List of Figures

List of Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
CDS	Coding sequence
Mb	Megabase*s*
Kb	Kilobase*s*
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 invovled 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 the 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 assemblies 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.

2 Background

This chapter provides a brief background on *Trichoderma* and gene prediction tools, providing context for this research. Section 2.3 discusses the evolutionary mechanisms that may have led to the increased production of secondary metabolites in *Trichoderma* species, such as horizontal gene transfer, gene duplication, and transposable elements. Section 2.2 discusses the novel *Trichoderma* genomes that were sequenced and assembled as part of this work, and how they relate to the research questions. Section 2.4 discusses the structure of eukaryotic genes how gene finders predict them, while section 2.5 provides an overview of secondary metabolites and their importance in *Trichoderma* species. Finally, section 2.6 provides an overview of comparative studies of gene prediction tools in *Trichoderma* species, and how this work aims to fill the gap in literature.

2.1 *Trichoderma* Species and their Features

Trichoderma are a genus of ascomycete fungi found in a wide variety of soils, and are well-known for their use in biomanufacturing of cellulases and hemicellulases as well as their roles as non-toxic, avirulent, opportunistic plant symbionts [3] [4]. Many of these features *Trichoderma* can be attributed to the large number of secondary metabolites produced by these fungi, which are used to interact with plants and other organisms in the soil [5]. *Trichoderma* species differ in the number of secondary metabolites they produce, with some species producing more than others [5]. This variation in secondary metabolite production is of great interest to researchers, as it can provide insight into the evolutionary processes that have shaped *Trichoderma* species over time.

There are many mechanisms organisms may leverage that result in gain or loss of gene function, and ultimately the evolution of a species. One well-studied mechanism is that of horizontal gene transfer (HGT), which is the process of acquiring genetic material from another organism, rather than through inheritance from a parent organism [6]. HGT is a common mechanism in bacteria, but it has also been observed in fungi, including *Trichoderma* species [6]. Another mechanism of interest is the duplication of genes, which can result in the production of more than one copy of a gene, leading to increased expression of that gene [6]. This is particularly relevant in the case of secondary metabolite production, as many *Trichoderma* species have been shown to have multiple copies of genes involved in secondary metabolite production [5]. Lastly, transoposable elements (TEs) are another mechanism of interest, as they can insert themselves into the genome and disrupt or enhance the expression of genes [6]. TEs have been shown to play a role in the evolution of *Trichoderma*

species, particularly in the case of secondary metabolite production [6].

Interestingly, both HGT and TEs tend to present themselves in regions where GC content differs from the rest of the genome [6]. Abnormal GC content is also associated with other genomic features, such as centromeres [7] and repetitive regions [8], both of which may have an effect on the production of secondary metabolites in *Trichoderma* species. As a result, it is important to consider the GC content of a genome when studying *Trichoderma* species, which forms the basis of the Research Question 3.8. Some studies have shown that *Trichoderma* species contain very few if any transposable elements, which is unusual for fungi [9]. It is hypothesized that TEs were lost in *Trichoderma* due to repeat induced point mutations (RIP), which is a process that occurs in fungi where TEs are silenced and eventually lost from the genome [9]. Genes responsible for RIP mutations also tend to target CA dinucleotides, converting them to TA nucleotides, which explains the low GC content observed in some regions of *Trichoderma* genomes [6].

Given the proximity of *Trichoderma* species to other fungi, bacteria, and plants in soils, it is possible that *Trichoderma* species acquired genes which are involved in response to antagonistic or sympathetic intercellular interactions from other organisms. While difficult to prove, studies have shown that HGT events have occurred in a number of fungal species [10]. It is also important to note that some *Trichoderma* species have higher numbers of genes involved in secondary metabolite production than others, indicating an evolutionary divergence, making comparative analysis of *Trichoderma* species and strains a useful tool for understanding the mechanisms behind secondary metabolite production [5].

2.2 Novel *Trichoderma* Genomes

Recently, two strains of *Trichoderma* have been identified in the prairie regions of Alberta and Saskatchewan. These two strains, named Tsth20 and DC1, have been found to have beneficial properties when used as an inoculant for plants in ~~the~~ soils. Tsth20 has been classified into the *Trichoderma harzianum* group, and may act as a bioremediation tool in soils contaminated with hydrocarbon content. DC1, while not yet classified, is believed to confer salt and drought tolerance to plants in dry, salty soils. While, bioremediation and resistance to drought tolerance has been investigated in other strains of *Trichoderma* [11], the study of new strains is still useful in the effort of discovering unique mechanisms used by *Trichoderma*. Little is known about the mechanisms at work in these strains, so DC1 and Tsth20 were sequenced at the Global Institute for Food Security in an initial attempt to better understand the mechanisms at work in these genomes. While this research does not explicitly identify genomic elements related the beneficial properties of these genomes, it may serve as a foundation for future research of *Trichoderma*. The assembly of these genomes come as a result of Research Question 3.1.

2.3 Evolutionary Landscape of *Trichoderma*

The evolutionary landscape of *Trichoderma* species is an important aspect of this work. DC1 and Tsth20 have not yet been classified into a specific *Trichoderma* species, although Tsth20 has been tentatively declared a *harzianum*, and further understanding of their evolutionary relationships to other *Trichoderma* species may help explain how DC1 and Tsth20 originated. Figure 2.1 shows the phylogenetic relationships among selected *Trichoderma* species as well as the outgroup *Fusarium graminearum*. The tree was constructed using cursory alignments of the Rpb2 gene, which is a commonly used gene for phylogenetic analysis in fungi [12]. Representative Rpb2 proteins from the various fungal species were extracted from the NCBI resource for fungal orthologs, while representative Rpb2 proteins from DC1 and Tsth20 were identified from the results generated later on in this work. *Trichoderma* species separate into a number of distinct clades, and these clades can vary between studies as evidenced by the low confidence in placement of *T. atroviride* and *T. gamsii*. Lack of confidence in evolutionary relationships is not uncommon in *Trichoderma* species, and is likely due to the high levels of horizontal gene transfer and gene duplication that have occurred in these species [6]. This makes *Trichoderma* species interesting to study but difficult to classify.

Phylogeny is also important when selecting genomes for comparative analysis of bioinformatics tools. Some tools may vary in their performance depending on the nature of the genome they are applied to, making a heterogeneous selection of genomes important for understanding the performance of the tools in different contexts. In this work, we selected three well-studied *Trichoderma* species, *T. reesei, T. harzianum*, and *T. virens* ~~for future comparisons~~, which are all commonly used in research and industry, but are relatively distant from each other evolutionarily according to more detailed studies [12].

2.4 Gene Predictions and Complementing Similarity Searches

Gene prediction methods are generally based on the idea that genes can be identified by searching for patterns in a sequence, which match an expected gene structure or model. These gene structures begin from a $5'$ start codon, continue through a series of exons and introns, and end with a $3'$ stop codon [13]. Flanking the gene structure are promoter regions, which are typically found upstream of the start codon, and untranslated regions (UTRs), which are found both upstream of the $5'$ start codon and downstream of the $3'$ stop codon. Promoter regions are important for the regulation of gene expression, while UTRs are important for the stability and translation of the mRNA transcribed by the gene [13]. Sequences are translated from the $5'$ start codon through to the $3'$ stop codon, splicing together exons and ignoring introns. An example of a eukaryotic gene structure is shown in Figure 2.2.

Gene prediction methods vary in their approach, but they generally fall into two categories: *ab initio* methods and evidence-based methods [14]. *Ab initio* methods rely on the identification of patterns in the sequence that match an expected gene structure, while evidence-based methods use prior information such

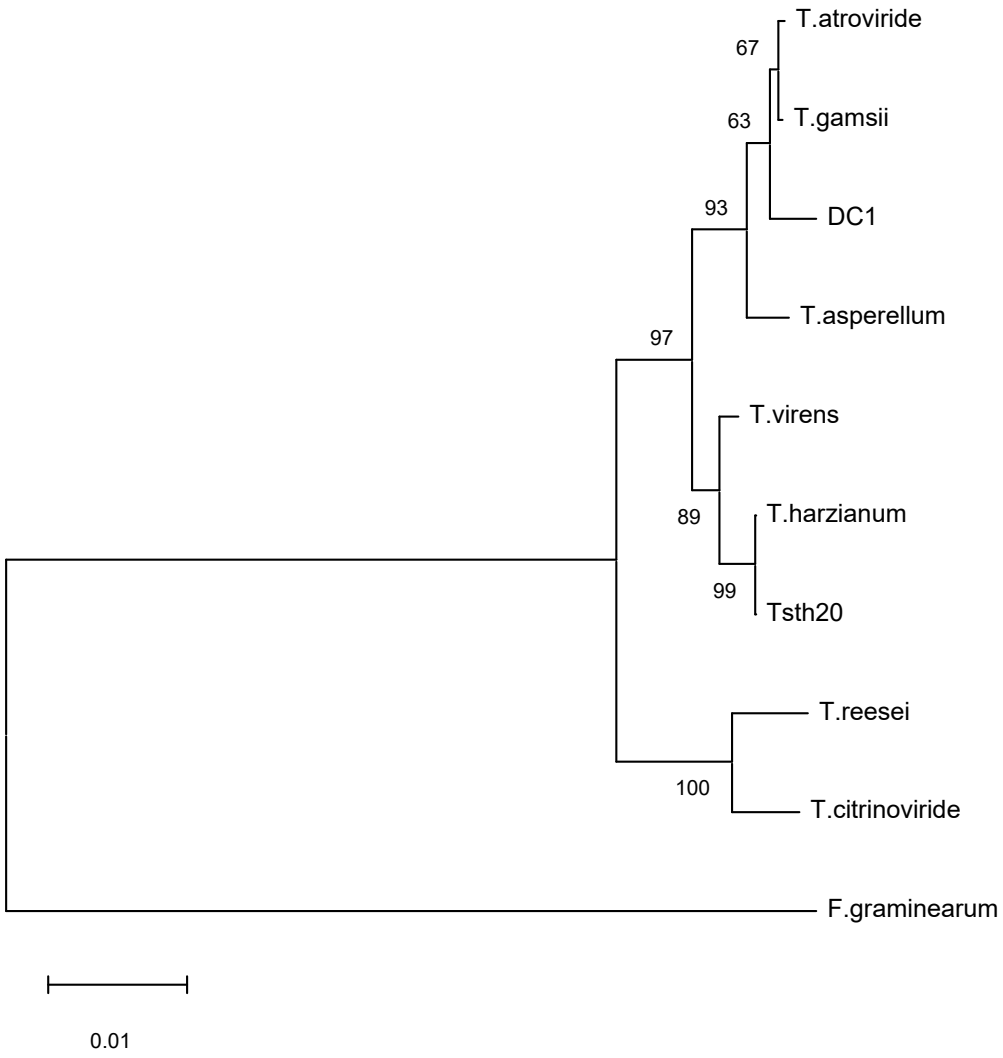

Figure 2.1: Phylogenetic tree illustrating relationships among selected *Trichoderma* species based on multiple sequence alignments of Rpb2. Evolutionary history was inferred using the Neighbour-Joining method [1]. The optimal tree with the sum of branch lengths = 0.146 is shown. The percentage of replicate 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.

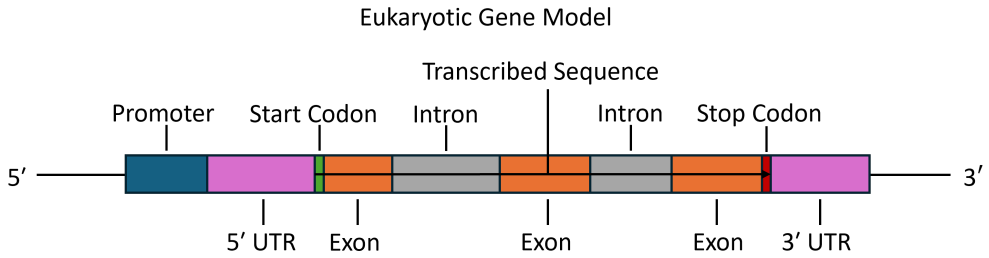

Figure 2.2: Example of a eukaryotic gene structure, showing a promoter region (blue), the 5′ and 3′ UTRs (purple), start codon (green), exons (orange), and introns (grey). The directed arrow indicates the direction of transcription.

as RNAseq data, expressed sequence tags (ESTs), and expressed protein sequences to identify genes within a new genome [14]. Gene prediction methods can also vary in their ability to identify different components of a gene, such as the promoter region, UTRs, and introns. Some methods may only identify the coding sequence (CDS) of a gene, while others may also identify the promoter region and UTRs [14]. Making things more complicated, different gene prediction tools may predict different combinations of exons and introns for a gene, and some tools may even predict multiple combinations of exons and introns, leading to multiple transcripts for the same gene. As a result, different transcripts may produce different proteins, or the same protein with different post-translational modifications, which can significantly impact the function of the resulting protein [14]. Given the already complex nature of secondary metabolites, it is important to consider the impact of variability in gene prediction methods, especially in the case of *Trichoderma* species and their secondary metabolite production. This consideration is one of the main motivators behind this research, and forms the basis of Research Question 3.5.

As mentioned earlier, gene finding tools generally fall into one of two categories: *ab initio* and evidence-based. *Ab initio* methods are often used when no prior information is available, and include tools such as GeneMark [15], GenScan [16], and GlimmerHMM [17]. Evidence-based methods are used when prior information is available [14], and include tools such as BLAST and other alignment tools to identify sequences in the genome that are similar to known sequences. More recently, hybrid tools have been developed that combine both *ab initio* methods and evidence from RNAseq data, ESTs, and protein sequences to identify genes in a genome [14]. These hybrid tools are often more accurate than either *ab initio* or evidence-based methods alone, as they can take advantage of the strengths of both approaches. Some examples of hybrid tools include AUGUSTUS [18], Braker2 [19], and MAKER [20]. GeneMark and Braker2 were selected as tools for comparison in this work as they are both widely used in the field, with high citation counts of over 2000 and 1600 respectively on individual papers, and also represent an *ab initio* method and a hybrid method, respectively. Braker2 was a particularly interesting option, as it combines GeneMark with AUGUSTUS as part of it's workflow. Purely evidence-based gene prediction methods were not selected for this work, as they require prior information such as RNAseq data, which was not available for DC1 and Tsth20. The NCBI RefSeq annotations for the additional *Trichoderma* assemblies were also used as a point of comparison as

they are generated using a combination of *ab initio* and evidence-based methods, and are considered to be high-quality annotations. RefSeq annotations may be generated by the NCBI Eukaryotic Genome Annotation Pipeline [21] or by other sources as long as it is approved by NCBI for the organism of interest. To simplify discussion surrounding the NCBI representative annotations, they will be referred to as RefSeq throughout the rest of this work.

While the basic structure of a gene may be present in a sequence, that does not mean that the gene codes for a functional protein, or in the case of secondary metabolites, a component that helps produce them. To assess potential function of a gene product, gene prediction methods are often used in conjunction with similarity searches, which compare the predicted genes to known genes in other organisms to identify potential functions, and serve as a form of validation for the predicted genes [13]. Similarity searches can come in several different forms, such as ~~blast~~ searches, which compare the predicted genes to a database of known proteins, or InterProScan, which compares the predicted genes to a database of known protein domains and binding motifs [13]. These similarity searches can provide additional information about the predicted genes, such as potential functions, and can also help to validate the predicted genes by comparing them to known genes in other organisms. Similarity searches can also be used to evaluate the completeness of the predicted genes. One example would be comparing predicted genes to a set of conserved single-copy orthologs expected in fungal genomes, such as those provided by BUSCO [22]. Together, these methods can provide a more complete picture of the gene predictions and their potential functions, and can help to identify potential targets for further research. This forms the basis of Research Questions 3.6, 3.7, and 3.9.

2.5 Secondary Metabolites

While cellular products essential to an organism's viability are of great interest, there are a vast number of cellular products that while not essential, still provide great benefit to the organism and may be necessary for survival in some situations [23] [5]. These other cellular products are known as secondary metabolites, and they are involved in several roles ranging from cellular signalling to antibiotic activity, making them a frequent subject of study in pharmaceutical research. Enzymes that produce secondary metabolites are comprised of non-ribosomal peptide synthetases (NRPS) and polyketide synthetases (PKS), which synthesize amino acids and other basic enzymatic building blocks into proteins [24]. Genes encoding the NRPSs and PKSs responsible for production of secondary metabolites are often found in clusters within a genome, but their products remain unknown [5].

The study of these products is difficult as the genes encoding the modules that make up NRPSs and PKSs are not expressed under normal laboratory conditions [5]. The evPSs and PKSs that have been studied have been shown to be large enzymatic structures containing several functional modules, each responsible for a specific step in the synthesis of a protein [5]. Studying these complicated mechanisms and their genes requires as close to a complete set of gene predictions as possible, as the absence of one gene encoding a module could

affect the entire protein complex. This concept of completeness of gene predictions is explored in Research Question 3.9. In addition, gene predictions can be processed further with tools such as antiSMASH, which can identify secondary metabolite biosynthetic gene clusters (BGCs) in a genome [25]. Comparison of predicted genes with known secondary metabolite BGCs can provide insight into completeness of the predicted genes with a specific focus on secondary metabolite production. This work is not covered in this thesis, but is worth noting for context as it is a common approach in the field of secondary metabolite research.

Another interesting feature of NRPSs and PKSs is the length of the genes encoding them, which can be quite long, often exceeding 10,000 base pairs in length [24]. Examining the outputs of gene prediction tools can provide insight into the lengths of predicted genes, and whether or not the tools are able to capture the full length of these large genes. This topic serves as the basis for Research Question 3.4.

2.6 Gene Prediction Tools in Fungi

With the biological context laid out, we can now turn our attention to the computational aspect of this work. Many *Trichoderma* species have been assembled and annotated in the past, mostly for the purpose of understanding the mechanisms behind secondary metabolite production [5]. Gene predictions in this context are typically compared against sequences from other organisms to derive functional information, and to validate the predicted genes [13]. In contrast, very few studies have been conducted comparing gene prediction tools in fungal genomes, and even fewer have been conducted in *Trichoderma* species. This is surprising, given the large number of gene prediction tools available, and the fact that many of them are used in the field of fungal genomics [14]. Comparisons of gene prediction tools are common when new tools are released, but these comparisons are often limited to a small number of tools, and tend to be focused on achieving higher counts of interesting genes, rather than comparing the results in a biological context [26]. Complicating the matter further, the quality of genome sequences available may vary, making evaluation of different gene prediction tools difficult when working with multiple genomes and lesser-studied fungal species with low quality assemblies. Given the gap in literature regarding comparative analysis of gene prediction tools in *Trichoderma* species, and the availability of the novel DC1 and Tsth20 genomes, this work aims to fill that gap by comparing the outputs of several gene prediction tools on these two genomes along with three other frequently studied *Tichoderma* species.

Another pitfall of existing literature is that many studies compare tools only against a reference annotation, such as the RefSeq annotations for some organisms. However, these reference datasets are not always available, may not be representative of the genome being studied, and may be incomplete. In addition, it is helpful to compare the outputs of gene prediction tools against each other, as this can provide insight into the strengths and weaknesses of each tool relative to one another. For these reasons, this work aims to compare outputs of gene prediction tools in an all-to-all manner, rather than an all-to-one manner.