Comparative analysis of Gene Finding tools when applied to *Trichoderma* genomes

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Master of Science
in the Applied Computing of Computer Science
University of Saskatchewan
Saskatoon

By
Connor Burbridge

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

placeholder

Acknowledgements

I would like to acknowledge both Dr. Kusalik and Dave Schneider for their excellent support and mentorship during both COVID and my MSc. project. I would also like to thank Brendan Ashby and Dr. Leon Kochian for providing both data and additional financial support.

Contents

List of Tables

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, 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

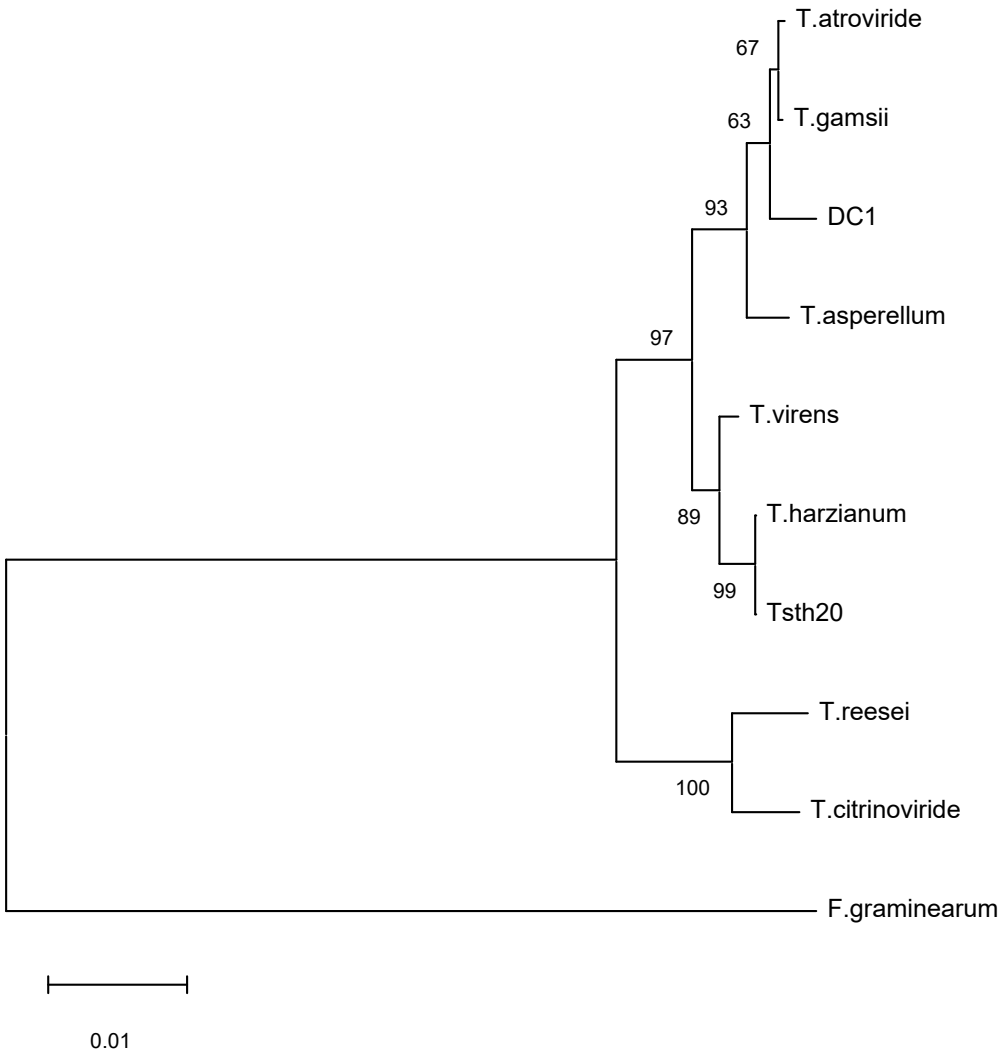


Figure 2.1: Phylogenetic tree illustrating relationships among selected *Trichoderma* species based on multiple sequence alignments of Rpb2. Evolutionary history was inferred using the Neighbour-Joining method [1]. The optimal tree with the sum of branch lengths = 0.146 is shown. The percentage of replicate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

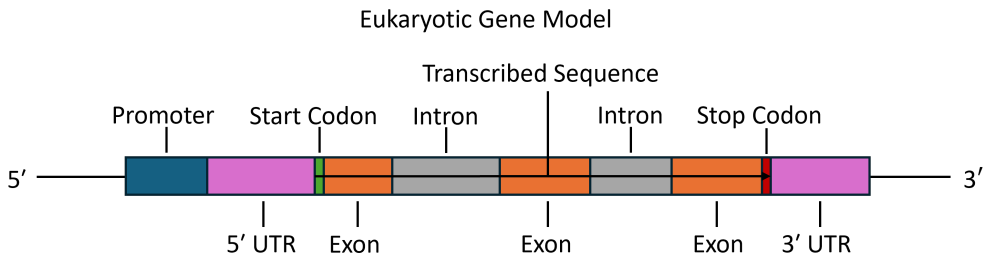


Figure 2.2: Example of a eukaryotic gene structure, showing a promoter region (blue), the 5′ and 3′ UTRs (purple), start codon (green), exons (orange), and introns (grey). The directed arrow indicates the direction of transcription.

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.

3 Research Questions

With an ever-increasing numer of gene prediction tools available, and the continuous discovery of new and unique genomic sequences, it is important to assess the performance of these gene finding tools. This chapter will address the following research questions, which will be answered in the context of *Trichoderma* genomes. The answers to these questions will provide insight into the performance of gene finding tools, and will help inform the selection of a gene finding tool for future work.

3.1 *Trichoderma* Assembly Results

Since gene finding tools operate on an assembled genomic sequence, it must follow that the results will be influenced by the supplied assembly. Before applying gene finders to the new assemblies, we should first evaluate the assemblies using general assembly metrics and compare with them existing assemblies. We ask: **how do assemblies of DC1 and Tsth20 compare to existing *Trichoderma* assemblies?** With these isolates being from the *Trichoderma* family, we expect assembly metrics to be similar in nature to existing assemblies from NCBI, but are they? In addition, what are the quantitative features of these assemblies?

3.2 Profiling of Gene Finding Tools

Different gene finding tools may predict different types of features associated with gene structures. The question arises: **which (if any) gene finders predict additional features outside of the standard gene model?** Additional features in this case include promoter sequences, transcription binding sites, activating sequences and other upstream or downstream sequences. In addition, different gene finding tools employ differing programming languages and algorithms which raises several questions. **How are these gene finding tools implemented? Is the software straightforward to install? Are the tools user-friendly? What is the processing time and memory consumption of different gene finding tools in the context of *Trichoderma*?**

3.3 Number of Features Predicted

One common method for evaluating gene finding tools is by comparing the number of genes and coding sequences each tool predicts. This gives insight into the sensitivity of each tool, as well as a sense of

completeness of the predicted gene set. We will compare the number of features predicted by each tool, and ask: **Do gene finders predict similar numbers of features in *Trichoderma* genomes?**

3.4 Lengths of Predicted Genes

Genome assemblies can contain a wide range of gene lengths, and for some users, genes of a specific length may be a key point of interest, as is the case in *Trichoderma*, where genes encoding non-ribosomal peptide synthases are generally quite large [24]. Thus, it is important to assess the lengths of predicted genes from different gene finding tools. We will compare the lengths of predicted genes from each tool and ask: **do gene finders predict genes of similar lengths in *Trichoderma*?**

3.5 Identifying Regions of Agreement and Disagreement

With several sets of gene predictions, it is important to assess the agreement and disagreement between the predictions of different gene finding tools. To answer this question, we will identify 'regions' of overlapping predictions. A region is defined as a set of overlapping gene predictions, where a gene prediction from one tool overlaps with a gene prediction from another tool. With overlapping regions identified, we can determine any agreement and disagreement in predictions between gene finding tools from which we can ask: **do gene finders agree on their predictions? If no, to what extent do they disagree?**

3.6 Validation of Predicted Genes via InterProScan

In an effort to validate, or at least provide supporting evidence for any given gene prediction we will apply InterProScan to coding sequences predicted by each of the gene finders to identify features associated with protein function. Genes will be considered as 'valid' if the gene's protein sequence contains binding sites, motifs, or other fucntional characteristics of proteins. Using the results from InterProScan, we ask the question: **in the context of *Trichoderma*, do proteins predicted by gene finders contain functional signatures?**

3.7 Similarity Searches of Predicted Genes with tblastn

In a similar manner to validation via InterProScan, we can use protein alignments to search for similarity between predicted genes and existing protein sequences. This will allow us to assess the degree of similarity between predicted genes and known proteins, which can provide further validation of gene predictions and help infer possible function. We ask: **do gene finders predict genes that are similar to known proteins in *Trichoderma***

3.8 Performance in AT-rich Genomic Sequence

Results from the assembly process show that AT-rich sequence content is prevalent in most *Trichoderma* assemblies we selected. These regions of assemblies present an interesting opportunity to assess gene finding performance in regions of anomalous GC content, since features such as transposable elements and repetitive sequences are often found in these regions. We ask: **do gene finders perform differently in AT-rich regions of *Trichoderma* genomes?**

3.9 BUSCO Completeness

The use of existing benchmarks for gene finding performance is useful when assessing performance of gene finding tools, particularly in the case of genes that should be evolutionarily conserved. **Do gene finders predict conserved single-copy orthologs expected in fungal genomes?**

3.10 Selection of a Gene Finding Tool

With all the results generated, we can provide insight to the question: **which gene finding tool should one choose?**

4 Data and Methodology

4.1 Methodology Overview

The general processing steps performed in this work can be grouped into four processing stages, or workflows. A simplified overview of these stages and the flow of data between them is shown in Figure 4.1. In ascending order, each processing stage is discussed in more detail in Sections 4.2, 4.3, 4.4, and 4.5.1.

4.2 Sequence Processing and Assembly Workflow

The sequence processing and assembly workflow is shown in Figure 4.2. Processing steps are represented by boxes, results of interest generated by this work are represented by folders, and datasets that were not produced by this work are represented by cylinders. Other workflows that use datasets produced by this workflow are depicted as arrows outside of the brown sub-graph.

Illumina©™ [27] sequences were first trimmed and filtered using Trimmomatic [28] to remove adapter content and low-quality sequences, as low-quality sequences may affect the assembly process, resulting in fragmented and erroneous assemblies. The processed Illumina©™sequences and Nanopore©™ [29] sequences were then supplied to the hybrid genome assembly tool NextDenovo [30]. As Ilumina sequences are of extremely high accuracy, especially after trimming, the processed Illumina©™data was then re-used to polish the assemblies with NextPolish [31]. Polishing is used because long-read sequences used in the assembly process are less accurate, at approximately 90-95%, than Illumina sequences at 99.9%, so the Illumina data can be used to correct any assembled sequences that may contain base-call errors. The resulting assemblies were then analyzed to identify AT-rich genomic sequence, analyzed with QUAST for general assembly metrics, and then used as input in the gene prediction workflow.

The processing steps Trimmomatic filtering, NextDenovo assembly, NextPolish Polishing and QUAST assembly assessment are described in Section 4.2.1. The AT-rich sequence identification process is described later in Section 4.2.2.

4.2.1 Sequence Pre-processing and Assembly

Genomic sequences used for assembly were generated for DC1 and Tsth20 using Nanopore©™ [29] and Illumina©™ [27] sequencing technologies by Brendan Ashby at the Global Institute for Food Security at the University of Saskatchewan. Short-read sequences were first examined with the FastQC tool [32] to

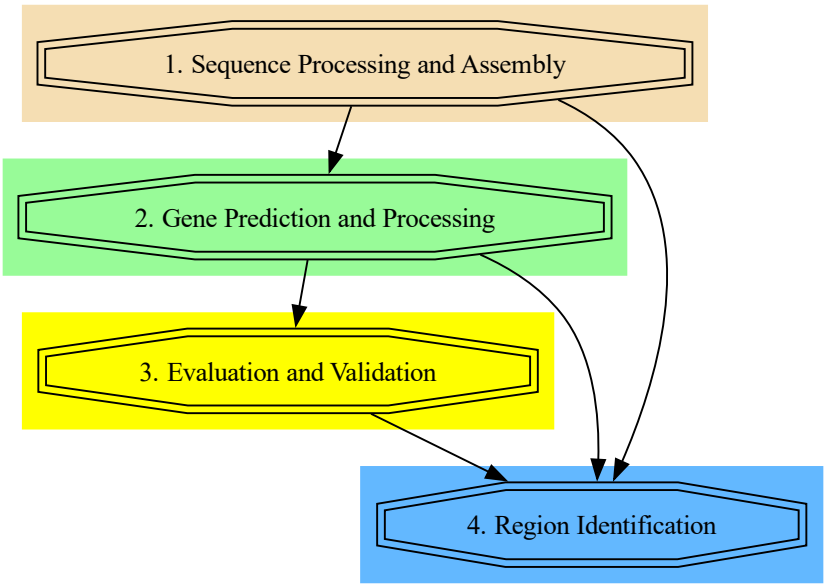


Figure 4.1: A simplified workflow used in this work. Each processing stage is is represented by an octagon, and the flow of data are indicated by directed edges between them. In this figure and subsequent figures, the sequence processing stage is coloured brown, the gene prediction and processing stage is coloured green, the evaluation and validation stage is coloured yellow, and the region identification process is coloured blue.

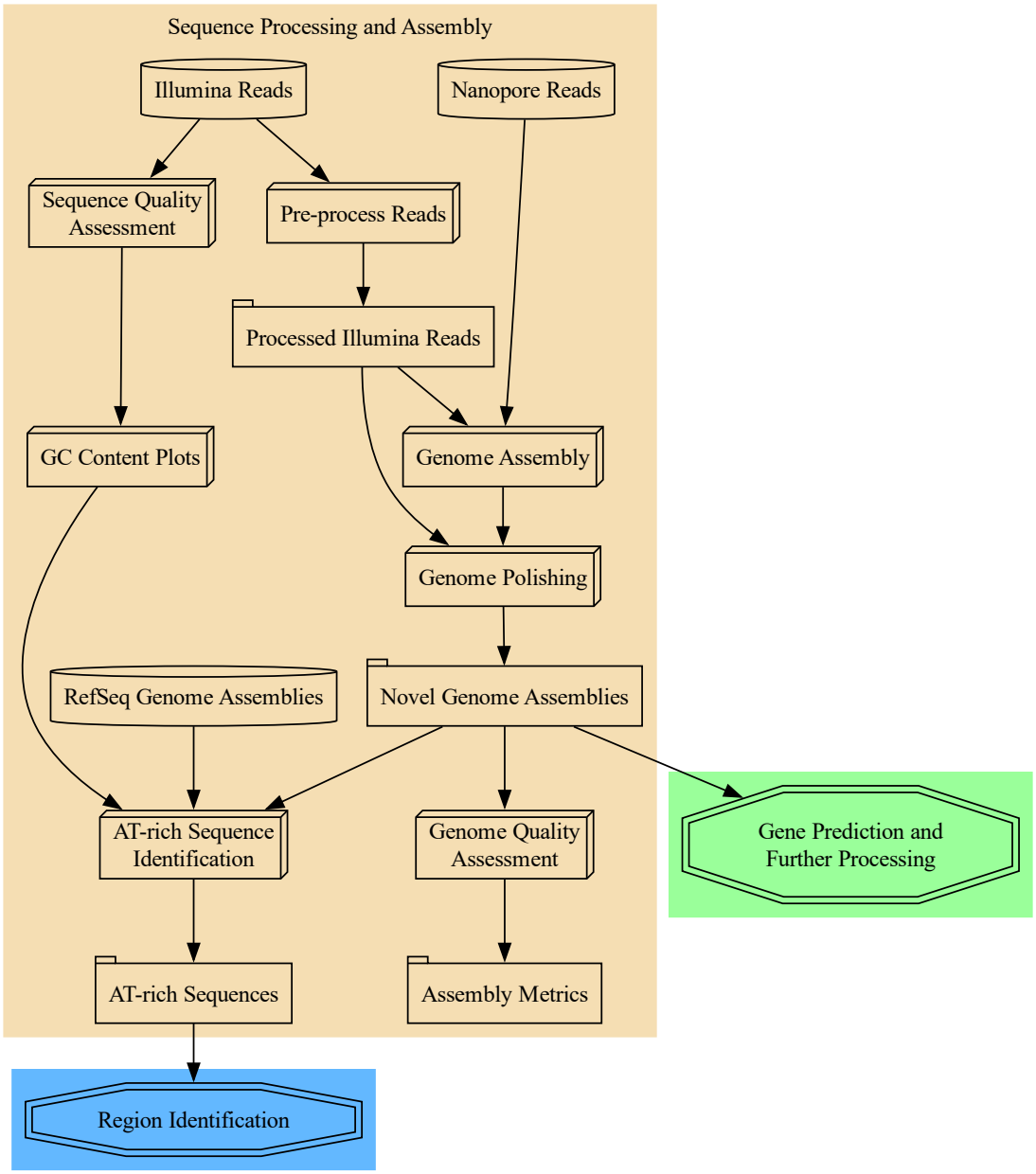


Figure 4.2: A workflow (in brown) depicting the steps taken to process input sequences, generate assemblies, and post-process those assemblies for use in other workflows. Processing steps are represented by boxes, datasets of interest that were produced by this work are represented by folders, and datasets 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.

evaluate sequencing quality and general metrics associated with the sequencing run. FastQC was run with default parameters. The short-read sequences were then trimmed to remove any adapter content or low quality sequences as they may lead to fragmented assemblies and erroneous base calls. Illumina©™sequencing data was filtered using Trimmomatic v0.38 [28] with filtering criteria as follows: SLIDINGWINDOW:4:28, LEADING:28, TRAILING:28, MINLEN:75. Only surviving paired-end reads were used for further analysis. The Nanopore©™data were not processed prior to assembly.

It has been shown that assemblies that use both short-read and long-read sequencing data during the assembly are of higher quality and less fragmented than assemblies using only one form of sequencing. Genomic Nanopore©™and Illumina©™sequences from DC1 and Tsth20 were assembled in to contigs using the hybrid assembly tool NextDenovo©™ [30] v2.5.0 with default parameters. Although hybrid genome assemblies are of higher quality, they may still contain base calling errors after assembly. The highly accurate short-reads can be used to correct base calling errors after assembly. To do this, the hybrid assemblies were polished with Illumina©™sequences using NextPolish [31] v1.4.1 and default parameters. Final assembly metrics were calculated using QUAST v5.0.2 [33] with default parameters.

4.2.2 Identifying AT-rich Genomic Sequence

Initial exploration of the Illumina sequence data generated for DC1 and Tsth20 showed bi-modal distributions of GC content in the sequences. An example of the GC distribution of Illumina sequences for DC1 is shown in Figure 4.3 compared to a theoretical normal distribution. To differentiate AT-rich sequences from normal sequences, a cutoff value was selected to isolate AT-rich sequences from the rest. The cutoff value was chosen as 28% GC content, which is roughly the point where the AT-rich sequence counts begin increasing in all assemblies. The details of this process are explained below.

AT-rich genomic sequences were identified using sliding windows of GC content over each *Trichoderma* assembly. Sliding windows were generated using the isochore program from the EMBOSS tool suite v6.6.0 [34]. The sliding window used was 250bp wide with a 50bp shift. A window was classified as AT-rich if the sequence contained lower than 28% GC content. The genomic sequences from AT-rich windows were mapped to a GFF file and processed using the region identification algorithm described in section 4.5.1.

4.3 Overview of Gene Prediction and Further Processing

To better understand the methods used in gene prediction and further processing, the steps have been assembled into a pipeline shown in Figure 4.4. Processing steps are represented by boxes, results of interest generated by this work are represented by folders, and datasets that were not produced in this work are represented by cylinders. Other workflows that use datasets produced by this workflow are depicted as arrows outside of the green sub-graph.

First, RefSeq assemblies and their associated annotations of closely related *Trichoderma* species were

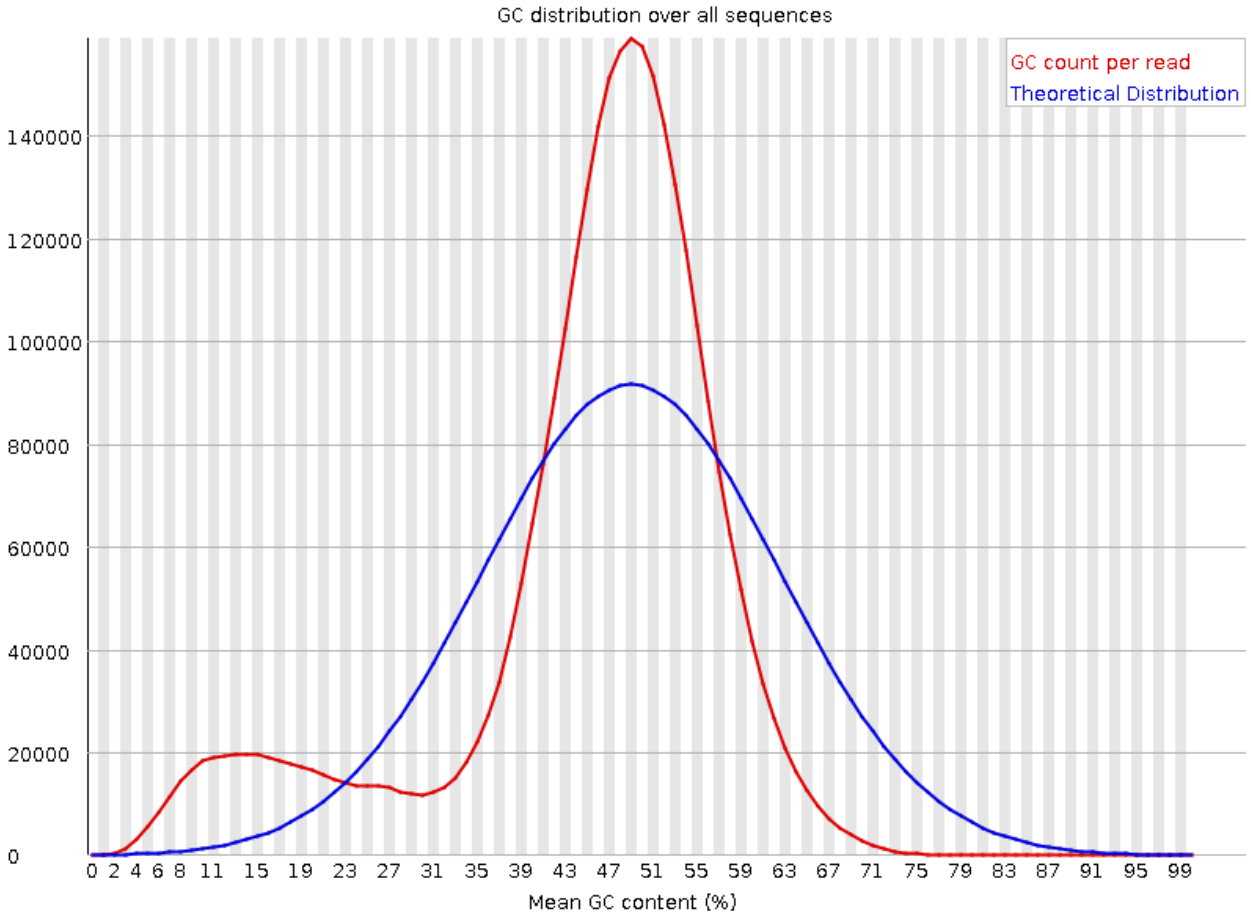


Figure 4.3: A plot of GC content of Illumina sequences generated for DC1. The X-axis indicates mean GC content of sequences while the Y-axis indicates the total number of sequences for a given mean GC content. A theoretical normal distribution is overlaid in blue.

selected from the National Center for Biotechnology Information (NCBI) for training and comparison. Both *ab initio* (GeneMark [15]) and evidence-based (Braker2 [19]) gene finders were selected for comparison to evaluate gene prediction behaviour of both methods in the context of *Trichoderma* genomes. Gene prediction was performed using both gene finders on the RefSeq genomes and the genomes of DC1 and Tsth20. Coding sequences output by the gene finders were then examined to better understand distributions of CDS lengths produced by different gene finders. Finally, all gene predictions were then supplied to the evaluation and validation stage as well as the region identification stage.

Datasets from the National Center for Biotechnology Information (NCBI) were selected for training and comparison, and are further discussed in Subsection 4.3.1. The GeneMark gene prediction process is described in Subsection 4.3.2, while the Braker2 gene prediction process is described in Subsection 4.3.3. The process used to examine CDSs is described in Subsection 4.3.4.

4.3.1 Selection of Datasets from NCBI

When evaluating gene-finders, it is important compare gene predictions to a standard, or control dataset, as a method of ground-truthing. Such datasets include reference genome sequences and their associated gene predictions from the NCBI as used in this work. In addition to comparing against control datasets, RNAseq data can also be used as training data for various tools as well, such as the evidence-based training used by Braker2. In this work, four forms of these inputs were selected from NCBI for comparison and training. The first two datasets selected were assemblies and associated gene predictions of three RefSeq *Trichoderma* organisms, with those organisms being *T. virens*, *T. harzianum*, and *T. virens*. NCBI RefSeq accession IDs for those assemblies are GCF_000167675.1_v2.0, GCF_003025095.1_Triha_v1.0, and GCF_000170995.1_TRIVI_v2.0, respectively. *T. reesei* was selected as it is the most well-studied *Trichoderma* species, and widely considered to be the gold-standard for *Trichoderma* genomes [35]. *T. harzianum* and *T. virens* were selected as additional datasets due to their prevalence in literature as well. It is important to note that while RefSeq is considered a high-quality dataset, we did not use it as a ground truth in this case, but rather an additional point of comparison as it it possible that RefSeq annotations may also contain errors or omissions. This aligns with the method of region identification used later in this work, which is aimed to be a generic and adaptable method of comparing gene predictions, or any feature in GFF3 format, from multiple sources.

The second set of selected inputs were RNAseq datasets from *T. reesei*, used as input in the training process for Braker2. The use of training data in gene finding applications is to produce a predictive model that factors in experimental evidence for a 'tailored' gene prediction model. These models can predict genes with more confidence and predict the correct gene structure more often [36]. The SRA accession IDs for sequences used are: SRR5229930, SRR8329344, SRR8329345, SRR8329346, and SRR8329347. Limited RNAseq datasets were used as there were few RNAseq datasets available at the beginning of this project.

The final set of inputs were protein sequences from RefSeq for the species *T. atroviride*, *Fusarium graminarium*, and *Saccharomyces cerevisiae*, which were for downstream validation of gene predictions. The NCBI

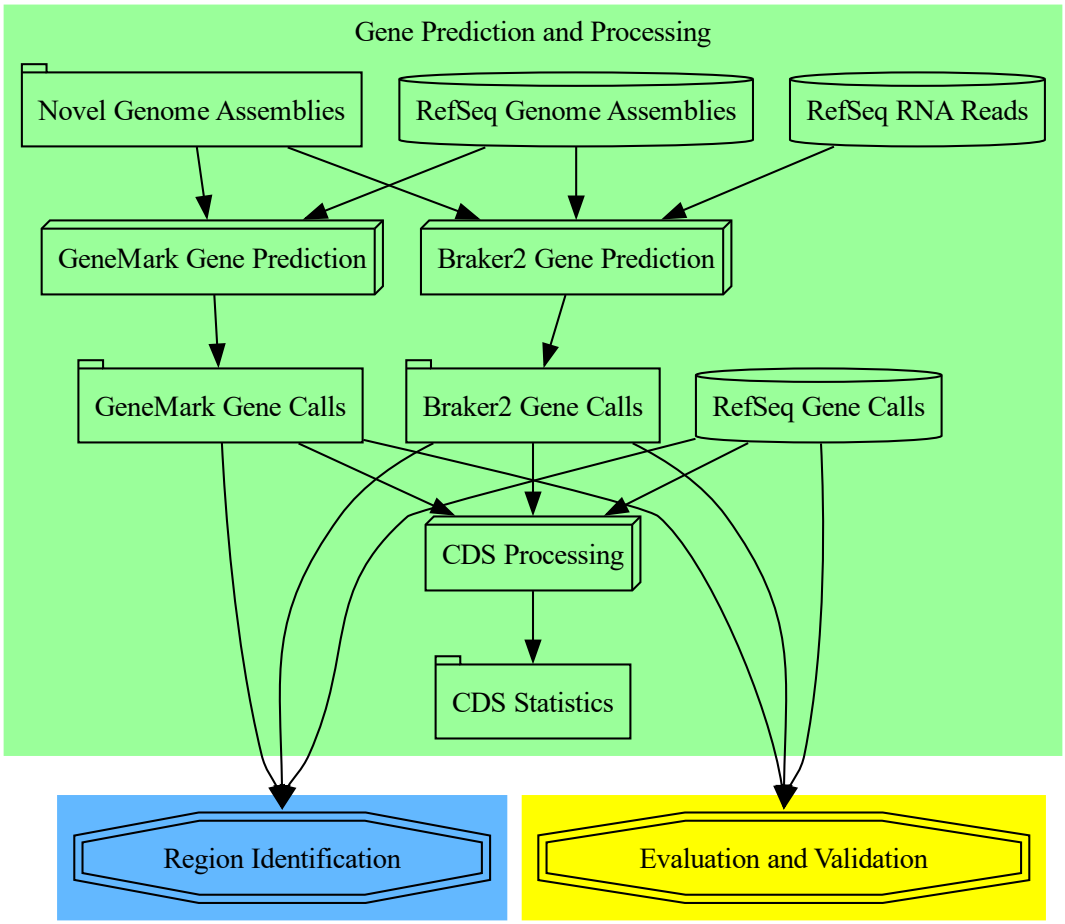


Figure 4.4: A workflow (green) depicting the steps taken to predict genes for use in other workflows and process CDS sequence lengths. Processing steps are represented by boxes, datasets of interest that were produced by this work are represented by folders, and datasets that were not produced in this 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.

accession IDs for these assemblies and associated protein sequences are GCF_000171015.1, GCF_000240135.3, and GCF_000146045.2, respectively.

4.3.2 Gene Prediction Using GeneMark-ES

Ab initio gene finding was performed using GeneMark-ES v4.71 [15] with DC1, Tsth20 and selected RefSeq *Trichoderma* assemblies used as input. Default parameters were used in all cases except for the inclusion of the fungal option, which allowed the use of a GeneMark model specific to fungal genomes.

4.3.3 Gene Prediction Using Braker2

For evidence-based gene finding, Braker2 v3.0.2 [19] was used with *Trichoderma reesei* selected as the reference organism for training. Illumina©™ [27] RNAseq datasets from *T. reesei* were downloaded from the NCBI short-read archive using the sra-toolkit [37] and were not trimmed or filtered prior to their use in Braker2 training. Default parameters were used to train a Braker2 model on *Trichoderma reesei* data except in the addition of the fungal option, which was used for this analysis for improved gene-finding performance in fungi. The trained Braker2 *T. reesei* prediction model was then applied to all assemblies with default parameters.

Braker2 also provides a UTR option to predict upstream sequence features, but that option is experimental and was left off for this work.

4.3.4 Statistical Analysis of CDS Lengths

The distribution of gene lengths predicted by a gene-finder is an important feature to consider when selecting a gene finding tool, as genes of a particular length may be of interest to researchers, but more importantly, gene-finders may be biased to certain distributions based on their underlying models. This is important as the distribution of predicted gene lengths may also play a role in *Trichoderma's* role as an avirulent plant symbiont. These biased distributions may differ from the true distribution of CDS lengths leading to propagation of sub-optimal gene-finding results and results that may not reflect the true distribution of gene lengths. To determine, if differences exist between distributions of predicted gene lengths from each tool, the $\log_{10}$ lengths of coding sequences (CDS) were subject to a two-sample Kolmogorov-Smirnov test [38]. The null hypothesis of this test is that the number of genes in AT-rich and normal GC content regions should be proportional to the length of AT-rich and normal GC content regions in the genome.

4.4 Overview of Evaluation and Validation of Gene Predictions

Following the gene prediction process, prediction results can then be evaluated and used as queries against exiting datasets as a form of validation.

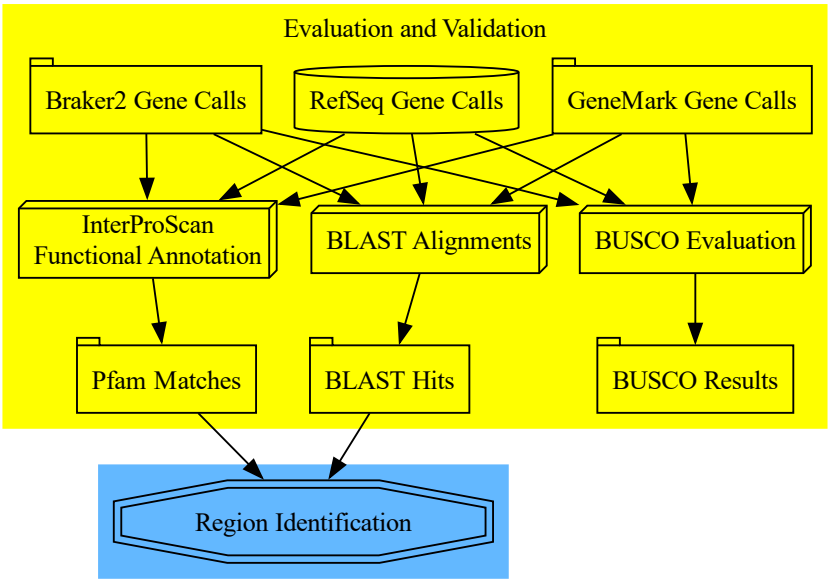


Figure 4.5: A workflow (yellow) depicting the steps taken to evaluate and validate predicted genes. Processing steps are represented by boxes, datasets of interest that were produced by this work are represented by folders, and datasets that were not produced in this work are represented by cylinders. 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.

Coding sequences from the previous stage were evaluated and validated using several tools. Firstly, protein products from predicted CDSs were processed with InterProScan to identify Pfam domains as evidence for the validity of gene predictions. Secondly, CDSs were analyzed with BUSCO to assess prediction of conserved fungal genes. Finally, RefSeq proteins from related fungal species were used in tblastn searches of the five *Trichoderma* assemblies selected for the main body of this work.

BUSCO evaluation, InterProScan Annotation, and BLAST alignments processes are described in more detail in Subsections 4.4.1, 4.4.2, and 4.4.3, respectively. Results from these processes are also used as inputs in the region identification workflow later.

4.4.1 Evaluating Gene-Finder Performance Using BUSCO

While novel gene predictions are of great interest in new subjects, it is also important to confirm that gene-finders are predicting genes that are expected to be present in the organism of interest [39]. To determine if this is true, BUSCO v6.0.0 [40] was selected as a tool to determine a gene-finder's ability to predict a set of conserved orthologous genes. The BUSCO Metaeuk pipeline was run using default parameters with the BUSCO sordariomycetes_Odb12 fungal lineage dataset to capture highly conserved genes in fungal genomes

closely related to *Trichoderma* species. This pipeline was applied to all genome assemblies used in this work.

4.4.2 Functional Annotation Using InterProScan

The presence of a functional component in a predicted gene's protein product is good evidence in favour of a gene-finder's performance. To identify functional components, the protein products from each gene finder's predictions were functionally annotated using InterProScan v5.65-97.0 [41] without the match lookup service and with the following analyses included: AntiFam-7.0, CDD-3.20, Coils-2.2.1, FunFam-4.3.0, Gene3D-4.3.0, Hamap-2023_01, MobiDBLite-2.0, NCBIfam-13.0, PANTHER-18.0, Pfam-36.0, PIRSF-3.10, PIRSR-2023_05, PRINTS-42.0, ProSitePatterns-2022_05, ProSiteProfiles-2022_05, SFLD-4, SMART-9.0, SUPERFAMILY-1.75. InterProScan was run on all predicted proteins using default parameters. The resulting protein annotations were mapped back to their genome assemblies and processed using the region identification algorithm in section 4.5.1.

4.4.3 Ground-truthing with the Basic Local Alignment Search Tool (BLAST)

The *Trichoderma* assemblies chosen for comparison were used as the subject sequences in tblastn [42] searches with the query proteins discussed in paragraph three of Section 4.3.1. Default parameters were used in the tblastn searches. Hits from the tblastn alignments were then filtered for alignments with a minimum of 30% query coverage and 30% identity. These criteria were selected as they filter out spurious alignments while also retaining short alignments that capture functional or conserved protein coding sequence. Remaining alignments were then processed using the region identification algorithm in 4.5.1. In the third row, are the AT-rich sequences identified in Section 4.2.2, the tblastn alignments identified in Subsection 4.4.3, and the Pfam matches from InterProScan identified in Subsection 4.4.2.

4.5 Region Identification Overview

A major portion of this work was the process of identifying regions of overlapping features as a method of comparing gene calls from different gene prediction tools. The definition of a region and the algorithm used to identify them is described in detail in Section 4.5.1, and an overview of the region identification workflow is shown in Figure 4.6. The top of the graph shows the gene calls produced by the workflow described in Subsection 4.3, which are used as the main input to the region identification process. The inputs and resulting regions of other previously generated datasets is represented within the rounded rectangles. These additional datasets may contain other information such as BLAST alignments or functional annotations and can be included as features in the region identification process. Regions output by the region identification process were then visualized with the Integrative Genomics Viewer, described in Section 4.5.3

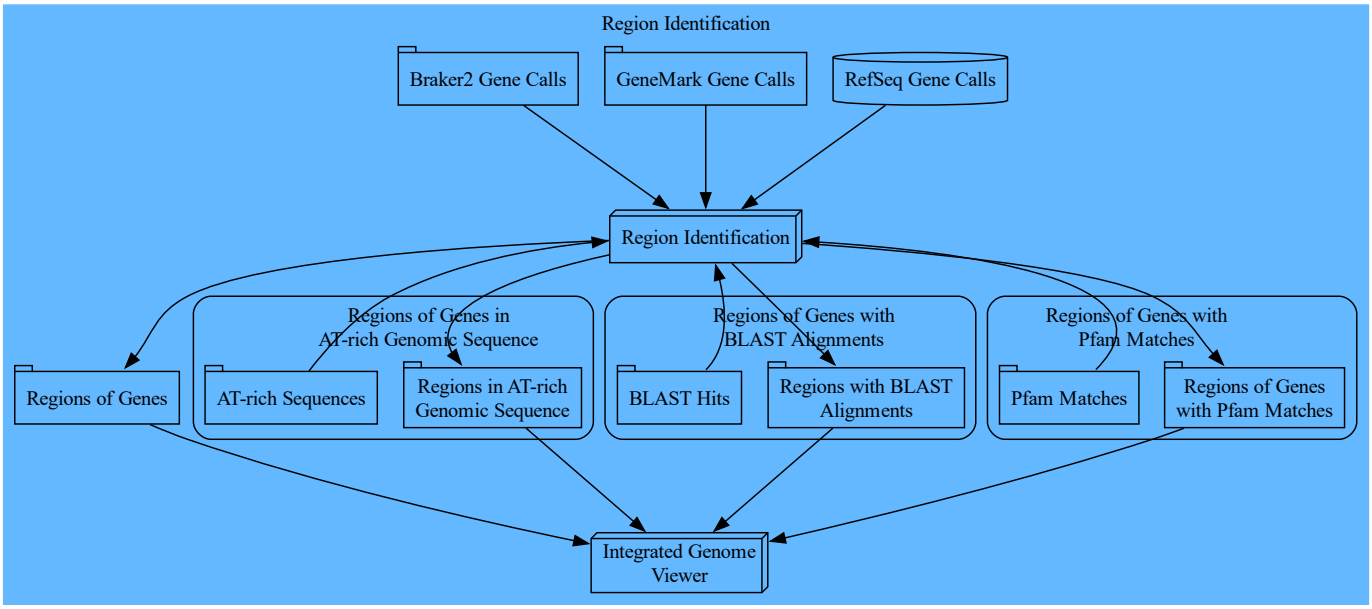


Figure 4.6: An overview of the region identification process for different datasets. Sections of the graph are outlined with rounded rectangles, indicating that the datasets within were processed along with gene calls and discussed in their own results sections. Processing steps are represented by boxes, 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.

4.5.1 Region Identification Algorithm

The gene predictions Subsections 4.3.1 4.3.3, 4.3.2, and the additional datasets from Subsections 4.2.2, 4.4.3, 4.4.2 were processed into regions of overlapping gene predictions. A region is defined as a pair of genomic coordinates containing one $5'$ and one $3'$ coordinate, between which are one or more overlapping features. We define a feature as a pair genomic coordinates containing one start and one stop coordinate that is associated with a gene prediction, annotation, or other information of interest.

Overlapping features were processed into regions using a concept similar to Allen's interval algebra [43]. The algorithm first sorts all features on each contig by start coordinate and then iterates over each feature, identifying overlaps with previous features to identify regions of overlapping features. We approached this as a graph problem, where features make up a set of nodes V, and spatial overlaps between features make up a set of edges E, forming the graph $G = \{V, E\}$. This algorithm performs a transitive closure R^+ of G, resulting in sub-graphs, or regions of features in which each node has a path to each other node in that sub-graph. Pseudo-code for the algorithm is shown in Algorithm 1. This processing was performed with Python v3.12.4 [44] and the GFF package from BCBio [45] for easy GFF parsing. The resulting regions were then further processed to identify trends and behaviours of gene finders in the context of other annotated information and features.

4.5.2 Regions in AT-rich Genomic Sequence

AT-rich genomic sequence regions were identified using the process described in Subsection 4.2.2 and were used input to the region identification process.

4.5.3 Visualization of Identified Regions

Results from the region identification process were visualized using the IGV genome browser v2.19.1 [46]. An example of an IGV screenshot is shown in Figure 4.7.

Algorithm 1 The general algorithm underlying the region identification process.

INPUT: list of GFF files, $GFFList$

RETURNS: list of regions, $regionList$

for `<each GFF in GFFList>` **do**

  $allFeatures \leftarrow GFF.features$

end for

$sortedFeatures \leftarrow sortStartCoord(allFeatures)$

$regionList \leftarrow []$

$firstFeature \leftarrow firstItem(sortedFeatures)$

$lastFeature \leftarrow lastItem(sortedFeatures)$

for `<each feature in sortedFeatures>` **do**

  if $feature = firstFeature$ **then**

    $currentRegion \leftarrow newRegion(feature)$

    $currentRegion.start \leftarrow feature.start$

    $currentRegion.end \leftarrow feature.end$

  else if $feature = lastFeature$ **then**

    $regionList.append(currentRegion)$

    $return\ regionList$

  else if $feature\ overlaps\ currentRegion$ **then**

    $currentRegion.updatePositions(currentRegion, feature)$

    $currentRegion.addFeature(feature)$

  else

    $regionList.append(currentRegion)$

    $currentRegion \leftarrow newRegion(feature)$

    $currentRegion \leftarrow feature.start$

    $currentRegion \leftarrow feature.end$

  end if

end for

$return\ regionList$

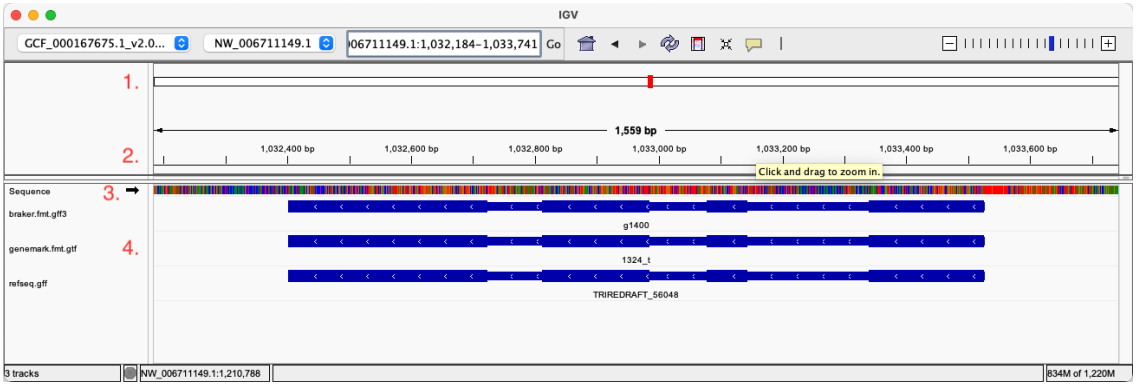


Figure 4.7: An example of an IGV screenshot used in portions of the results from this work. The top half of the image displays information about the reference sequence with track one showing the position of the viewer relative to the reference sequence, and track two showing the numerical positions of the reference. The lower half of the image displays more tracks in the form of additional layers of information mapped to the reference sequence. Track three shows a track displaying the nucleotide at each position if the scale of the viewer is small enough. Track four comprises several tracks, each containing features, the regions in blue, from a GFF file. The tracks in this example are limited to gene predictions, although tracks in other IGV screenshots may include features from other tools such as InterProScan and will be discussed in their respective sections. Individual tracks will be referred to in increasing numerical order in the track list beginning from the top. The nucleotide track is not included in the order.

5 Results and Discussion

5.1 Overview

Results from this work are ordered in a particular manner relative to the order in which data was processed. Since there are novel *Trichoderma* assemblies included in this work, the results begin with an overview of the assemblies of DC1 and Tsth20 in Section 5.2. Following this, is Section 5.3, which discusses the installation, application and other features associated with the gene finders selected for this work. Results of the initial gene finding process are then presented in Section 5.4 followed by more detailed analysis of predicted CDS lengths in Section 5.5. Gene predictions from each gene finder and assembly combination are supplied to the BUSCO metric process in Section 5.6. Results from the initial spatial comparison of gene predictions using the region identification process are then presented in Section 5.7. The region identification process is then applied with additional information from tblastn alignments, InterProScan annotation, and GC sequence composition in Sections 5.8, 5.9 and 5.10, respectively. A final comparison of all results and selection of an appropriate gene finder are then discussed in the Conclusion 6.1.

RefSeq assemblies in these results are referred to by their scientific name for context and not by their associated NCBI accession IDs. Instead, NCBI accessions for the RefSeq assemblies and their associated annotations are listed here: *T. reesei* - GCF_000167675.1_v2.0, *T. harzianum* - GCF_003025095.1_Triha_v1.0, GCF_000170995.1_TRIVI_v2.0.

5.2 Assemblies of DC1 and Tsth20

For general assembly metrics of DC1 and Tsth20, the QUAST [**?**] tool was used. Results from QUAST are shown in Table 5.1, from which we can make several observations. In DC1 and Tsth20, the total contig counts are an order of magnitude smaller when compared to the other NCBI RefSeq assemblies, indicating highly contiguous assemblies from nextDenovo [30] and nextPolish [30]. This is likely due to the use of Nanopore sequencing in the assemblies of DC1 and Tsth20. The total assembled lengths of DC1 and Tsth20 are similar when compared to RefSeq assemblies, ranging from 38Mb to 42Mb, except in the case of *T. reesei*, which is known to have a significantly smaller genome length [4], at roughly 33Mb.

The largest contig size for each assembly varies greatly. DC1 and Tsth20 have the largest contigs of all assemblies being considered, which is again likely due to the inclusion of long-read sequencing data in the assembly process. The N50 values for all assemblies are above 1Mb, with DC1 and Tsth20 N50s being

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

Table 5.1: General assembly metrics produced by QUAST[**?**] (a genome quality assessment tool).

at minimum two times larger than others assemblies. N50 lengths nearing chromosomal lengths indicates that assemblies of DC1 and Tsth20 are better assembled than the other *Trichoderma* assemblies. While chromosome-scale contigs are not the sole indicator of genome quality, it does provide confidence in the quality of the input data and resulting assemblies. In general, the assemblies of DC1 and Tsth20 are of similar length to existing *Trichoderma* assemblies and the number of contigs reported match the number of 'chromosome' scale contigs reported in other work [4].

During initial investigation of the inputs to the assembly process, we observed that the Illumina reads have a bimodal distribution of GC content as shown in Figure5.1. To see if this observation extended to the assemblies as well, 250 bp sliding windows were used to calculate GC content for all assemblies included in this analysis. The results of this analysis are shown in Figure 5.1. AT-rich sequences are visualized on the left peak of the distributions sequences containing 90 percent AT content. Of the included assemblies, AT-rich sequences were identified in DC1, Tsth20, *T. reesei* and *T. harzianum*, with *T. virens* deviating from the other assemblies showing very few AT-rich windows. This lack of AT-rich sequence in *T. virens* nucleotide composition may indicate potential assembly issues or even misclassification of the organism. In addition to the confirmation of increased AT-rich sequence content in most assemblies, it appears that the distribution of GC content in *T. reesei* differs from the other assemblies. The curve of GC content for *T. reesei*, visualized in green in Figure 5.1, is shifted to the right of the other *Trichoderma* assemblies, indicating fewer AT-rich windows and more balanced nucleotide composition in its assembly. While the left tail of the curve also shows an increase in AT-rich sequence composition, with it's peak is again located farther right on the X-axis than other *Trichoderma* assemblies. We can not explain exactly why *T. reesei*'s curve differs so much at this time, but it is likely due to the much smaller *T. reesei* assembly and it's properties. Investigation of these AT-rich sequences is continued in Section 5.10, where only sequences containing less than greater than 72% AT nucleotide content are considered, as the distributions begin to deviate from the normal distribution at that point.

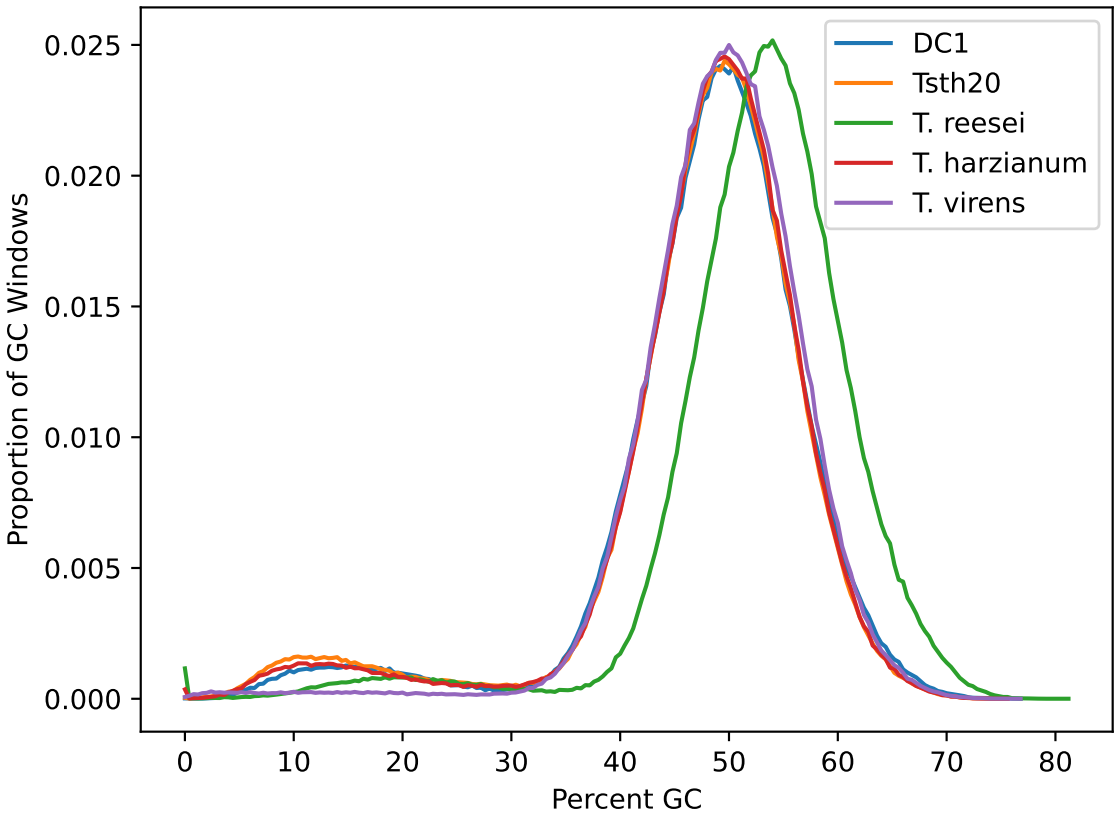


Figure 5.1: Plots showing proportions of sliding windows of GC content for each *Trichoderma* genome assembly.

5.3 Installation and Profiling Gene Finding Tools

While Braker2, GeneMark and RefSeq all predict coding sequences of possible genes for a provided input, the implementation of each tool is different, requiring more or less effort to install and run than other tools. This section discusses the implementations, installation procedures and execution of GeneMark and Braker2 in more detail. For more information on the versions and parameters used in this processing, please see Section 4.3.3. The computing platform used in this research is hosted and managed by University of Saskatchewan services, which is modelled around the HPC platform used by the Digital Research Alliance of Canada and software is managed similarly.

GeneMark [15] is a gene finding tool developed by the Georgia Institute of Technology with packages prepared for Linux and MacOS. It is provided as licensed product in the form of a package which can be downloaded from their website after submitting a form. Once the necessary information is submitted, the user is provided with a key that must be placed in the appropriate location once the software is downloaded and unpacked. The core controlling methods of GeneMark are written in Perl, accompanied by several Python scripts and compiled executables. GeneMark was tested by the developers with Perl version 5.10, and Python 3.3. A number of Perl dependencies are also required, which can be installed via CPAN. The user will have to know which implementation of GeneMark they are wanting to use for their application, as GeneMark has several variations it can run depending on the desired application. In this work, the GeneMark-ES variant of GeneMark was executed, as it is the self-training *ab initio* GeneMark method for eukaryotic organisms. Options required by GeneMark at runtime are documented in the help message, and simple enough that any user with familiarity of bioinformatics tools should be able to run GeneMark, although documentation for use is only provided by the help message when running the program and not online. In regards to run-time, running the GeneMark-ES pipeline on DC1 with 56 threads finished in 16 minutes. GeneMark's outputs consist of a GTF or GFF file of predicted genes as well as a number of other outputs related to the run.

Braker2 [19] is hosted on GitHub as a repository that receives relatively frequent updates with Braker3 being released while working before the completion of this work. Braker is maintained by Katharina Hoff from the University of Greifswald and is available under the Open Source Artistic License. Installation of the repository is a straightforward pull from GitHub. As with GeneMark, Braker2 uses a combination of Perl, Python and other executables in its regular use. Downloading the repository itself is not enough for execution, as Braker2 relies on a number of dependencies and bioinformatics tools including Perl and (Perl dependencies), Augustus [47], BamTools [48], BedTools [49], GeneMark [15], StringTie [50], GFFRead [51] and several others. Manual installation of these dependencies would be difficult, time consuming and in general advised against. In this case, many of Braker2's requirements are satisfied by modules already included in the environment, making installation relatively simple if one is familiar with the Digital Research Alliance of Canada's software stack. This case still required installation of some Perl modules in addition to loading necessary modules. Alternatively, one could use a package manager like Conda to handle installation of packages. This is perfectly

reasonable, but also requires knowledge specific to Conda, which can be complicated and frustrating for users with little software management experience. Once installation is finished, the Braker2 pipeline is relatively straightforward to run as well, with excellent documentation included both online and through the built in help message. The training, including RNAseq alignments, and gene finding pipeline in this case took 1 hour and 17 minutes using 60 threads. Applying the Braker2 trained gene finding model to DC1 with 60 threads took 21 minutes to complete. Run times will of course vary depending on processing power available to the end user, but in the case of *Trichoderma* genomes, users can expect quick results with relatively little computing power. Once annotation is complete, Braker2 produces a GFF file containing predicted genes along with CDS sequences and amino acid sequences their protein products.

In summation, Braker2 and GeneMark are not direct plug-and-play software packages. Users should expect to encounter issues when getting these programs running in addition to normal downloading and unpacking of software packages so some expertise is recommended. Neither Braker2 nor GeneMark require users to compile software, however Braker2's dependencies may require additional compilation and attention. Once installed, both tools are relatively simple to use with documentation available for both on the command-line and excellent documentation available for Braker2 on their GitHub page. Outputs from both tools are similar although Braker2 has the ability to output coding and amino acid sequences for downstream processing. Both tools run in reasonable amounts of time, where in the case of smaller genomes such as *Trichoderma*, users can expect results within a few hours to a day depending on number of computing cycles available to them.

5.4 Initial Gene Finding Results

Prior to discussing gene finding results, we will first define the terms gene and coding sequence in the context of this work. We refer to a gene as a set of start and stop coordinates in a genomic sequence. This definition of a gene simplifies processing, although the definition could be expanded in the future to include introns, exons and potential up and downstream sequences as well. Gene finders may also predict isoforms, or alternative splice variants of a gene based on evidence provided in training, resulting in multiple potential coding sequences for a gene. In this work, we refer to coding sequences as the set of all coding sequences predicted by a gene finder.

Counts of genes and coding sequences predicted by Braker2, GeneMark and RefSeq are shown in Figure 5.2, Figure 5.3 and Table 5.2. Immediately we see that Braker2 predicts far fewer genes in all assemblies, except in the case of *Trichoderma reesei.* This is possibly due to the effects of training the Braker2 gene model using data from *Trichoderma reesei*, which has a significantly smaller genome in comparison to other *Trichoderma* assemblies, although genome size is not always indicative of gene content. Regardless, we observe a difference in the number of genes predicted by Braker2 in comparison to GeneMark and RefSeq. The number of genes predicted by GeneMark and RefSeq are similar, except in the case of *T. harzianum*, in which RefSeq predicts roughly 17% more genes than GeneMark. Braker2 consistently predicts more coding sequences than

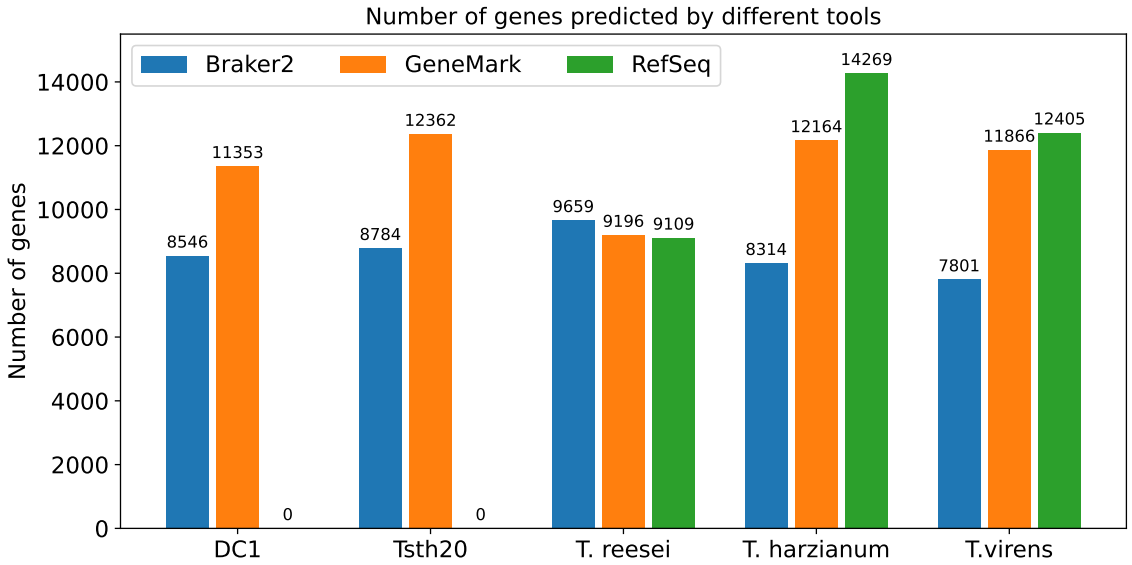


Figure 5.2: Number of genes predicted by each gene finder for each *Trichoderma* genome assembly.

Assembly	Braker2		GeneMark		RefSeq	
	Genes	CDS	Genes	CDS	Genes	CDS
DC1	8546	8637	11353	11353	N/A	N/A
Tsth20	8784	8858	12362	12362	N/A	N/A
T. reesei	9659	10175	9196	9196	9109	9118
T. harzianum	8314	8385	12164	12164	14269	14090
T. virens	7801	7863	11866	11866	12405	12406

Table 5.2: Number of genes and coding sequences (isoforms) predicted by each gene finder for each *Trichoderma* genome.

GeneMark and RefSeq. RefSeq also appears to predict multiple coding sequences for each gene but in fewer numbers than Braker2. Coding sequence prediction counts in *T. harzianum* from RefSeq are also interesting, with RefSeq predicting fewer coding sequences than genes. Why this occurs is unknown but may warrant further investigation.

5.5 Distribution of Predicted Coding Sequence Lengths

To better understand and compare the distributions of CDS sequences predicted by Braker2, GeneMark, and RefSeq, the cumulative distribution function for the lengths of CDS sequences for each gene finding tool are shown in Figures 5.4, 5.5 and 5.6. The $\log_{10}$ values of gene lengths were used as the abscissa for a better visulazation of the distributions. In DC1, the curves from Braker2 and GeneMark follow each other closely, with the only variation being genes of short length, where Braker2's curve extends beyond that of GeneMark,

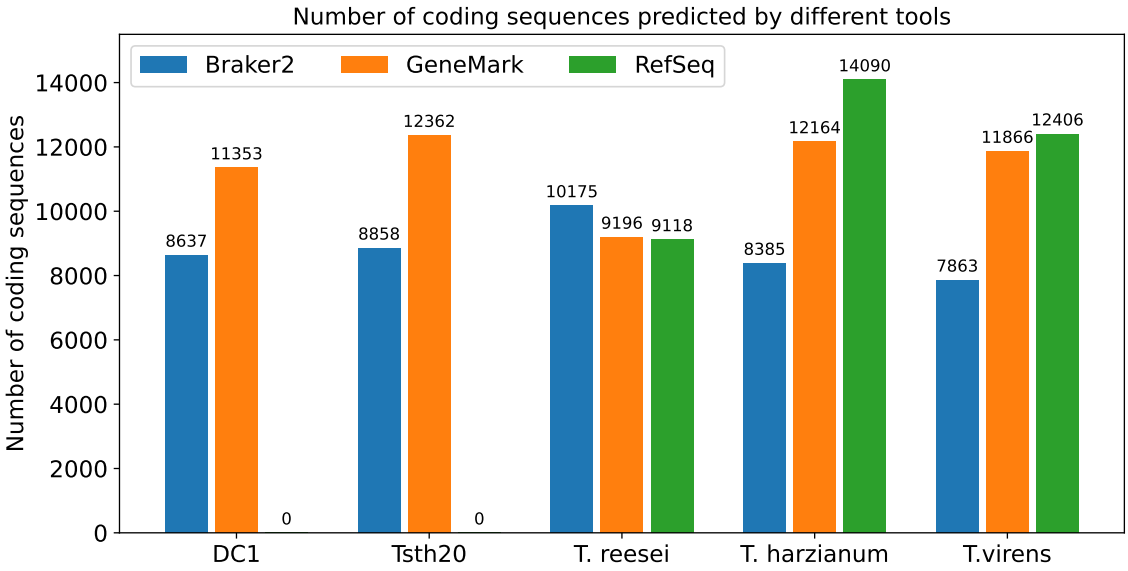
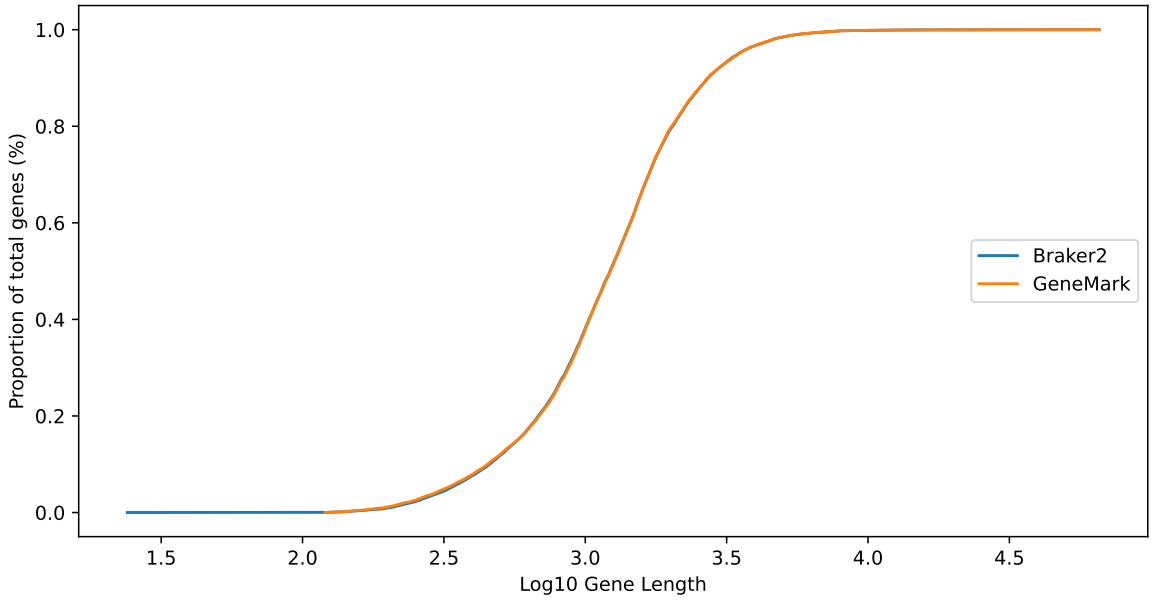


Figure 5.3: Number of coding sequences predicted by each gene finder for each *Trichoderma* genome assembly.

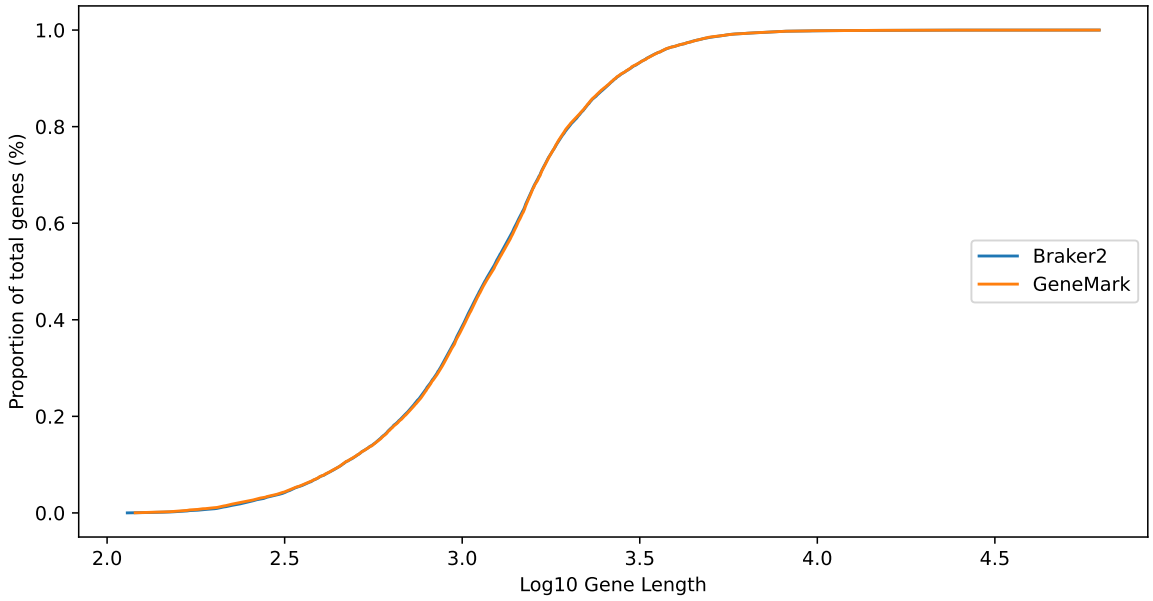
indicating that Braker2 predicts the shortest genes in all assemblies except *T. virens*. In the case of Tsth20, the curves are nearly identical. In *T. reesei*, we see disagreement in the curves for shorter genes, with Braker2 appearing to predict a larger fraction of shorter genes than GeneMark and RefSeq. The right sides of the curves trend toward similar predicted gene lengths. In *T. harzianum*, RefSeq deviates from GeneMark and Braker2, predicting more genes of short length, while Braker and GeneMark appear to be in near-complete agreement except in the case of very short genes. Finally, in *T. virens*, we see the RefSeq curve predicting a larger fraction of shorter genes once again, although the deviation is not as drastic as in *T. harzianum*.From these plots, we can say that visually, gene finding tools appear to predict different lengths of genes. We also observe that Braker2 typically predicts the shortest genes of all three gene finding methods.

To confirm that CDF curves differ, two-sided two-sample Kolmogorov-Smirnov[**?**] tests were performed using the $\log_{10}$ transformed gene lengths, with the null hypothesis being that frequency of genes predicted in AT-rich and normal genomic sequence are same. Results are presented in Table 5.3. In the cases of DC1 and Tsth20, we see that in agreement with Figure 5.4, Braker2 and GeneMark do not produce statistically different lengths of genes, which is interesting considering that the Braker2 includes experimental evidence for another assembly while GeneMark does not. In *T. reesei*, Braker2's predicted gene lengths are significantly different from both RefSeq and GeneMark, which is also evident in the CDF plots. The same cannot be said for *T. harzianum* and *T. virens*, where RefSeq is significantly different from both GeneMark and Braker2, which are not significantly different from each other. It is also notable that RefSeq and GeneMark predict similar gene lengths in *T. reesei*, but not in *T. harzianum* and *T. virens*.

It can clearly be stated that these gene finding tools predict different distrbutions of gene lengths, paritcularly in *T. reesei, T. harzianum* and *T. virens*. Why that may be the case is difficult to answer without

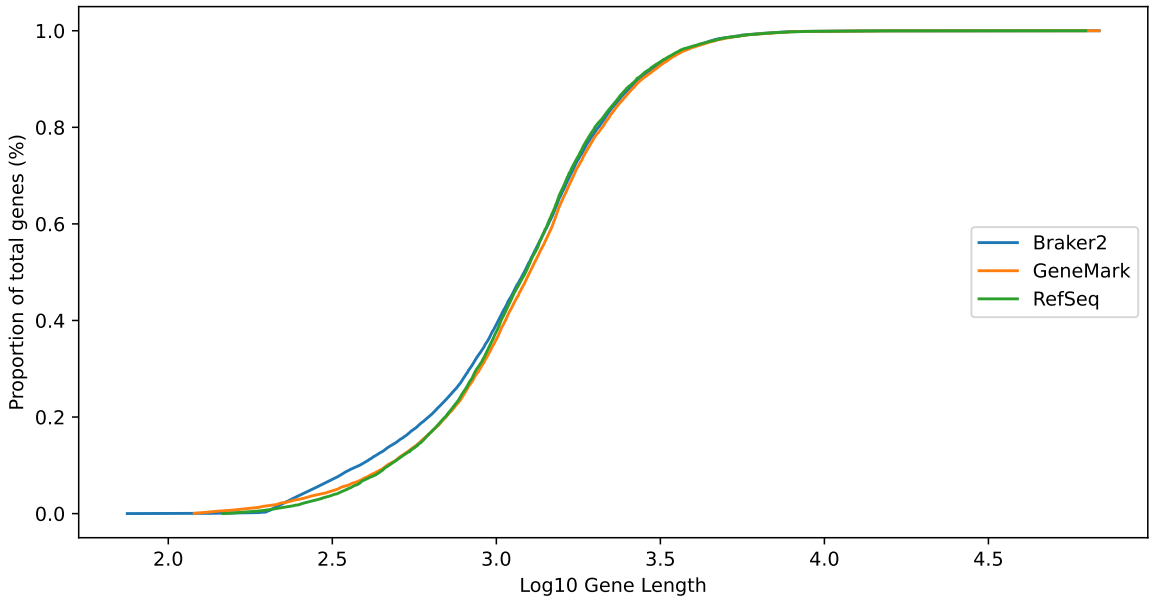


(a) DC1

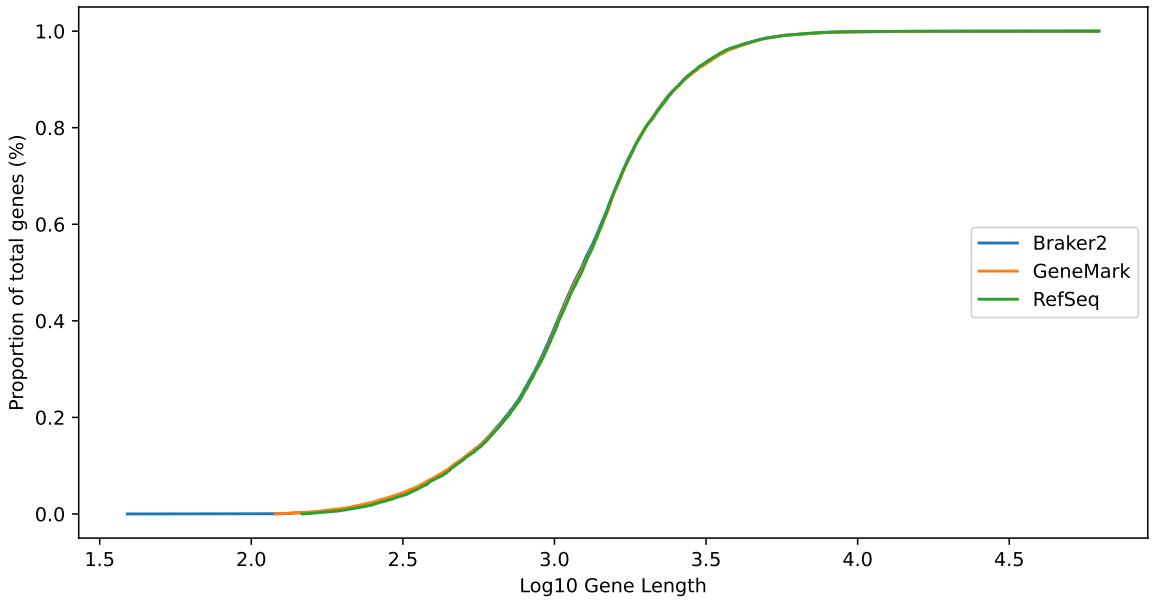


(b) Tsth20

Figure 5.4: Plots of the cumulative distribution function for CDS lengths produced be each gene finding tool applied to DC1 and Tsth20

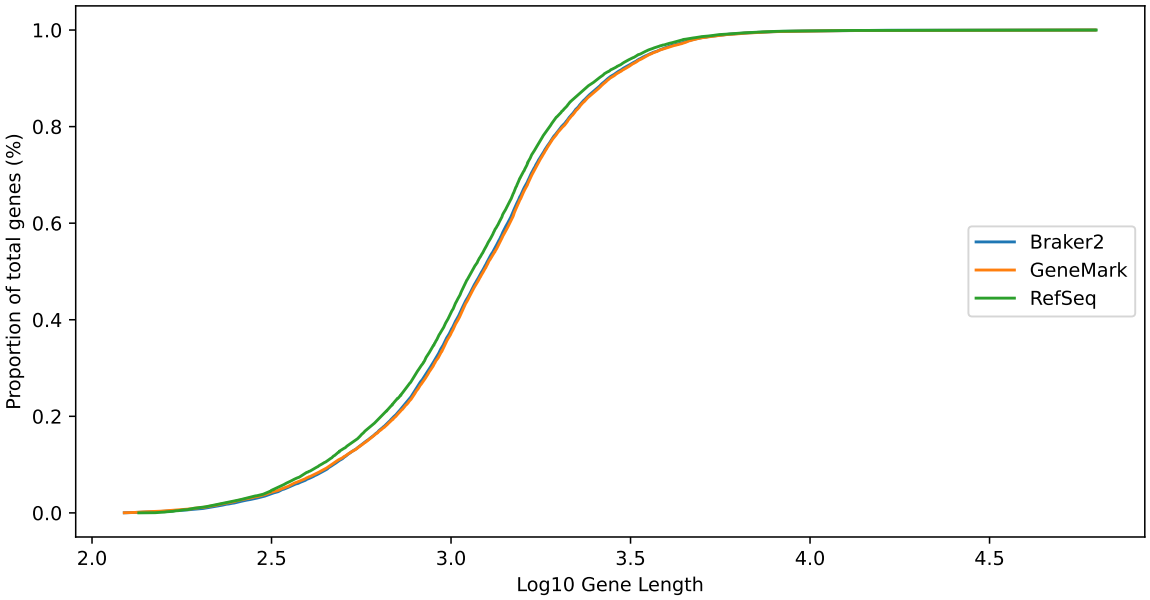


(a) *T. reesei*



(b) *T. harzianum*

Figure 5.5: Plots of the cumulative distribution function for CDS lengths produced be each gene finding tool applied to *T. reesei* (GCF_000167675.1_v2.0) and *T. harzianum.* (GCF_003025095.1_Triha_v1.0)



(a) *T. virens*

Figure 5.6: Plots of the cumulative distribution function for CDS lengths predicted by each gene finding tool when applide to *T. virens* (GCF_000170995.1_TRIVI_v2.0).

Genome	Tool #1	Tool #2	*P*-value
DC1	Braker2	GeneMark	0.999
Tsth20	Braker2	GeneMark	0.965
T. reesei	Braker2	GeneMark	$9.481 * 10^{-07}$
T. reesei	GeneMark	RefSeq	0.002
T. reesei	Braker2	RefSeq	$1.340 * 10^{-07}$
T. harzianum	Braker2	GeneMark	0.863
T. harzianum	GeneMark	RefSeq	$4.313 * 10^{-52}$
T. harzianum	Braker2	RefSeq	$4.674 * 10^{-55}$
T. virens	Braker2	GeneMark	0.635
T. virens	GeneMark	RefSeq	$7.352 * 10^{-12}$
T. virens	Braker2	RefSeq	$1.794 * 10^{-09}$

Table 5.3: Table of *P*-values from two-sided two-sample Kolmogorov-Smirnov tests between gene finding tools.

deeper invesitgation. There is clearly an underlying difference between RefSeq and the other gene finding tools. Braker2 and GeneMark tend to be in agreement, except in the case of *T. reesei*, for which Braker2 was specifically trained. We observe that when Braker2 is applied to an assembly from which the training data originated, Braker2 predicts a larger fraction of shorter genes. Conversly, we observe that when Braker2 is trained with experimental evidence and applied to a *Trichoderma* assembly from which the training data did not originate, the distributions of predicted genes lengths do not significantly differ from the *ab initio* gene finder GeneMark.

5.6 BUSCO Results

Results of BUSCO analysis using the sordariomycetes_Odb12 dataset provided by BUSCO are presented in Figure 5.7, and Table 5.4. The results indicate that all gene sets considered in this analysis have a BUSCO completeness of 94.1% or higher, with a maximum completeness of 98.4% in the case of Braker2 and DC1. In general, Braker2 and RefSeq have the most BUSCO complete sets of gene predictions of the three tools considered. Interestingly, Braker2 produces far more duplicated BUSCO matches than both GeneMark and RefSeq. Examining the BUSCO output logs, this appears to be due to Braker2 predicting more than one coding sequence for some genes predictions, resulting in multiple similar proteins. Interestingly, the coding sequences in the RefSeq annotations seem to miss more genes than the other two gene finders while also having a higher number of fragmented BUSCO genes. This may be due to human curation of the RefSeq datasets, or the Gnomon gene prediction pipeline used by NCBI to produce these annotations. Further investigation is required to determine the exact cause of this discrepancy. Finally, it appears that *T. reesei* tends to have slightly lower BUSCO completeness than the other *Trichoderma* species considered in this analysis, regardless of gene finder used. Why this is the case is unknown, but may be due to the evolutionary distance between *T. reesei*, the Gnomon annotation process, or potentially the fragmented nature of the *T. reesei* assembly used in this analysis. While these results do not capture the entire set of genes possibly present in these *Trichoderma* assemblies, they do confirm that the gene finders are at minimum predicting many evolutionarily conserved fungal genes.

While BUSCO matches are a good metric for general performance of gene finders, it is also important to investigate BUSCO proteins that were missing from the gene predictions. Of interest are BUSCO proteins that are systematically missed by a gene finder in multiple or all assemblies, and whether or not these BUSCO proteins are also missed by other gene finders. Table 5.5 lists BUSCO IDs and their corresponding annotations that were not found in any set of predictions from Braker2, GeneMark or RefSeq. There are 17 BUSCO proteins that were not found by any of the three gene finders in any of the five *Trichoderma* assemblies. The annotations of these BUSCO proteins do not indicate any obvious reason why they would be systematically missed by all three gene finders, and further investigation is required to determine why these genes are not being predicted. It is possible that these genes are simply not present in these *Trichoderma*

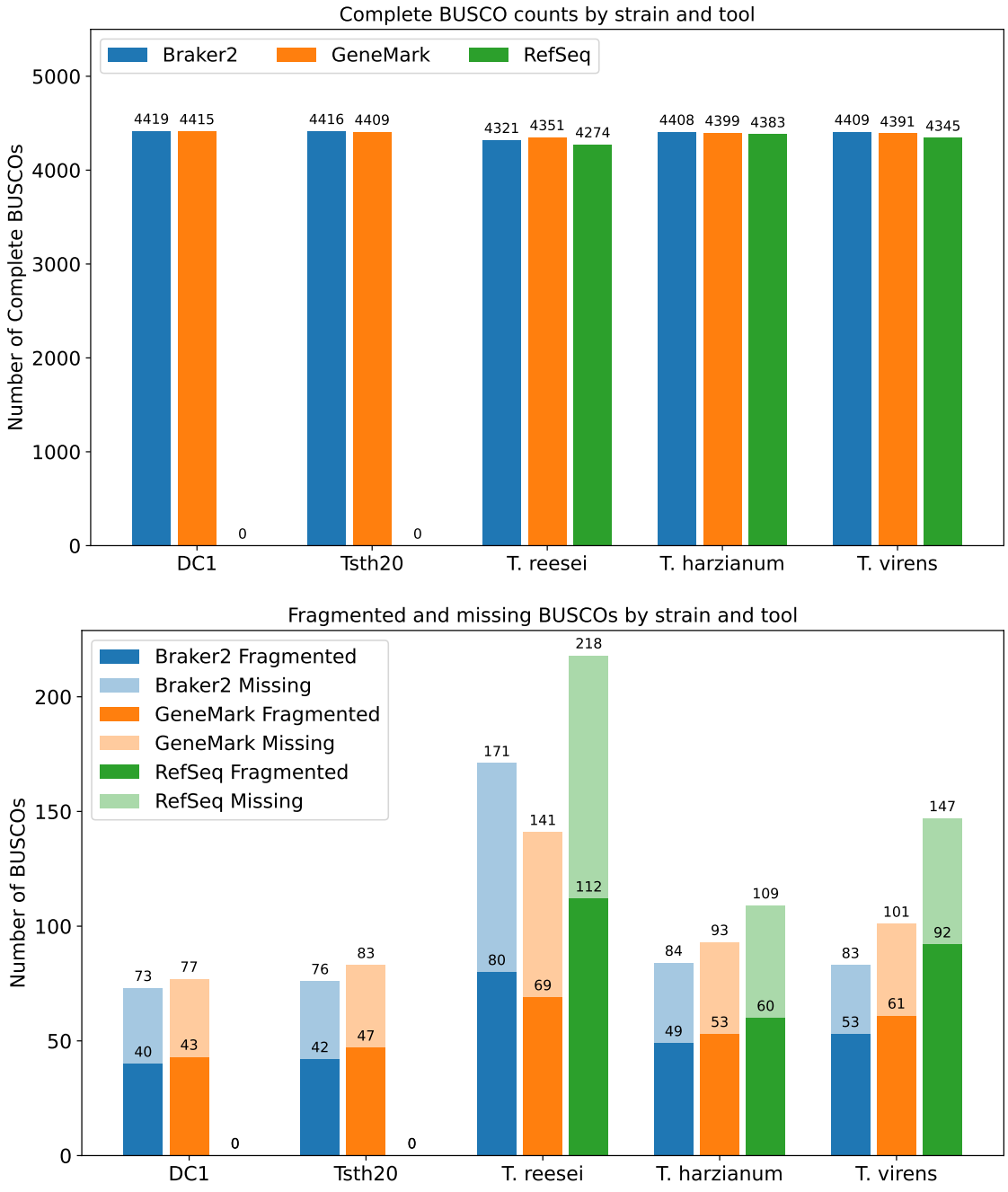


Figure 5.7: BUSCO complete and missing gene counts for each gene finder across all *Trichoderma* genome assemblies. There were a total of 4492 markers in the selected BUSCO dataset. For DC1 and Tsth20, RefSeq annotations are not available, so their values are set to 0. The selected sordariomycetes_Odb12 dataset contains 4492 markers.

Strain	Complete	Single	Duplicated	Fragmented	Missing
DC1	4419	3547	872	40	33
Tsth20	4416	3585	831	42	34
T. reesei	4321	3587	734	80	91
T. harzianum	4408	3572	836	49	35
T. virens	4409	3530	879	53	30

(a) Braker2

Strain	Complete	Single	Duplicated	Fragmented	Missing
DC1	4415	4401	14	43	34
Tsth20	4409	4392	17	47	36
T. reesei	4351	4345	6	69	72
T. harzianum	4399	4382	17	53	40
T. virens	4391	4369	22	61	40

(b) GeneMark

Strain	Complete	Single	Duplicated	Fragmented	Missing
T. reesei	4274	4267	7	112	106
T. harzianum	4383	4366	17	60	49
T. virens	4345	4321	24	92	55

(c) RefSeq

Table 5.4: Results from BUSCO using the fungal analysis option organized by gene finding tool. The sordariomycetes_Odb12 dataset contains 4492 markers. For more information on the categories assigned by BUSCO, please refer to the documentation.

species, but this may be unlikely given the evolutionary conservation of BUSCO proteins. Each gene finder also has one or more BUSCO proteins that are missed in all assemblies but not by one or both of the other gene finders. These genes are presented in Table 5.6. Again, the annotations of these BUSCO proteins do not indicate any obvious reason why they would be systematically missed by a particular gene finder, and further investigation is required to determine why these genes are not being predicted. While not presented here, we also identified a number of other genes that are missed by the gene finders which may provide insight into their performance and the evolution of *Trichoderma*. Overall, it appears that there are very few BUSCO proteins that are systematically missed by all gene finders, indicating that the gene finders are generally capable of predicting the majority of conserved fungal genes.

BUSCO ID	Annotation
191658at147550	Rossmann-fold NAD(+)-binding protein
250641at147550	Aspartic-type endopeptidase
279527at147550	Phosphatidic acid-preferring phospholipase A1, contains DDHD domain
578862at147550	Transcription factor
579967at147550	Alpha-L-rhamnosidase C
627082at147550	Ferric reductase
628206at147550	Zinc finger domain-containing protein
636196at147550	ATP-dependent RNA helicase
646880at147550	Lipase class 3
647592at147550	Conserved hypothetical protein
652375at147550	Conserved hypothetical protein
657983at147550	Conserved hypothetical protein
658912at147550	Aromatic amino acid aminotransferase
659938at147550	HAUS augmin-like complex subunit 1
672116at147550	Nitrogen regulatory protein AreA, GATA-like domain
677025at147550	Mating-type switching protein Swi10
689133at147550	Myosin heavy chain

Table 5.5: BUSCO IDs and their corresponding annotations that were not found in any set of predictions from Braker2, GeneMark or RefSeq.

Braker2, GeneMark and RefSeq all demonstrate excellent coverage of the BUSCO fungal protein set, indicating that these gene finders are capable of predicting genes that are expected to be present in these assemblies. From this we can say that the foundations of the underlying gene models used by each gene finder are solid. Braker2 produces more duplicate matches than GeneMark and RefSeq, but this is likely due to multiple isoforms of possible genes being present in the input data. Despite excellent coverage of the BUSCO fungal proteins, all three gene finders miss some BUSCO proteins in their predictions. Finally, we

BUSCO ID	Tool	Annotation
282444at147550	RefSeq	Conserved hypothetical protein
291262at147550	RefSeq	HAUS augmin-like complex subunit 6, N-terminal
632369at147550	RefSeq	Peptidyl-prolyl cis-trans isomerase, FKBP-type
632579at147550	RefSeq	Cyanate hydratase
672871at147550	RefSeq	MAU2 chromatid cohesion factor
677279at147550	RefSeq	Pentatricopeptide repeat domain-containing
688724at147550	GeneMark/Braker2	Putative protein of unknown function

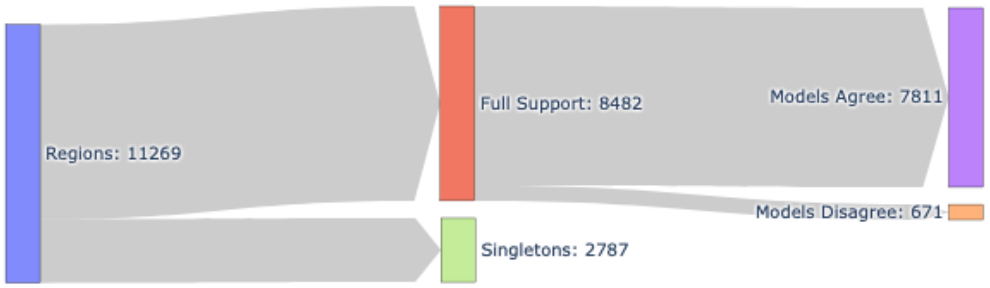
Table 5.6: Additional BUSCO IDs along with their corresponding annotations not found in any set of predictions. The tools that missed each BUSCO ID are also listed.

reiterate the possibility that human curation of RefSeq datasets is responsible for these differences, but this requires further investigation.

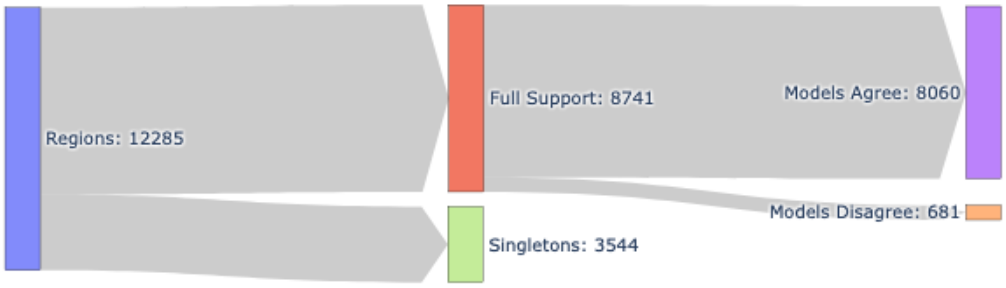
5.7 Region Identification

Visual breakdowns of the types of regions identified and their counts for each *Trichoderma* assembly are shown in Figure 5.8. We will discuss the types of regions, beginning with fully supported regions. These regions, labelled 'Full Support', are sections of genomic sequence where all three gene finders agree that a gene is present in some form. These fully supported regions are then broken down into two categories, based on whether or not the models from each gene finder agree on the start and/or stop positions of the gene model. Regions that have support from more than one gene finder, but not all, are labelled regions with 'Partial Support'. These regions are also broken down in to two subtypes based on whether or not the gene predictions agree on the start and/or stop positions of the gene. Regions with support from only one gene finder are labelled as singletons.

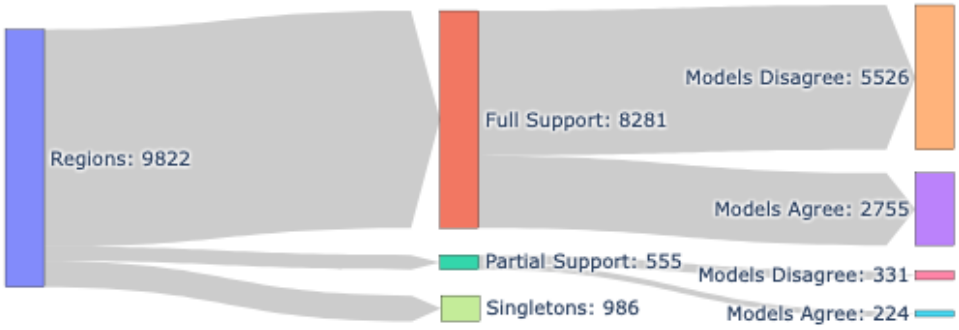
Looking at DC1 and Tsth20 in Figure 5.8, it appears that Braker2 and GeneMark predict genes in the same regions in the 69% and 65% of cases, respectively. For regions with full support in DC1 and Tsth20, the models also tend to agree on the start and stop positions of the gene in that region. This is not generally the case as we will see later when RefSeq is also considered. There are no regions with partial support for DC1 and Tsth20, as there are only two gene finders considered. *Trichoderma reesei* contains the fewest number of regions, which seems to scale appropriately with the total number of predicted genes and assembly length. Eighty-four percent of the regions are fully or partially supported by Braker2, GeneMark and RefSeq, with only 10 percent of the regions being single tool predictions. The most interesting observation here is the extent to which fully supported regions disagree on start and stop positions of the gene(s) in each region. There is clearly a difference in the gene models being produced by these gene finders. If there is also more disagreement on the start and stop positions of a gene, then there is likely disagreement on the number and



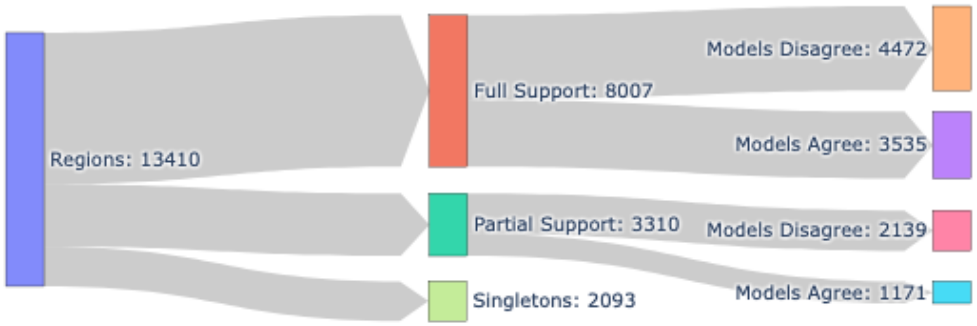
(a) DC1



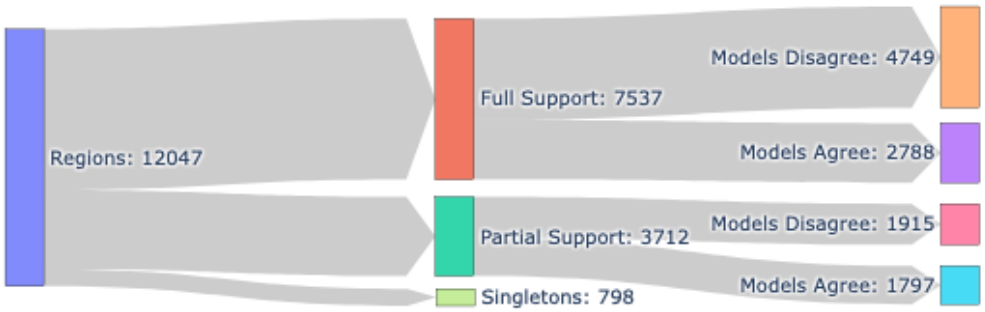
(b) Tsth20



(c) *T. reesei*



(d) *T. harzianum*



(e) *T. virens*

Figure 5.8: Figures showing breakdowns of genomic regions identified by the region finding process. Regions, in blue, are categorized based on support from gene finders. Regions with supporting predictions from all gene finders included in this analysis are labelled with 'Full Support' and colored red. Fully supported regions are then broken down into regions where gene models agree on the start and stop positions of the genes (purple), and those that do not (orange). Regions labelled with 'Partial Support' (turquoise) are regions with supporting gene predictions from two gene finders. Partially supported regions are also broken down into regions in which gene finders agree on the start and stop positions of the gene (cyan), and those that do not agree (pink). Regions with gene predictions from only one gene finder are labelled as singletons (green).

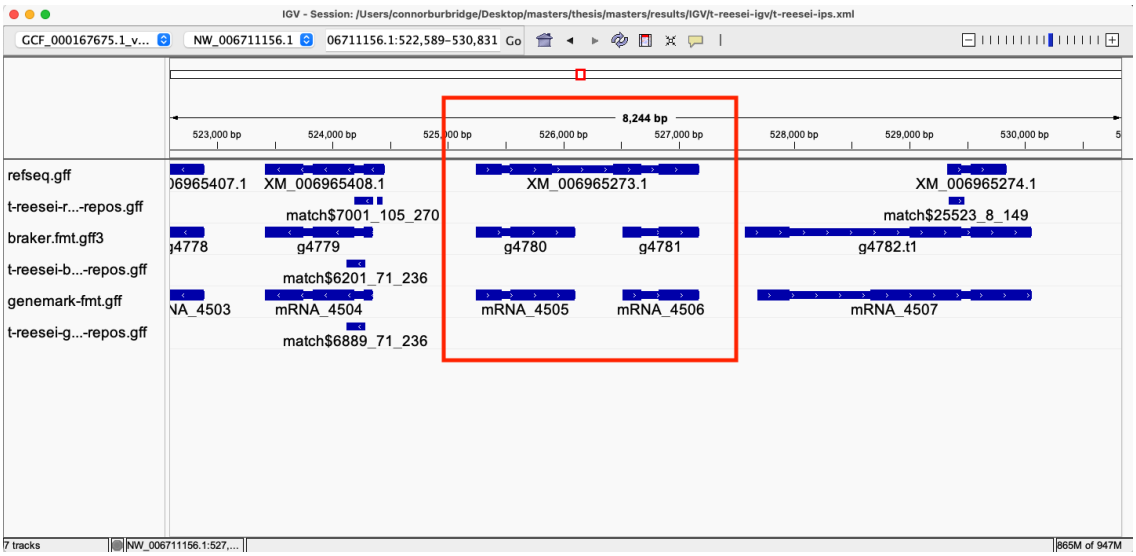


Figure 5.9: An IGV screenshot from *T. reesei* showing a region containing five gene predictions in which one of the gene finders disagrees with the other two, highlighted in the red box.

location of exons and introns within the gene model as well. In the case of partially supported regions, there is roughly 50% to 60% agreement on start and stop positions, but that may come as less of a surprise as there is already disgreement on presence by definition. Gene finding behaviour in regions from *T.harzianum* and *T. virens* differ from the other assemblies in the split between fully and partially supported regions. There seems to be fewer regions with full support from all gene finders and the reason why is unclear. The gene models in each region, whether fully or partially supported, still tend to disagree on start and stop positions of the genes more often than agree.

It is also worthwhile investigating some potentially interesting cases of strange gene calls in regions. One such case is when a region contains more than three individual predicted genes. The null hypothesis, so to speak, is that if all gene finders agree exactly on the presence and position of a gene in a region, then one should expect exactly three predictions in that region. We observe that this is not always true. Cases of more than three gene predictions in a region were present in regions from all processed assemblies. Figure 5.9 shows an example of one of these regions, in which the single gene predicted by RefSeq spans multiple gene predictions from both Braker2 and GeneMark. In the case of DC1 and Tsth20, there were very few regions that matched this scenario, with 19 and 6 regions containing more than 3 gene predictions, respectively. Conversly, *T. reesei, T.harzianum and T.virens* contain many such regions, with assemblies reporting 546, 899 and 521 regions with greater than three gene predictions, respectively. While having predictions from each gene finding tool present in a region is a strong indicator for the presence of a gene, having more than one gene prediction per gene finder raises questions about which model (or models) is correct and why disagreement exists.

Another interesting set of criteria to investigate is whether or not gene finders always predict genes on the same strand within a region. We observe that regions containing predictions on different strands do exist,

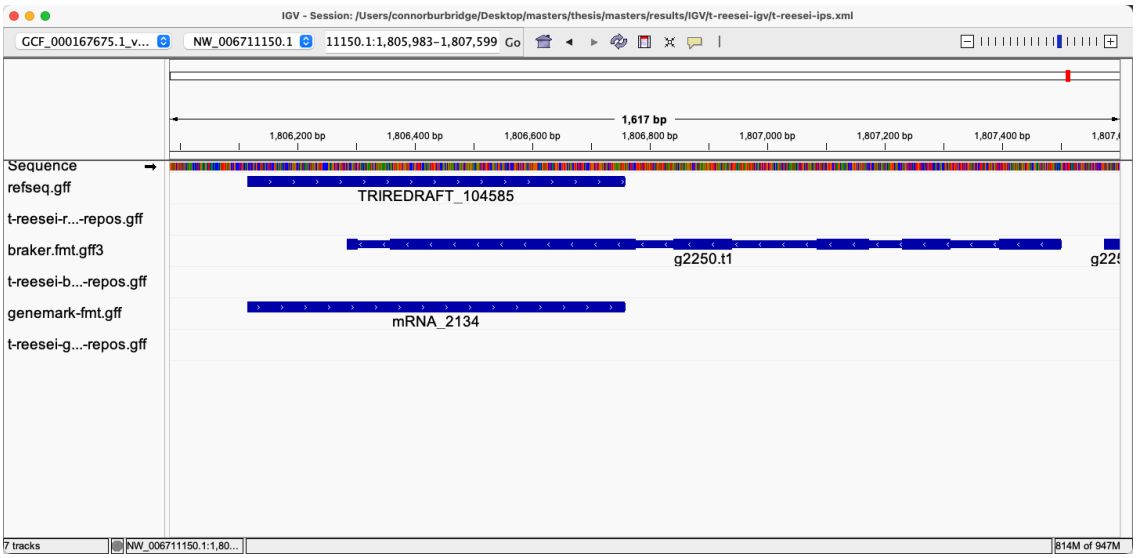


Figure 5.10: An IGV screenshot of *T. reesei* showing a region which contains gene predictions on opposite strands.

and are observed in predictions from all assemblies included in this analysis. Figure 5.10 shows an example of a region containing gene predictions on opposing strands. As in the case of regions with more than three gene predictions, DC1 and Tsth20 have fewer mixed strand regions with 29 regions identified in DC1 and 46 regions identified in Tsht20. Results from *T. reesei, T. harzianum* and *T. virens* show more regions in which this property is true, with region counts of 203, 533 and 293 respectively. Under the assumption that gene finders should predict the same genes on the same strands, it is unexpected to find so many of these cases, although it is possible that these overlapping sense and anti-sense genes exist naturally [**?**].

In summary, gene finders agree partially or completely on the presence of a gene in the vast majority of cases. While gene finders generally agree on the presence of a gene, they tend to disagree on the underlying gene model more often than they agree, except in in the case of DC1 and Tsth20. This is likely due to only including two gene finders in their analysis rather than three, resulting in fewer opportunities for disagreement. This observation is true when applied to the start and stop positions of the gene, but further investigation and comparison of intronic and exonic sequences between genes may provide more insight. It is also possible that genome quality may affect agreement between gene finders as the underlying model may be trained on contigs that are larger or of different structure than the assembly it is applied to. This is particularly true in the cases of DC1 and Tsth20, which are composed of larger contigs in comparison to the assemblies from NCBI. Finally, regions identified in this analysis do not always fit the ideal scenario of one gene prediction from each tool per region. Regions in which there are more than three gene predictions were observed in all assemblies included in this work. In addition, we observe cases where gene finders predict genes on different strands.

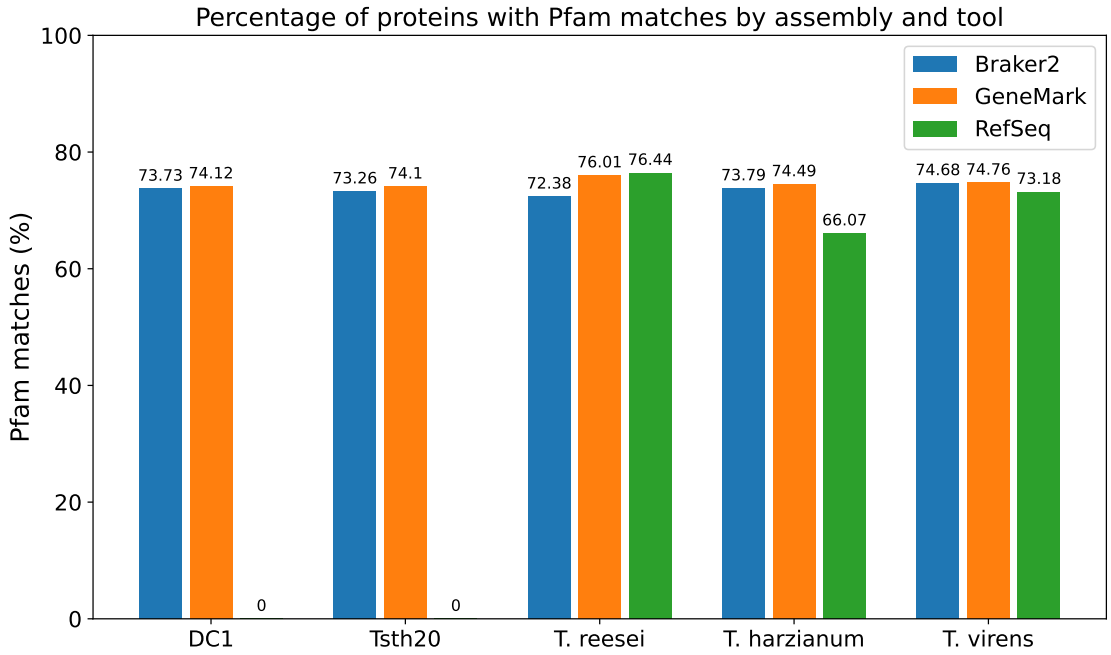


Figure 5.11: Barplot showing the percentage of predicted proteins with Pfam matches from InterProScan for each gene finder across all *Trichoderma* genome assemblies. For DC1 and Tsth20, RefSeq annotations are not available, so their values are set to 0.

5.8 InterProScan as Supporting Evidence for Predicted Genes

Pfam hits from InterProScan analysis are presented in Figure 5.11 and Table 5.7, from which we can identify one major trend. In general, roughly 73-76% of proteins predicted by Braker2, GeneMark and RefSeq contain a match to a Pfam entry, except in the case of *T. harzianum*, which reports a considerably lower proportion of genes with Pfam matches in the RefSeq dataset. Why the proportion of RefSeq proteins with Pfam matches in *T. harzianum* is so low is unknown. This is promising performance for the gene finders as the RefSeq annotations demonstrate similar proportions of Pfam hits to predicted proteins. The total counts may be deceiving however, as predictions from Braker2 may result in more than one protein product per gene, whereas in the case of GeneMark, only one protein is produced per gene model. From visual inspection of Pfam hits mapped back to the references, it appears that in general, when gene finders agree that a gene is present, InterProScan reports the same Pfam match in all three predictions. An example of agreement between Braker2, GeneMark, RefSeq and InterProScan is shown in Figure 5.12. It is important to note that the position of the Pfam match in IGV does not indicate the true position of the Pfam match in the gene, but provides an indication of the presence of Pfam matches. Pfam matches are offset from the start of the gene based on the start and end position of the Pfam match in the protein sequence. For example, if a Pfam match has a start position 10 amino acids into the protein sequence, the corresponding start position in the resulting GFF is 10bp downstream from the start of the gene.

Assembly	Braker2 (%)	GeneMark (%)	RefSeq (%)
DC1	73.73	74.12	N/A
Tsth20	73.26	74.10	N/A
T. reesei	72.38	76.01	76.44
T. harzianum	73.79	74.49	66.07
T. virens	74.68	74.76	73.18

Table 5.7: Table showing the fractions of predicted genes with Pfam annotations from InterProScan as a percent

While cases of complete agreement are abundant, cases of disagreement also exist and in strange forms. In many cases, while the gene finders agree on the presence of a gene, only the RefSeq protein product contains a match to the Pfam database. Why this may be the case is unclear, and may warrant further investigation. There are also many cases in which InterProScan reports the same Pfam hits for individual proteins, but the gene models in the region do not agree. There are even cases such as the region shown in Figure 5.13, where Braker2 and GeneMark agree that two genes and their associated proteins and Pfam matches are separate, but RefSeq only reports one gene with multiple Pfam hits. There are also several cases where two tools are in agreement with proteins containing Pfam hits while another is not. Even more interesting are cases such as the one shown in Figure 5.14, in which Braker and GeneMark predictions contain Pfam matches while RefSeq does not report a gene at all. This may be the due to experimental data used in the RefSeq training process or a result of curation. Regardless of the tool, these cases demonstrate well that gene finders are not always in agreement even on well known proteins, to the point that predictions are not present even though protein products from other gene finders contain known Pfam matches.

In summary, the protein products from Braker2 and GeneMark predictions do contain matches to the Pfam database. The proportions of matches to total proteins are similar to that of the RefSeq annotation, sitting between 65 and 75 percent. While Pfam hits to Braker2, GeneMark and RefSeq proteins generally agree, we do observe regions in which there is disagreement in several forms.

5.9 BLAST Results

As a control, proteins from three related organisms were used as queries to tblastn[**?**] on each *Trichoderma* assembly. The organisms and their associated NCBI accession IDs are as follows: *T. atroviride* - GCF_000171015.1, *F. graminarium* - GCF_000240135.3, *S. cerevisiae* - GCF_000146045.2. Results from the tblastn runs are presented in Figure 5.15 and Table 5.8. Initial tblastn results appear promising for both the *T. atroviride* and *Fusarium* datasets. All assemblies considered contain at minimum 89% of the reference protein sequences in the case of *T. atroviride* and a minimum of 75% in the case of *Fusarium*. In the case of *S. cerevisiae*, a minimum of 57% of reference proteins matched. It appears that the percentage of

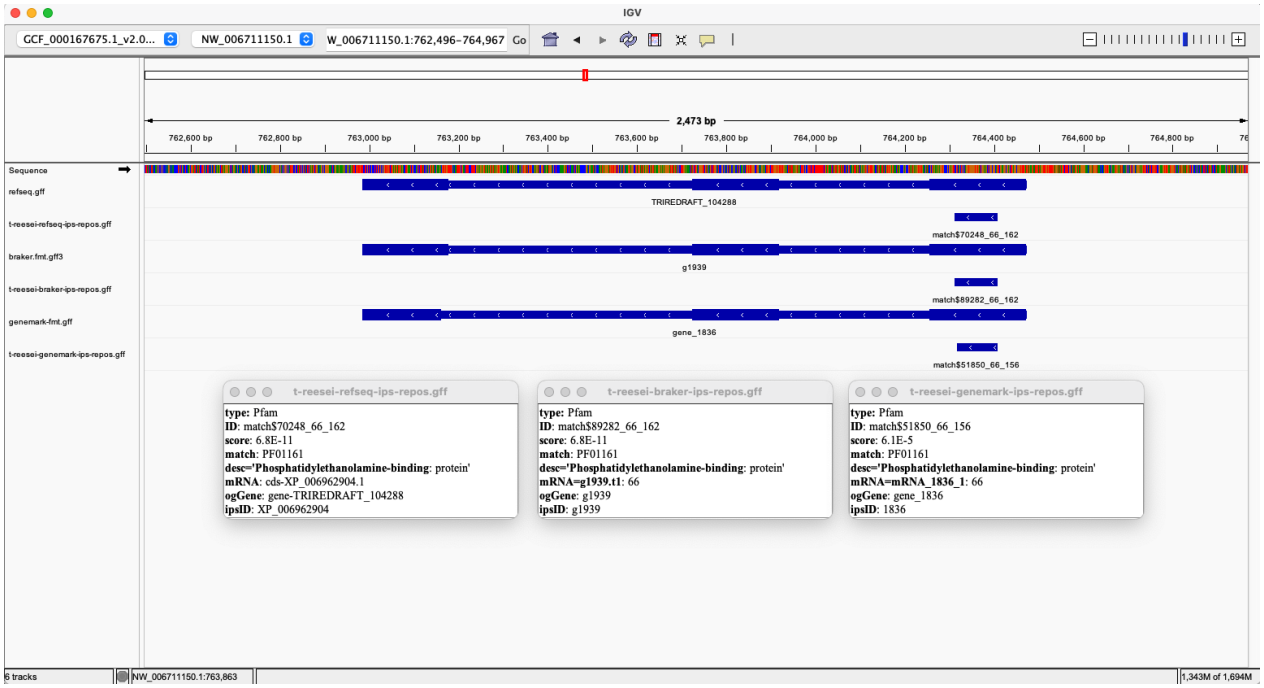


Figure 5.12: An IGV capture showing complete agreement between gene finders for both gene model and protein Pfam hits. The longer segments are predicted genes and the smaller segments are annotated Pfam matches, all with agreeing start and stop positions. The **desc** in each sub-window indicate agreeing Pfam annotations.

Reference	Ref. Proteins	DC1	Tsth20	*T. reesei*	*T. harzianum*	*T. virens*
Trichoderma atroviride	11807	11552	11080	10601	11081	11078
Fusarium granminarium	13312	10327	10429	10064	10434	10490
Saccharomyces cerevisiae	6014	3537	3517	3445	3509	3500

Table 5.8: tblastn hits from reference protein sequences to selected *Trichoderma* assemblies. Hits are reported if the alignment length is greater than 30% of the reference protein length and if 30% of the alignment length has identical matches.

hits decreases as evolutionary distance increases. These results provide rough validation that the assemblies contain potential for protein coding sequences.

While successful tblastn hits from input proteins are useful, it is worthwhile exploring those hits in the context of regions and gene predictions within them. Counts of genes from regions with tblastn hits are shown in Figures 5.16a, 5.16b, 5.16c, and Table 5.9. In a similar trend to Table 5.8, genes in regions with tblastn hits decrease as evolutionary distance increases. In DC1 and Tsth20, we observe that Braker2 predicts more genes in regions with tblastn hits than GeneMark. RefSeq is not included for DC1 and Tsth20, as there is no RefSeq annotation available. In *T. harzianum* and *T. virens*, Braker2 and RefSeq appear to have similar counts of genes in regions with tblastn hits. Interestingly, in the case of *T. reesei*, there are more genes from GeneMark in regions with tblastn hits, which is not the case in *T. harzianum* and *T. virens*.

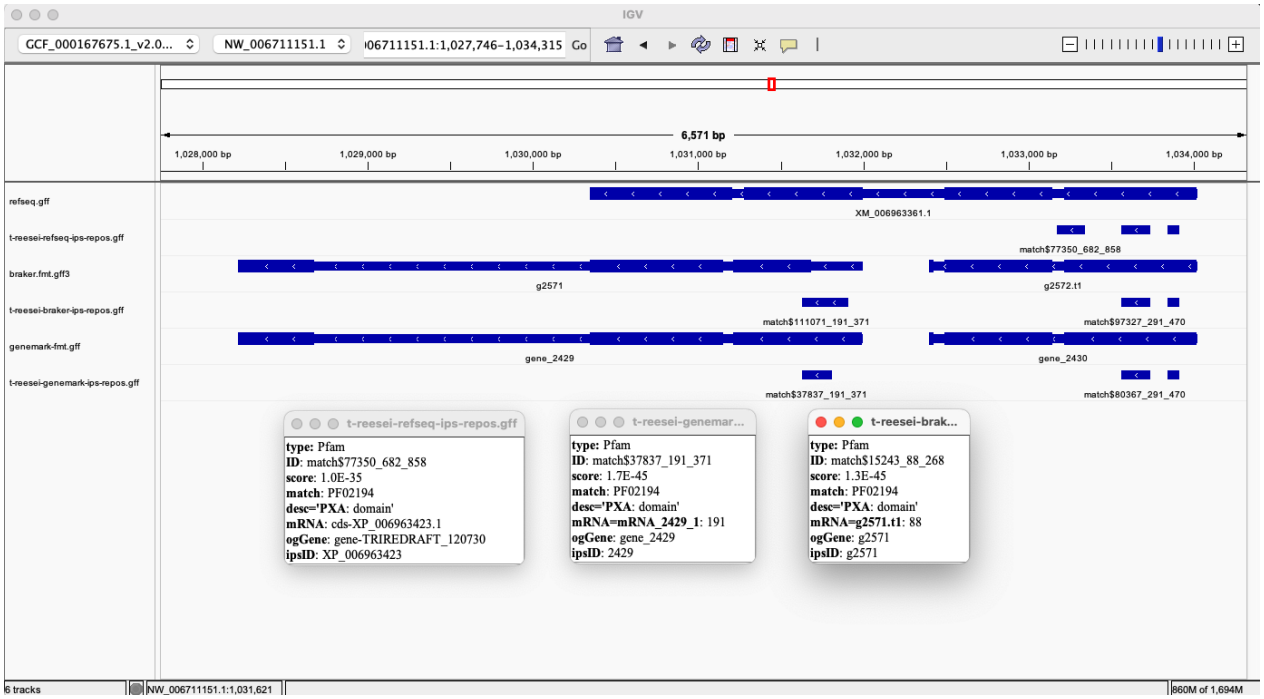


Figure 5.13: An IGV capture showing Braker2 and GeneMark reporting two genes and their resulting proteins and Pfam hits as separate, while RefSeq reports one gene, one protein and three Pfam matches.

Another interesting observation is comparing the coverage, or completeness, of the input and query sequence. As mentioned, it appears that high proportions of query sequences are being mapped to the subject, but the gene predictions are not as well represented in regions where those tblastn hits occur. A likely reason for this is inappropriate alignment parameters. More flexible gap parameters and the use of a different scoring matrix would likely produce better results, but optimization of parameters was not performed in this analysis. It is also possible that the post-alignment filtering criteria were inappropriate, resulting in short alignments that are true in nature, but may not be indicative of the presence of a truly similar protein.

Overall, it appears that using tblastn with proteins from related organisms may be a viable option as a control when no reference or model organism is available. At the very minimum, tblastn hits can be used as another piece of supporting evidence for the existence of a gene, with one caveat being imperfect coverage of predictions. The issue of imperfect coverage may be mitigated by parameter optimization. It is also possible that one could employ a similarity-based cutoff on the proportion of predicted genes in regions with tblastn hits as an indicator of gene finding performance, however this would require further analysis, preferably with more closely related organisms.

5.10 Genes in Regions of Anomalous GC Content

Figure 5.17 and Table 5.10 shows the results of applying the same region finding process from earlier to segments of the genome identified as AT-rich. AT-rich is defined as a region of genomic sequence with percent

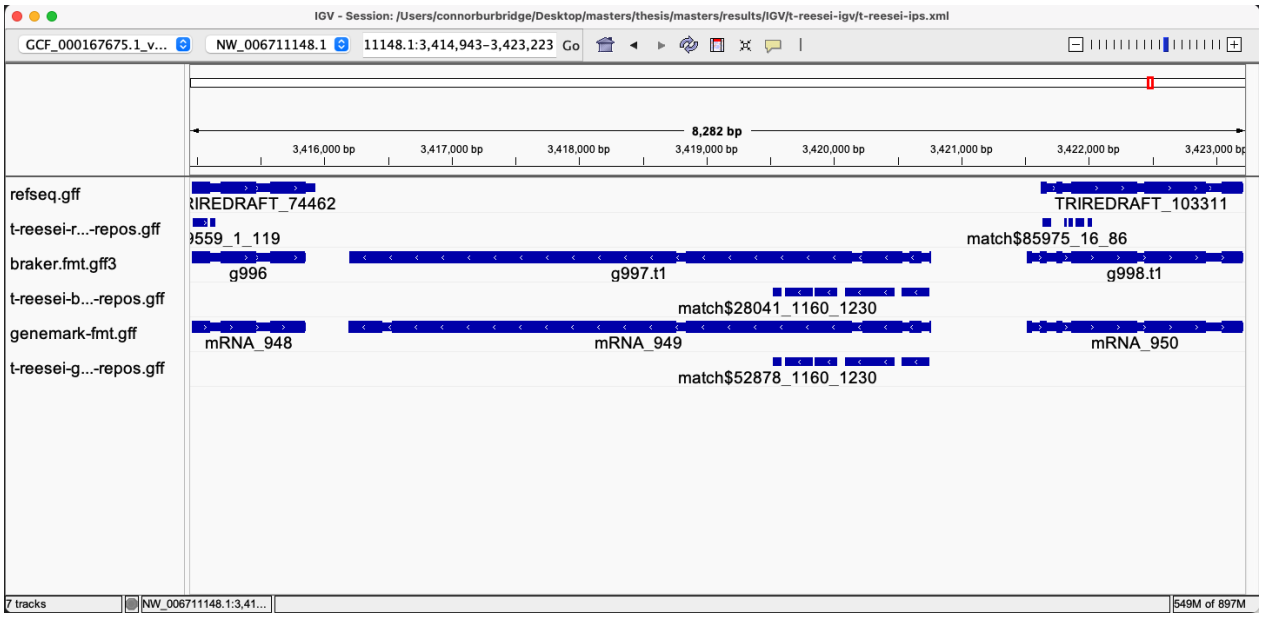


Figure 5.14: An IGV capture showing a scenario where GeneMark and Braker2 agree on a gene model with supporting Pfam evidence and RefSeq does not report any gene.

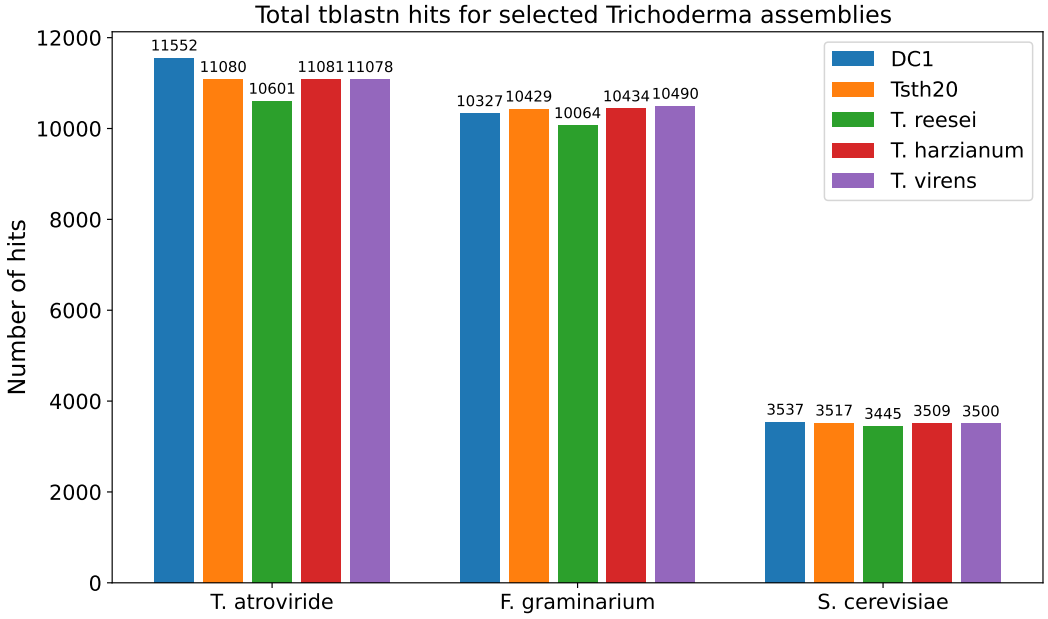
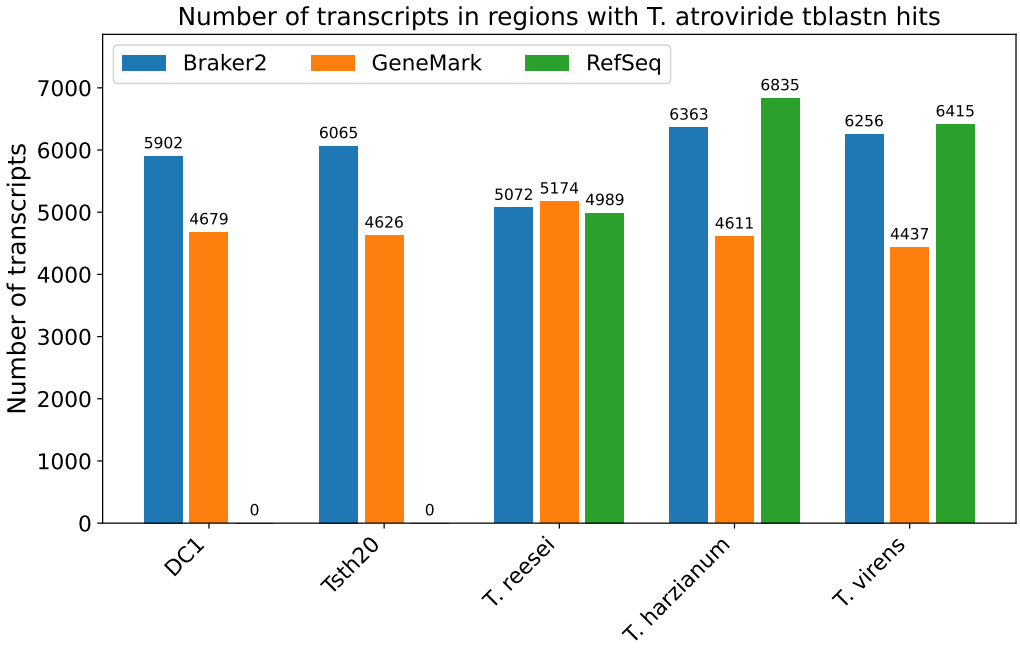
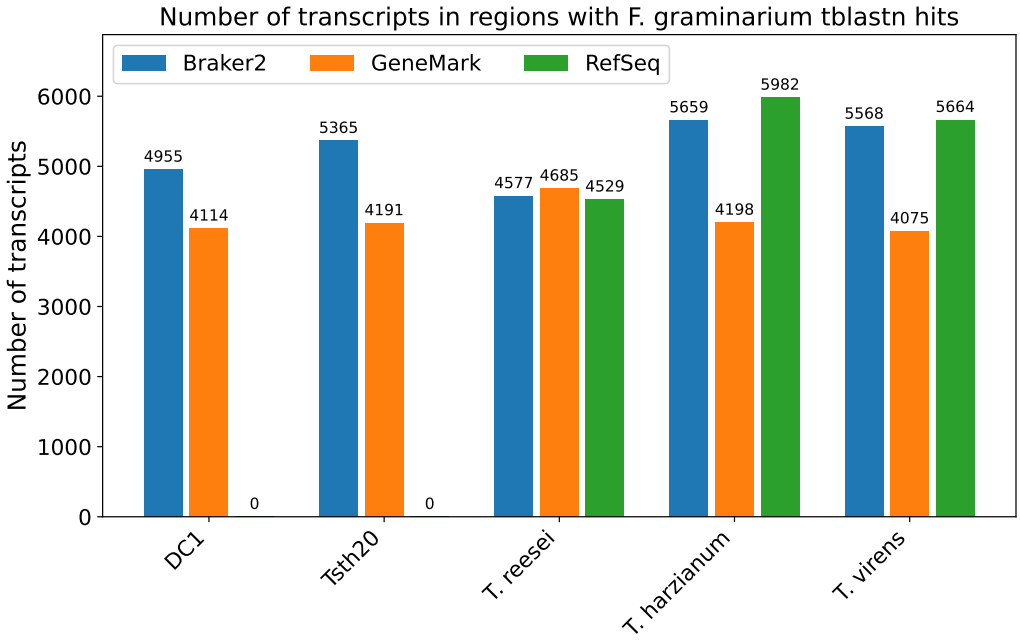


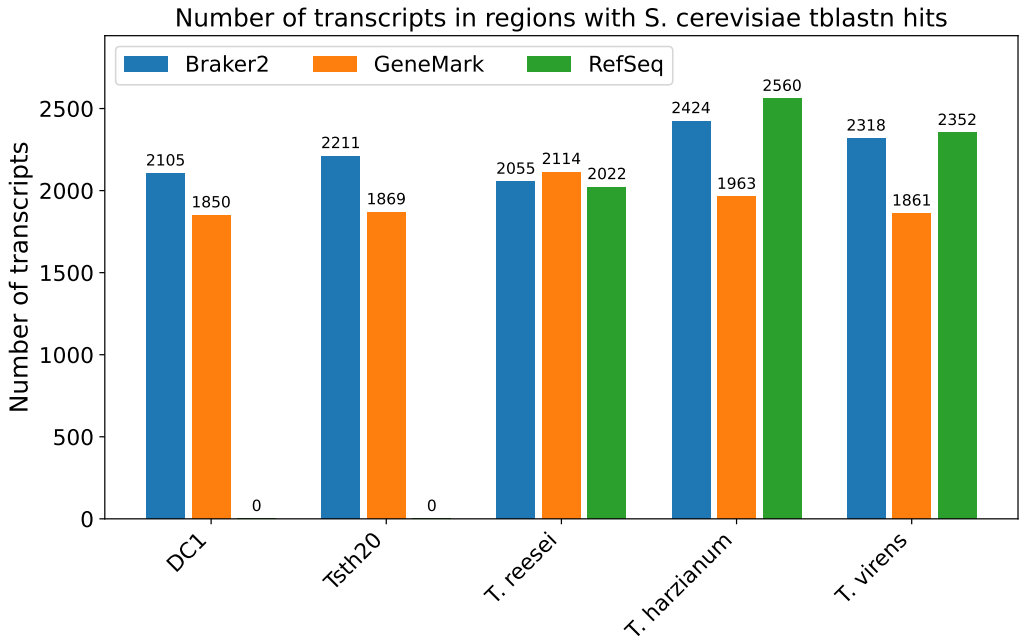
Figure 5.15: Total tblastn hits for reference proteins from *T. atroviride*, *F. graminarium*, and *S. cerevisiae* against *Trichoderma* assemblies.



(a) *T. atroviride* tblastn hits



(b) *F. graminarium* tblastn hits



(c) *S. cerevisiae* tblastn hits

Figure 5.16: tblastn hits for reference proteins from *T. atroviride*, *F. graminarium*, and *S. cerevisiae* against *Trichoderma* assemblies. For DC1 and Tsth20, RefSeq annotations are not available, so their values are set to 0.

Subject	Query	Braker2	GeneMark	RefSeq
DC1	*T. atroviride*	5902	4679	N/A
DC1	*F. graminarium*	4955	4114	N/A
DC1	*S. cerevisiae*	2105	1850	N/A
Tsth20	*T. atroviride*	6065	4626	N/A
Tsth20	*F. graminarium*	5365	4191	N/A
Tsth20	*S. cerevisiae*	2211	1869	N/A
T. reesei	*T. atroviride*	5072	5174	4989
T. reesei	*F. graminarium*	4577	4685	4529
T. reesei	*S. cerevisiae*	2055	2114	2022
T. harzianum	*T. atroviride*	6363	4611	6835
T. harzianum	*F. graminarium*	5659	4198	5982
T. harzianum	*S. cerevisiae*	2424	1963	2560
T. virens	*T. atroviride*	6256	4437	6415
T. virens	*F. graminarium*	5568	4075	5664
T. virens	*S. cerevisiae*	2318	1861	2352

Table 5.9: Counts of transcripts predicted by Braker2, GeneMark and RefSeq in regions containing tblastn hits from *Ttrichoderma atroviride*, *Fusarium granminarium*, and *Saccharomyces cerevisiae*.

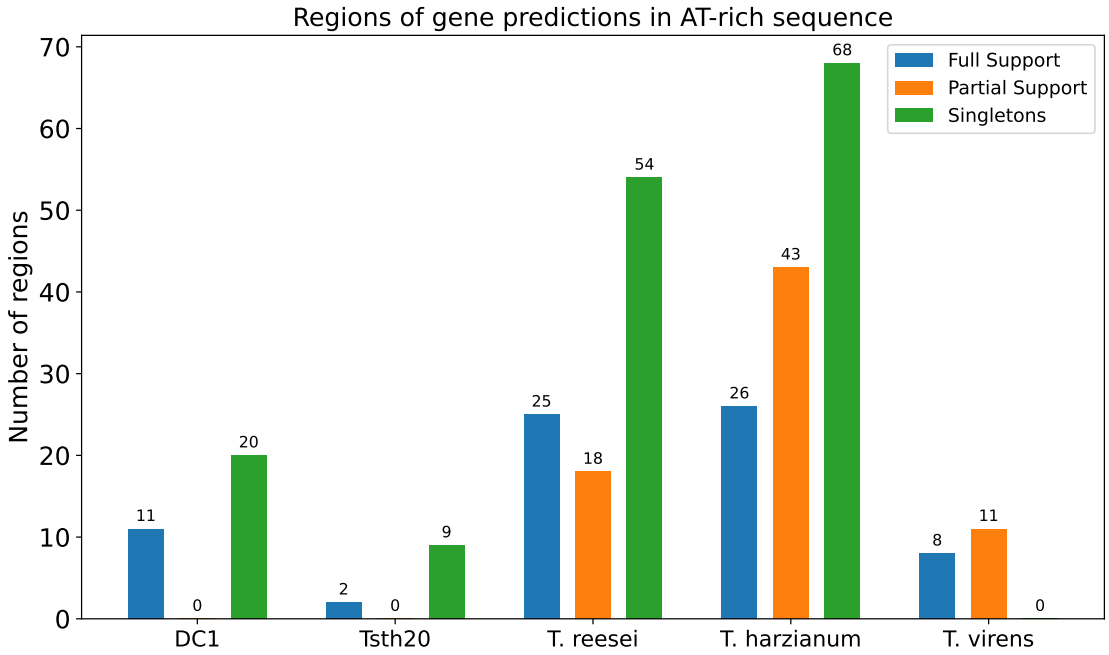


Figure 5.17: Regions of full, partial, and no agreement in AT-rich genomic sequence. It is important to note that in the cases of DC1 and Tsth20, full support indicates supporting gene predictions from both GeneMark and Braker2 and as such, there are no regions with partial support.

GC composition less than 28% as determined in Section 5.2. In comparison to results from Section 5.7, we see that overall there are far fewer regions with gene predictions in AT-rich genomic segments. In DC1, Tsth20, there are very few regions with full support from both Braker2 and GeneMark, but more singletons. Again, as in Section 5.7, there are no regions with partial support as only two gene finding tools were applied to those assemblies. *T. reesei* and *T. harzianum* report more regions in AT-rich genomic sequence than the other assemblies, but with the majority of regions belonging to the singleton category. *T. virens* is an interesting case, reporting a similar numbers of regions and genes in AT-rich genomic sequence as DC1 and Tsth20. *T. virens* is also the only assembly to report zero singleton gene predictions. Why *T. virens* differs from the other RefSeq assemblies is unclear. In general, there are few regions with gene predictions in AT-rich genomic segments, and within these regions, the majority of cases are isolated singleton gene predictions. These observations differ greatly from those made in nucleotide composition agnostic approach in Section 5.7.

In addition to a breakdown of regions in AT-rich genomic sequence, understanding which gene finders predict more or fewer genes in these regions may be of interest. It is possible that an HMM, trained on genomic sequence with varying nucleotide content, may predict genes differently to an HMM trained only on sequences with uniformly distributed nucleotide composition as the variation in nucleotide composition may affect the various states and relationships between them in an HMM. Figure 5.18 and Table 5.11 shows the number of genes predicted by each gene finding tool in regions of AT-rich genomic sequence. GeneMark

Assembly	Full Support	Partial Support	Singletons	No. Genes
DC1	11	N/A	20	42
Tsth20	2	N/A	9	13
T. reesei	25	18	54	194
T. harzianum	26	43	68	265
T. virens	8	11	0	49

Table 5.10: Regions of full and partial agreement as well as singleton regions 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.

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

Table 5.11: Number of genes predicted by Braker2, GeneMark and RefSeq in AT-rich genomic sequence from each assembly.

appears to predict the fewest genes in AT-rich regions, while RefSeq appears to predict the most. Braker2 lies somewhere in the middle. Again, *T. virens* appears as an odd case, with very few predictions from all gene finders. Why *T. virens* differs from the other RefSeq assemblies is unclear.

Finally, to test the probability of any given gene prediction falling in an AT-rich genomic sequence, a two-sided binomial test was performed to determine if the number of genes predicted in AT-rich sequences is proportional to fraction of genomic sequence they comprise. The null hypothesis in this case is that the number of genes in AT-rich sequence is proportional to the total AT-rich sequence in the genome. The results of the test are shown in Table 5.12. In all cases, it appears that the gene finding tools selected for this analysis do not predict the same proportion of genes in AT-rich genomic sequence as in typical genomic sequence.

In summary, very few regions with gene predictions are present in AT-rich genomic sequence. Additionally, genes predicted in these regions tend to be isolated and not supported by other gene finders, although some agreement is observed. In terms of number of genes predicted, RefSeq tends to predict the most genes in these AT-rich regions while GeneMark predicts the fewest. Lastly, the selected gene finding tools do not predict genes in AT-rich sequences in proportion to the fraction of genomic sequence they comprise.

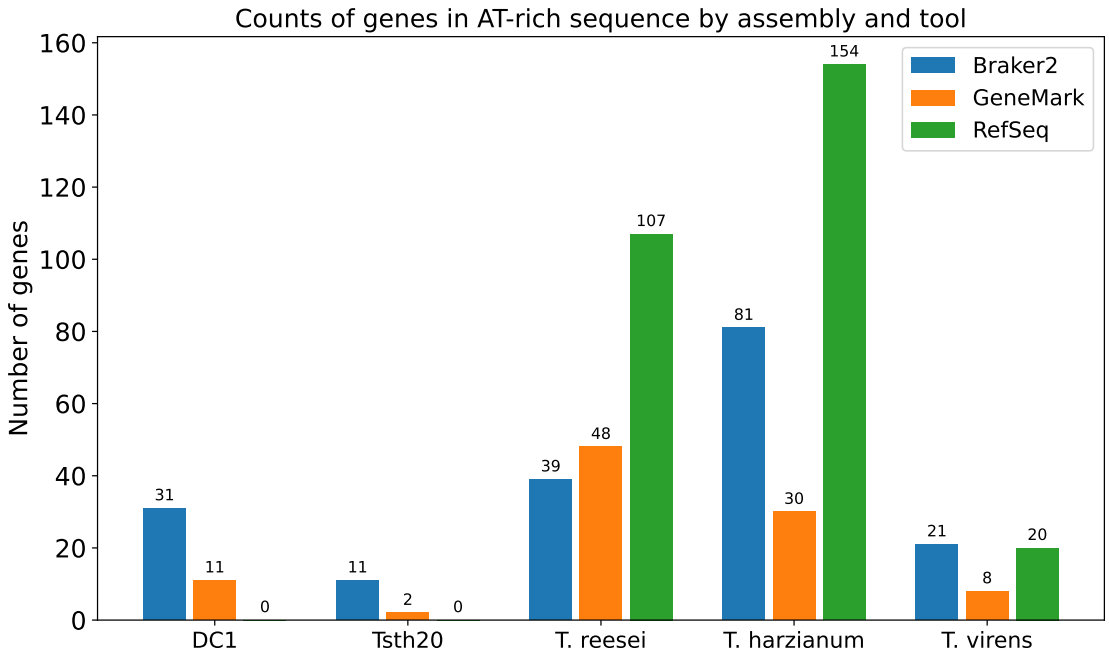


Figure 5.18: Number of genes predicted by each gene finder in AT-rich genomic sequence.

Tool	DC1	Tsth20	*T. reesei*	*T. harzianum*	*T. virens*
Braker2	$9.56 * 10^{-181}$	$1.14 * 10^{-259}$	$2.68 * 10^{-96}$	$4.05 * 10^{-140}$	$1.35 * 10^{-35}$
GeneMark	$5.12 * 10^{-216}$	0.0	$5.66 * 10^{-49}$	$5.37 * 10^{-219}$	$5.31 * 10^{-35}$
RefSeq	N/A	N/A	$1.29 * 10^{-49}$	$2.44 * 10^{-205}$	$7.40 * 10^{-33}$

Table 5.12: p-values produced from a two-sided binomial test for each combination of tool and assembly.

6 Conclusions

6.1 Selection of Gene Finding Tool

With all of these results, it makes sense to explore the question of which gene finding tool one should choose for optimal gene prediction performance. Comparisons drawn in this section are made in the context of *Trichoderma* assemblies and may not extend to other datasets. Observations from the results portion of this work have been converted to scores relative to the other gene finders. Scores range from zero to three, with values being: 0 - failing performance, 1 - passing performance, 2 - good performance, and 3 - excellent performance. Results from this work are summarized in Table 6.1. Ignoring availability and use of the gene finding tools, it would appear that RefSeq performs the best in the remaining categories, earning top marks in every category except in it's ability to predict very short genes. Braker2 earns second place; however, this does not capture Braker2's failure in predicting accurate numbers of genes in DC1, Tsth20, *T. harzianum* and *T. virens*. GeneMark comes in last, excelling only in number of genes predicted and Pfam support for the genes that it predicts.

Relating these observations to use-case scenarios, in the case that your organism of interest has a RefSeq annotation associated with it, the RefSeq gene prediction process appears to produce the best set of predictions. If users also have experimental evidence, such as RNAseq data under experimental conditions, it may be worthwhile training a Braker2 model and predicting genes with Braker2 to supplement the already well performing RefSeq gene predictions. In a similar case, if the organism is of RefSeq status, but the assembly in question is novel to the work being performed, training and prediction using a Braker2 model with experimental evidence either from the novel genome or from the RefSeq accession in addition to use of the RefSeq annotation would likely be the best option. If the organism of interest is not a RefSeq individual but training data is available for that organism, Braker2 is the next best option, although it is important to note that the application of a trained Braker2 prediction model to an organism from which the training data did not originate is not advised based on the results presented in this work (see 5.4). While it is true that the *T. reesei* genome differs from other *Trichoderma* genomes, it's status as a representative RefSeq organism makes it somewhat of a gold standard. In this case, applying a gene model trained using evidence from the gold standard produces biased numbers of genes predicted in other *Trichoderma* genomes. While Braker2 technically scores the second highest, users must be very careful when selecting training data, and ensure that the training data either comes from the organism of interest, or comes from a very closely related organism with a highly similar genome. In the case that no appropriate training data is available, GeneMark

is still an option, and users can be confident that the tool predicts a reasonably accurate number of genes with supporting Pfam matches. It is also important to note that GeneMark does not perform as well in AT-rich regions as Braker2 and RefSeq, does not predict isoforms, and systematically fails to predict some BUSCO orthologs.

When availability of a gene finding tool becomes a concern and RefSeq is not considered, the scores drop significantly for Braker2 and GeneMark as seen in the final row of Table 6.1. In this situation, Braker2 still outperforms GeneMark. If the organism of interest is not considered a representative RefSeq individual, but supporting evidence specific to that organism or a very closely related organism is available, a trained Braker2 prediction model will perform well. Again, in the case that the organism is not a RefSeq individual and no appropriate training data is available, GeneMark is still a reasonable option even with the previously identified caveats.

In summary, these results indicate that if your organism is a RefSeq organism, use the RefSeq annotation. If no RefSeq predictions are available but appropriate training data is, one should use Braker2. If the training data is of questionable similarity or not available at all, users can fall back on *ab initio* gene finders such as GeneMark, which while not ideal, still predict genes with supporting evidence.

6.2 Future Work

While this work is extensive, there are still many areas that could be expanded on. The sheer number of gene finding tools available means that many more could be compared to the ones presented here. In addition, supplying more training data to Braker2 may improve gene finding performance, and training on RNAseq from different organisms could be used to improve the performance of Braker2. It is clear from the results that Braker2 performed worse than GeneMark and the RefSeq predictions, which was likely due to the training data being used from *T. reesei*. Different types of training data such as RNAseq, expressed sequence tags (ESTs), and protein sequences could also be used to train models that support external evidence. Another limitation of this work is that the gene finders were run with default parameters, and it is possible that tuning the parameters of the gene finders could improve performance.

Following the gene prediction process, there are a number of different analyses that could be performed on the predicted genes. For example, functional annotation of the predicted genes could be performed using tools such as BLAST2GO to assign Gene Ontology (GO) terms and KEGG pathways to the predicted genes. This could provide further insight into the functional roles of the predicted genes and their potential applications. More importantly in the context of this work, analysis of predicted genes with a tool like antiSMASH could be used to identify biosynthetic gene clusters (BGCs) in the predicted genes.

Expanding the number of genomes used in this work would also be beneficial, as the results presented here are based on a limited number of genomes. Including more genomes from different species and strains of *Trichoderma* could provide a more comprehensive understanding of gene finding performance across the genus.

Category	Braker2	GeneMark	RefSeq
Availability	3	3	0
Ease of install	1	2	0
Ease of use	3	3	0
# of genes predicted	0	3	3
# of transcripts predicted	3	0	2
Predicts shortest genes	2	1	0
Predicts more shorter genes	1	0	3
BUSCO Performance	2	1	3
Performance in AT-rich sequence	2	1	3
Predictions with InterProScan support	3	3	3
Final Score (Publicly Available)	20	17	N/A
Final Score (Ignoring Availability)	13	9	17

Table 6.1: Table with scores attributed to performance of each gene finder in several categories. The score definitions for performance are as follows: 0 - fail, 1 - pass, 2 - good, 3 - excellent. Since RefSeq is not publicly available, it is marked as N/A in the publicly available final scores. The dashed line separates categories associated with availability and use from categories describing gene prediction performance.

Additionally, comparing the performance of gene finders on other fungal genomes could provide insights into the generalizability of the findings.Examining the performance of gene finders on different assemblies of the same genome could also be useful, as different assemblies may have different levels of completeness and accuracy. This could help to identify tangible benefits of using one assembly over another, and could also provide insights into the limitations of gene finders when applied to different assemblies.

Finally, the results of this work could be used to inform future research in the field of fungal genomics, particularly in the case of the DC1 and Tsth20 genome assemblies. The results of these genome assemblies are high-quality, and contain near-chromosomal scale sequences. The predicted genes from these genomes are a rich resource for future research, and could be used to identify novel genes and gene families in DC1 and Tsth20. Further understanding of the mechanisms behind salt and drought tolerance as well as degradation of hydrocarbon in sub-optimal environments could be achieved by studying the predicted genes in these genomes.

References

[1] N Saitou and M Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, July 1987.

[2] Joseph Felsenstein. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. *Evolution; International Journal of Organic Evolution*, 39(4):783–791, July 1985.

[3] Sheridan L. Woo, Rosa Hermosa, Matteo Lorito, and Enrique Monte. Trichoderma: A multipurpose, plant-beneficial microorganism for eco-sustainable agriculture. *Nature Reviews Microbiology*, 21(5):312–326, 2023.

[4] Christian P. Kubicek, Andrei S. Steindorff, Komal Chenthamara, Gelsomina Manganiello, Bernard Henrissat, Jian Zhang, Feng Cai, Alexey G. Kopchinskiy, Eva M. Kubicek, Alan Kuo, Riccardo Baroncelli, Sabrina Sarrocco, Eliane Ferreira Noronha, Giovanni Vannacci, Qirong Shen, Igor V. Grigoriev, and Irina S. Druzhinina. Evolution and comparative genomics of the most common Trichoderma species. *BMC Genomics*, 20(1):485, 2019.

[5] Prasun K. Mukherjee, Benjamin A. Horwitz, and Charles M. Kenerley. Secondary metabolism in Trichoderma – a genomic perspective. *Microbiology*, 158(1):35–45, 2012.

[6] Carla Gonçalves, Chris Todd Hittinger, and Antonis Rokas. Horizontal Gene Transfer in Fungi and Its Ecological Importance. In Yen-Ping Hsueh and Meredith Blackwell, editors, *Fungal Associations*, pages 59–81. Springer International Publishing, Cham, 2024.

[7] Miroslav Plohl, Nevenka Meštrović, and Brankica Mravinac. Centromere identity from the DNA point of view. *Chromosoma*, 123(4):313–325, August 2014.

[8] David J. Winter, Austen R. D. Ganley, Carolyn A. Young, Ivan Liachko, Christopher L. Schardl, Pierre-Yves Dupont, Daniel Berry, Arvina Ram, Barry Scott, and Murray P. Cox. Repeat elements organise 3D genome structure and mediate transcription in the filamentous fungus Epichloë festucae. *PLOS Genetics*, 14(10):1–29, 2018.

[9] Christian P. Kubicek, Alfredo Herrera-Estrella, Verena Seidl-Seiboth, Diego A. Martinez, Irina S. Druzhinina, Michael Thon, Susanne Zeilinger, Sergio Casas-Flores, Benjamin A. Horwitz, Prasun K. Mukherjee, Mala Mukherjee, László Kredics, Luis D. Alcaraz, Andrea Aerts, Zsuzsanna Antal, Lea Atanasova, Mayte G. Cervantes-Badillo, Jean Challacombe, Olga Chertkov, Kevin McCluskey, Fanny Coulpier, Nandan Deshpande, Hans von Döhren, Daniel J. Ebbole, Edgardo U. Esquivel-Naranjo, Erzsébet Fekete, Michel Flipphi, Fabian Glaser, Elida Y. Gómez-Rodríguez, Sabine Gruber, Cliff Han, Bernard Henrissat, Rosa Hermosa, Miguel Hernández-Oñate, Levente Karaffa, Idit Kosti, Stéphane Le Crom, Erika Lindquist, Susan Lucas, Mette Lübeck, Peter S. Lübeck, Antoine Margeot, Benjamin Metz, Monica Misra, Helena Nevalainen, Markus Omann, Nicolle Packer, Giancarlo Perrone, Edith E. Uresti-Rivera, Asaf Salamov, Monika Schmoll, Bernhard Seiboth, Harris Shapiro, Serenella Sukno, Juan Antonio Tamayo-Ramos, Doris Tisch, Aric Wiest, Heather H. Wilkinson, Michael Zhang, Pedro M. Coutinho, Charles M. Kenerley, Enrique Monte, Scott E. Baker, and Igor V. Grigoriev. Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of Trichoderma. *Genome Biology*, 12(4):R40, April 2011.

[10] David A. Fitzpatrick. Horizontal gene transfer in fungi. *FEMS Microbiology Letters*, 329(1):1–8, 2012.

[11] Biancamaria Senizza, Fabrizio Araniti, Simon Lewin, Sonja Wende, Steffen Kolb, and Luigi Lucini. Trichoderma spp.-mediated mitigation of heat, drought, and their combination on the Arabidopsis thaliana holobiont: A metabolomics and metabarcoding approach. *Frontiers in Plant Science*, 14, 2023.

[12] Xiao-Ya An, Guo-Hui Cheng, Han-Xing Gao, Xue-Fei Li, Yang Yang, Dan Li, and Yu Li. Phylogenetic Analysis of Trichoderma Species Associated with Green Mold Disease on Mushrooms and Two New Pathogens on Ganoderma sichuanense. *Journal of Fungi*, 8(7):704, July 2022.

[13] Brendan Loftus. 4 - Genome Sequencing, Assembly and Gene Prediction in Fungi. In Dilip K. Arora and George G. Khachatourians, editors, *Fungal Genomics*, volume 3, pages 65–81. Elsevier, 2003.

[14] Girum Fitihamlak Ejigu and Jaehee Jung. Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. *Biology*, 9(9), September 2020.

[15] Mark Borodovsky and Alex Lomsadze. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Current Protocols in Bioinformatics*, Chapter 4(SUPPL.35):4.6.1–4.6.10, September 2011.

[16] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268(1):78–94, April 1997.

[17] W. H. Majoros, M. Pertea, and S. L. Salzberg. TigrScan and GlimmerHMM: Two open source *ab initio* eukaryotic gene-finders. *Bioinformatics*, 20(16):2878–2879, November 2004.

[18] Mario Stanke and Stephan Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics (Oxford, England)*, 19 Suppl 2:ii215–225, October 2003.

[19] Tomáš Brůna, Katharina J. Hoff, Alexandre Lomsadze, Mario Stanke, and Mark Borodovsky. BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR genomics and bioinformatics*, 3(1):lqaa108, March 2021.

[20] Brandi L. Cantarel, Ian Korf, Sofia M.C. Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18(1):188–196, January 2008.

[21] The NCBI Eukaryotic Genome Annotation Pipeline. https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/process/.

[22] Mosè Manni, Matthew R. Berkeley, Mathieu Seppey, and Evgeny M. Zdobnov. BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, 1(12):e323, 2021.

[23] Arryn Craney, Salman Ahmed, and Justin Nodwell. Towards a new science of secondary metabolism. *The Journal of Antibiotics*, 66(7):387–400, 2013.

[24] Hisayuki Komaki and Tomohiko Tamura. Polyketide Synthase and Nonribosomal Peptide Synthetase Gene Clusters in Type Strains of the Genus Phytohabitans. *Life*, 10(11):257, October 2020.

[25] Kai Blin, Simon Shaw, Hannah E Augustijn, Zachary L Reitz, Friederike Biermann, Mohammad Alanjary, Artem Fetter, Barbara R Terlouw, William W Metcalf, Eric J N Helfrich, Gilles P van Wezel, Marnix H Medema, and Tilmann Weber. antiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research*, 51(W1):W46–W50, July 2023.

[26] Byoungnam Min, Igor V Grigoriev, and In-Geol Choi. FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation. *Bioinformatics*, 33(18):2936–2937, September 2017.

[27] Simon Bennett. Solexa Ltd. *Pharmacogenomics*, 5(4):433–438, June 2004.

[28] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

[29] Yunhao Wang, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, 39(11):1348–1365, 2021.

[30] Jiang Hu, Zhuo Wang, Zongyi Sun, Benxia Hu, Adeola Oluwakemi Ayoola, Fan Liang, Jingjing Li, José R. Sandoval, David N. Cooper, Kai Ye, Jue Ruan, Chuan-Le Xiao, Depeng Wang, Dong-Dong Wu, and Sheng Wang. NextDenovo: An efficient error correction and accurate assembly tool for noisy long reads. *Genome Biology*, 25(1):107, 2024.

[31] Jiang Hu, Junpeng Fan, Zongyi Sun, and Shanlin Liu. NextPolish: A fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*, 36(7):2253–2255, April 2020.

[32] Simon Andrews. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[33] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, 29(8):1072–1075, April 2013.

[34] P. Rice, I. Longden, and A. Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, 16(6):276–277, June 2000.

[35] Vijai Kumar Gupta, Andrei Stecca Steindorff, Renato Graciano de Paula, Rafael Silva-Rocha, Astrid R. Mach-Aigner, Robert L. Mach, and Roberto N. Silva. The Post-genomic Era of Trichoderma reesei: What's Next? *Trends in Biotechnology*, 34(12):970–982, December 2016.

[36] Chengzhi Liang, Long Mao, Doreen Ware, and Lincoln Stein. Evidence-based gene predictions in plant genomes. *Genome Research*, 19(10):1912–1923, October 2009.

[37] SRA-Toolkit. https://hpc.nih.gov/apps/sratoolkit.html.

[38] Kolmogorov–Smirnov Test. In *The Concise Encyclopedia of Statistics*, pages 283–287. Springer New York, 2008.

[39] Mosè Manni, Matthew R. Berkeley, Mathieu Seppey, and Evgeny M. Zdobnov. BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, 1(12):e323, 2021.

[40] Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, October 2015.

[41] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 2014.

[42] E. Michael Gertz, Yi-Kuo Yu, Richa Agarwala, Alejandro A. Schäffer, and Stephen F. Altschul. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biology*, 4(1):41, December 2006.

[43] Rina Dechter. Chapter 12 - Temporal Constraint Networks. In Rina Dechter, editor, *Constraint Processing*, pages 333–362. Morgan Kaufmann, 2003.

[44] Python Software Foundation. Python.

[45] Brad Chapman. BCBio Python Package.

[46] James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26, 2011.

[47] Mario Stanke, Ana Tzvetkova, and Burkhard Morgenstern. AUGUSTUS at EGASP: Using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology*, 7(1):S11, 2006.

[48] Derek W. Barnett, Erik K. Garrison, Aaron R. Quinlan, Michael P. Strömberg, and Gabor T. Marth. BamTools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics (Oxford, England)*, 27(12):1691–1692, June 2011.

[49] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[50] Mihaela Pertea, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3):290–295, 2015.

[51] Geo Pertea and Mihaela Pertea. GFF Utilities: GffRead and GffCompare. *F1000Research*, 9, 2020.