

# List of Figures



2.1	Phylogenetic tree illustrating relationships among selected <i>Trichoderma</i> species based on multiple sequence alignments of Rpb2. Evolutionary history was inferred using the Neighbour-Joining method [1]. The optimal tree with the sum of branch lengths = 0.146 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches [2]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. An example branch of length 0.01 is shown. . . . .	6
2.2	Example of a eukaryotic gene structure, showing a promoter region (blue), the 5' and 3' UTRs (purple), start codon (green), exons (orange), and introns (grey). The directed arrow indicates the direction of transcription. . . . .	7
4.1	A simplified workflow used in this work. Each processing stage is is represented by an octagon, and the flow of data are indicated by directed edges between them. In this figure and subsequent figures, the sequence processing stage is coloured brown, the gene prediction and processing stage is coloured green, the evaluation and validation stage is coloured yellow, and the region identification process is coloured blue. . . . .	14
4.2	A workflow (in brown) depicting the steps taken to process input sequences, generate assemblies, and post-process those assemblies for use in other workflows. Processing steps are represented by boxes, datasets of interest that were produced by this work are represented by folders, and datasets that were not produced in this work are represented by cylinders. Other workflows are represented by green and blue arrows outside of the brown graph. Directed edges indicate the flow of data. . . . .	15
4.3	A plot of GC content of Illumina sequences generated for DC1. The X-axis indicates mean GC content of sequences while the Y-axis indicates the total number of sequences for a given mean GC content. A theoretical normal distribution is overlaid in blue. . . . .	17
4.4	A workflow (green) depicting the steps taken to predict genes for use in other workflows and process CDS sequence lengths. Processing steps are represented by boxes, datasets of interest that were produced by this work are represented by folders, and datasets that were not produced in this work are represented by cylinders. Other workflows are represented by blue and yellow arrows outside of the green graph. Directed edges indicate the flow of data. . . . .	19
4.5	A workflow (yellow) depicting the steps taken to evaluate and validate predicted genes. Processing steps are represented by boxes, datasets of interest that were produced by this work are represented by folders, and datasets that were not produced in this work are represented by cylinders. Other workflows using results from these processing steps are represented by a blue arrow outside of the yellow graph. Directed edges indicate the flow of data. . . . .	21
4.6	An overview of the region identification process for different datasets. Sections of the graph are outlined with rounded rectangles, indicating that the datasets within were processed along with gene calls and discussed in their own results sections. Processing steps are represented by boxes, datasets of interest that were produced by this work are represented by folders, and datasets that were not produced by this work are represented by cylinders. Directed edges indicate the flow of data. . . . .	23
4.7	IGV example . . . . .	26
5.1	Plots showing proportions of sliding windows of GC content for each <i>Trichoderma</i> genome assembly. . . . .	29
5.2	Number of genes predicted . . . . .	32
5.3	Number of coding sequences predicted . . . . .	33
5.4	CDF plots part 1. . . . .	34
5.5	CDF plots part 2. . . . .	35
5.6	CDF plots part 3. . . . .	36

5.7	BUSCO complete and missing gene counts for each gene finder across all <i>Trichoderma</i> genome assemblies. There were a total of 4492 markers in the selected BUSCO dataset. For DC1 and Tsth20, RefSeq annotations are not available, so their values are set to 0. The selected sordariomycetes.Odb12 dataset contains 4492 markers. . . . .	38
5.8	Breakdown of identified regions . . . . .	43
5.9	Example of a region with many gene calls . . . . .	44
5.10	Predictions on opposing strands . . . . .	45
5.11	Percentage of proteins with Pfam matches . . . . .	46
5.12	Agreeing Pfam matches . . . . .	48
5.13	Split Pfam matches . . . . .	49
5.14	RefSeq absence with IPS evidence . . . . .	50
5.15	Total tblast hits . . . . .	50
5.16	tblastn hits for reference proteins from <i>T. atroviride</i> , <i>F. graminearum</i> , and <i>S. cerevisiae</i> against <i>Trichoderma</i> assemblies. For DC1 and Tsth20, RefSeq annotations are not available, so their values are set to 0. . . . .	52
5.17	Regions of full, partial, and no agreement in AT-rich genomic sequence. It is important to note that in the cases of DC1 and Tsth20, full support indicates supporting gene predictions from both GeneMark and Braker2 and as such, there are no regions with partial support. . .	53
5.18	Number of genes predicted by each gene finder in AT-rich genomic sequence. . . . .	55

# List of Abbreviations


DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
CDS	Coding sequence
Mb	Megabases
Kb	Kilobases
WGS	Whole genome shotgun (sequencing)
NGS	Next generation sequencing
PCR	Polymerase chain reaction
GMO	Genetically modified organism
CPU	Central processing unit
GC	Guanine cytosine
Mb	Mega-base
HMM	Hidden Markov model
UTR	Untranslated region
GFF	General feature format
BUSCO	Benchmarking universal single-copy orthologs
RSMI	Root, soil and microbial interactions
MITE	Minature inverted-repeat transposable element
TIR	Terminal inverted repeat
TDR	Terminal direct repeat
BLAST	Basic local alignment search tool
NCBI	National center for biotechnology information
HGT	Horizontal gene transfer
TE	Transposable elements

# 1 Introduction

*Trichoderma* species are fungi from the Hypocreaceae family that are commonly found in soil and aid in the decomposition of plant matter. They are also known for their ability to produce a variety of enzymes involved in a number of different roles, including cellulose degradation, nitrogen fixation, antibody production, and cellular signalling. Some of these species are also used in agriculture as biocontrol agents against plant pathogens. Some of the enzymes responsible for these roles are produced in large quantities, making them of interest to both industry and academia. In particular, the elevated production of secondary metabolites in *Trichoderma* species has been shown to have a positive effect on plant growth and health, making them an attractive option for inoculation of agricultural crops. The genomes of these species are of interest to researchers in the field of bioinformatics, as they contain a wealth of information about the genes and proteins involved in these processes. These genomes must first be annotated to identify the genes and their functions, which is a non-trivial task. In addition, there are a number of different tools available for genome annotation, each with its own strengths and weaknesses. Gene finding tools are commonly compared to one another in common model organisms, but there is little work done comparing these tools in fungi, and even fewer in *Trichoderma* species. The sequencing of two novel *Trichoderma* genomes, DC1 and Tsth20, provides an opportunity to compare the performance of several gene finding tools in these species.

In this work, we explore the following research areas:

- How do the assemblies of the DC1 and Tsth20 genomes compare to other *Trichoderma* species?
- Profiling of gene finding tools: how are they implemented, and what features do they predict?
- How many genes and coding sequences do gene finders identify in the selected genomes?
- Do gene finders agree on the lengths of predicted genes?
- Do gene finders agree on their predictions, and if not, to what extent do they disagree?
- Do the predicted genes contain functional signatures, as identified by InterProScan?
- Do predicted genes show similarity to known proteins in *Trichoderma*?
- How do gene finders perform in benchmarking tests, such as BUSCO?
- How do gene finders perform in AT-rich genomic sequence?

We first assembled the DC1 and Tsth20 genomes, and applied two gene finding tools, GeneMark and Braker2, to these genomes. The mbles of DC1 and Tsth20 appear to be of high quality, with near-chromosomal scale contigs as well as N50 and L50 values greater than the RefSeq assemblies of *T. reesei*, *T. harzianum* and *T. virens*. We also applied the gene finding tools Braker2 and GeneMark to the DC1 and Tsth20 genomes, and compared the results to the RefSeq annotations of *T. reesei*, *T. harzianum* and *T. virens*. We found that gene finding tools are rarely in complete agreement with one another, and that the gene models produced by these tools can differ significantly in terms of start and stop positions, number of exons and introns, and presence of alternative splicing. All three sets of gene predictions performed well in BUSCO tests, and the gene finding tools also performed well in InterProScan tests, with all gene finders identifying proteins with Pfam matches. We also compared gene finding performance in AT-rich genomic sequence, and found that gene finding tools did not perform well in these regions, which may provide insight into evolutionary processes in *Trichoderma* species. Finally, we present recommendations for selection of a gene finding tool based on the results of this work. This work has produced a rich dataset of gene predictions and annotations for the DC1 and Tsth20 genomes, which can be used to further investigate the biology of these species and the performance of gene finding tools in fungi.

