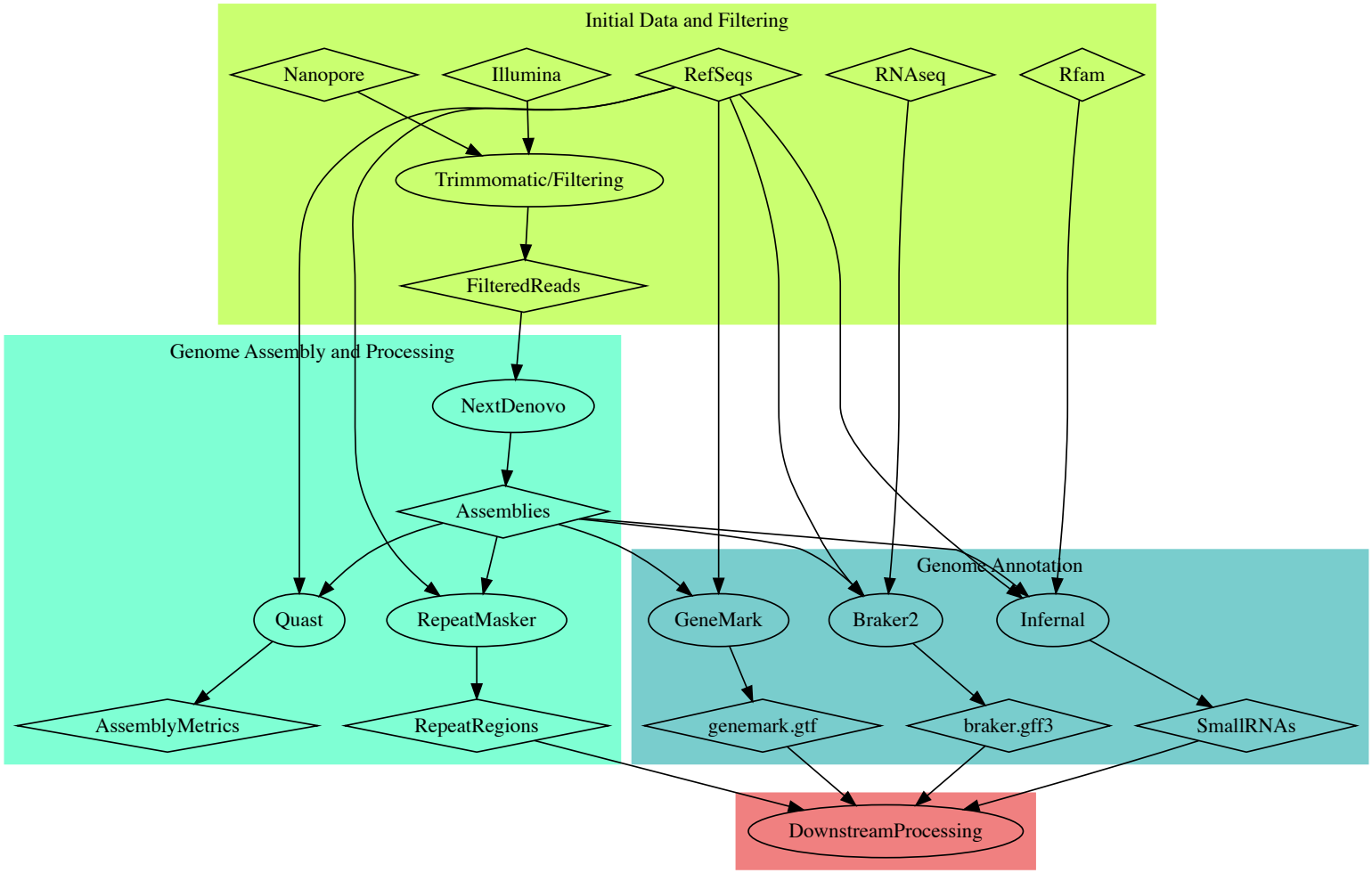
- Genome assembly is 'relatively' straight-forward
- **However, the choice of a tool for gene finding or annotation is uncertain**
- There has been relatively little comparative analysis for gene finding tools in fungi, and even fewer for *Trichoderma*
- **This raises questions. How do different gene finding tools perform in fungi and Trichoderma in particular?**

Project Goal

This project aims to evaluate several different gene finding tools in the context of Trichoderma genomes

- Gene finding tools currently selected are GeneMark-ES and Braker2
- The selected tools aim to include a mix of *ab initio*, evidence-based and hybrid gene finding methods
- This list is not final and will include at least one more tool for comparison

Methodology



Downstream Analysis of Gene Finding Predictions

- A plan for evaluating and comparing the selected tools needs to be developed
- Metrics for comparison will include both quantitative and qualitative observations

Quantitative Metrics

- Total genes predicted
- Total transcripts predicted
- Genes predicted in repetitive regions
- Analysis of low GC content regions
- Genes overlapping small RNAs
- Length of gene models predicted
- Comparison to genes predicted in other fungal species (Yeast)
- Run times and memory usage

Qualitative Metrics

- Features of the gene finding tools
- Ease of software installation and their dependencies
- Ease of use
- Popularity among other research

Assembly Metrics

NextDenovo

Strain	Total Contigs	Total Length	Largest Contig	GC%	N50	L50
DC1	8	38.53 Mb	11.47 Mb	47.96	5.67 Mb	3
Tsth20	7	41.48 Mb	8.0 Mb	47.32	6.50 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

Initial Gene Finding Results

DC1

Tool	Gene Count	Transcript Count
Braker2	8291	8813
GeneMark-ES	11351	11351

Tsth20

Tool	Gene Count	Transcript Count
Braker2	8436	8965
GeneMark-ES	12370	12370

Initial Gene Finding Results

T. harzianum

Tool	Gene Count	Transcript Count
Braker2	8314	8385
GeneMark-ES	12164	12164

T. reesei

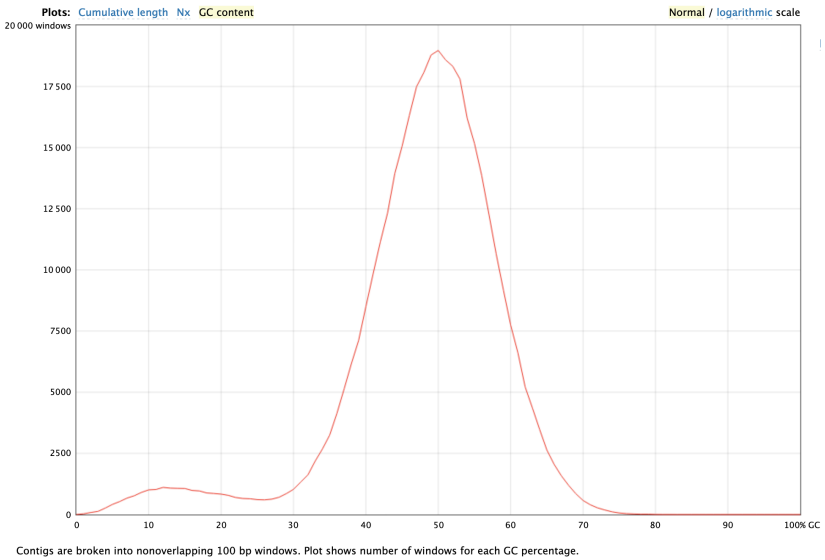
Tool	Gene Count	Transcript Count
Braker2	9659	10175
GeneMark-ES	9203	9203

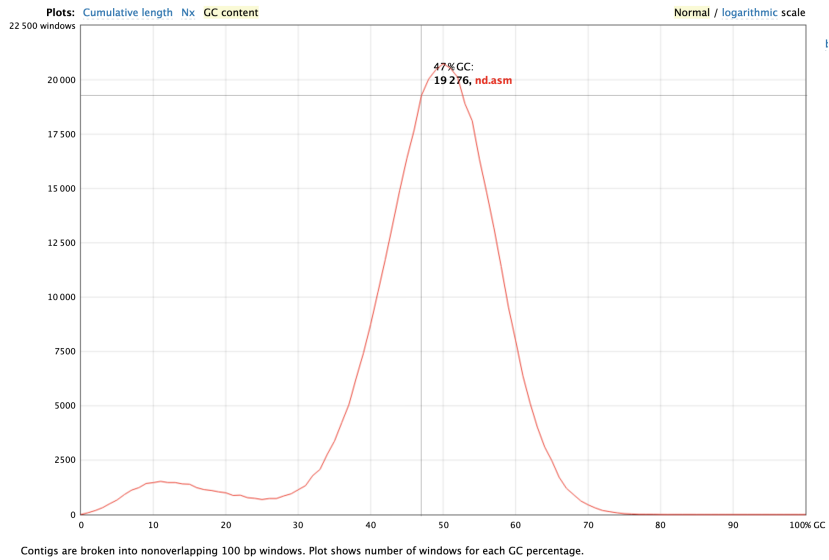
T. virens

Tool	Gene Count	Transcript Count
Braker2	7801	7863
GeneMark-ES	11877	11877

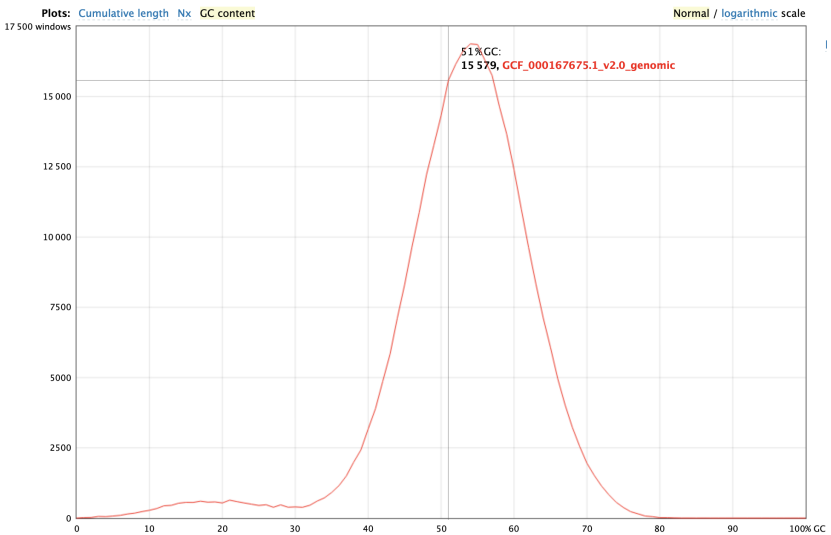
Low GC Content in *Trichoderma* Genomes

- Low GC content of some *Trichoderma* genomes
  - Example of DC1 and Tsth20 GC content below

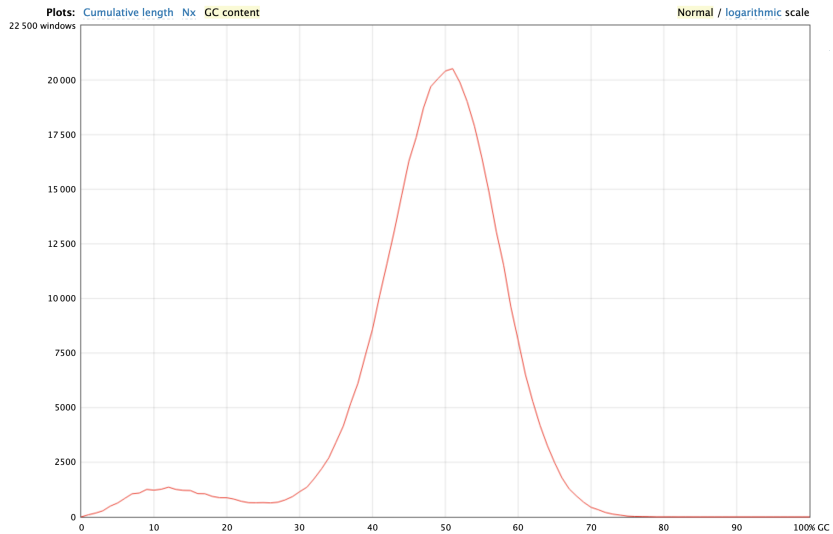




Low GC content in *T. reesei* and *T. harzianum*

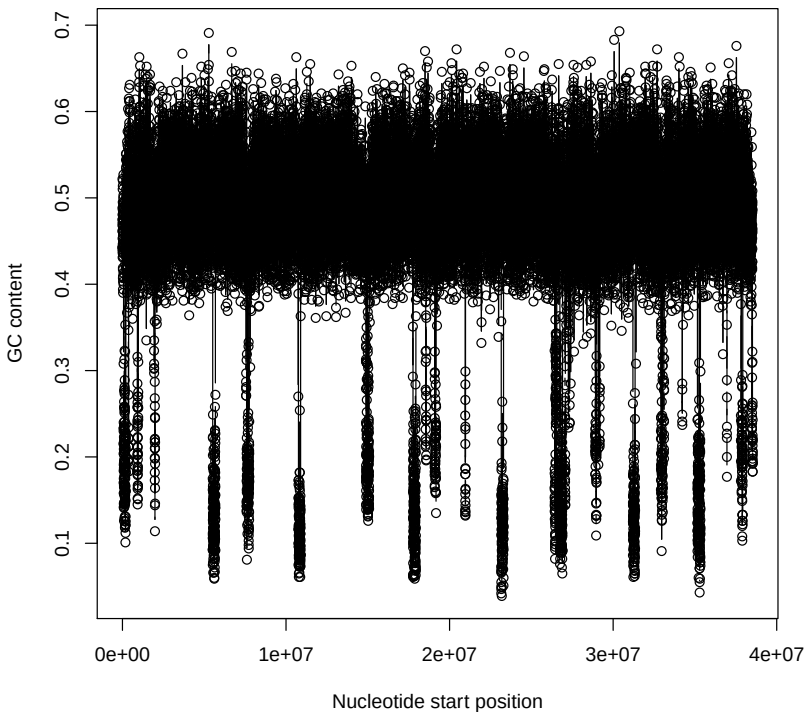


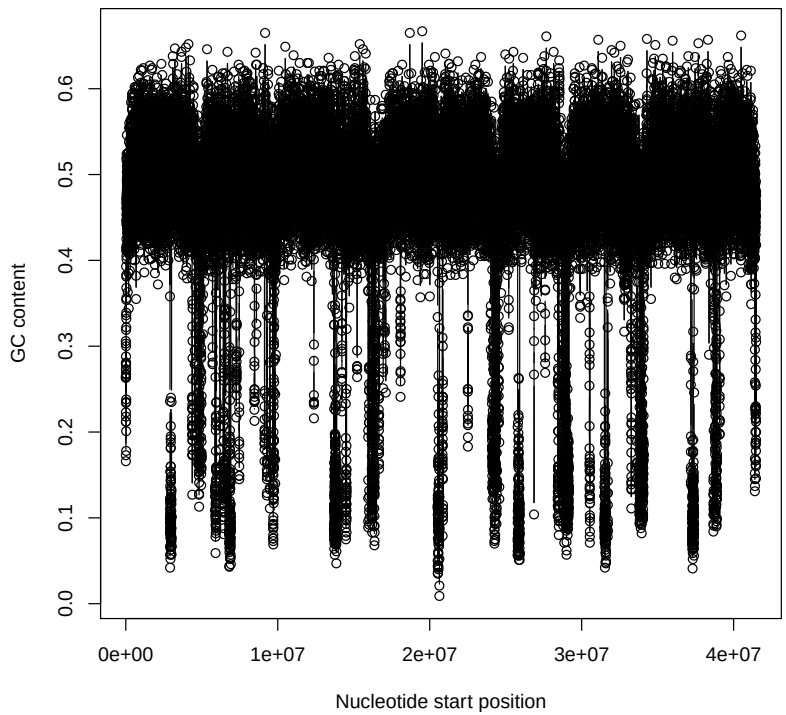
Contigs are broken into nonoverlapping 100 bp windows. Plot shows number of windows for each GC percentage.



Contigs are broken into nonoverlapping 100 bp windows. Plot shows number of windows for each GC percentage.

Low GC Content Regions in *Trichoderma* (DC1 and Tsth20)





Why is this useful to RSMI/GIFS?

- Assemblies and statistics of novel *Trichoderma* strains DC1 and Tsth20 made available
- Multiple sets of gene calls, sRNAs and repeat annotations made available
- Analysis of GC content in combination with gene content and repeat regions
- Unfortunately, I can't tell you which genes are responsible for these strains resistance to high salt content soils and bioremediation. At least not yet!

What Next?

- Genes predicted in repetitive regions
- Genes overlapping smallRNAs
- Length of gene models predicted
- Comparison to genes predicted in other fungal species (Yeast)
- Run times and memory usage
- Annotation of small RNAs using Infernal
- Annotation with another gene finding tool (possibly NCBI)

Questions?