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[1]

[1]USask NSID: cbe453 , USask ID no. 11162928 , Supervisors: Dave Schneider & Tony Kusalik,

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 acknoweldge 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

1 Introduction

The study of organisms in Biology is a highly complex complex process involving many disciplines. To better understand how these organisms function, we must braker down the problem into different sub-problems. One important sub-problem in biology is the understanding of the molecular tools and processes used by cells to function in normal and abnormal environmental scenarios or stress conditions. These conditions may include disease and environmental stress for example. However, previous work in biology has shown that the underlying backbone of information, known as the genome, can vary widely between different organisms in many aspects, such as overall structure, length, ploidy, methylation, and overall gene content. All of these aspects can affect the survival of an organism in any given environment, some of which may be interesting to researchers. As an example, one of the most popular and extensively studied diseases is cancer. Through the study of human genomes, researchers identified a key gene, named TP53, involved in the suppression of tumours. Functional mutations affecting this gene can result in increased risk of cancer. Understanding how and why TP53 suppresses tumours can provide insight into future cancer prevention methods and treatments. These genes can then be mapped to the genome of the organism to identify its location.

This general workflow can be applied to features of interest from organisms in all branches of life. To facilitate this process of identifying genes from a genome, the process of genome annotation was developed. In this case, instead of identifying a gene of interest and then mapping it back to the genome, gene finders identify potential genes from a reference genome before truly knowing their function or if the candidate is truly untilized by the organisms. This set of potential genes acts as a reference for future research. However, to generate a reliable set of possibe genes, a gene finder must be supplied with a suitable high quality genome. In many cases, researchers may be studying a specific strain or variety of organism that differs from the reference assembly and annotation available for the organism of interest. Rather than using the reference assembly for analysis, it may be beneficial to generate a new assembly for the unique variety or strain, which must then be passed through a gene finding tool. With the variety of gene finding tools and approaches, the choice of an appropriate tool can affect the resulting gene set. This problem raises a question. Do the results from gene finding tools differ? And more importantly, how should one compare results from gene finding tools when there is no reference annotation for a specifc variety in question? Most genome annotation tools benchmark their performance in comparison to an existing reference annotation, which is usually considered to be of suitably high quality. This is not possible in the case of unique assemblies, and so the devleopment of a comparative methodology is in order.

2 Background

2.1 Genomics

Genomics is a wide area of study focusing on the genomes of organisms from all varieties of life. A genome is the fundamental set of 'rules' used to create what we know as life. One can think of a genome as a set of instructions that our cells use in order to complete the tasks that make us function. A genome is comprised of tightly bundled sequences of DNA, which are stored in the nucleus of cells. These bundles of DNA contain sections known as genes, which can be thought of as the tools described by the set of instructions. These tools carry out a vast number of processes ranging from no known function at all to genes that are key in protecting against diseases cancer. Genomes can vary widely in size, ranging from small bacterial genomes of roughly 4 Mb up to approximately 149000 Mb. Piecing together genomes provides numerous opportunities to understand other 'omics' within cells, such as proteomics, metabolomics, transcriptomics and epigenomics.

2.1.1 Sequencing

Sequencing is a pivotal form of data used in nearly all applications of Bioinformatics. To understand the processes used by organisms for day to day survival or in unique circumstances, we must have an initial set of data points to work with. These sequences, referred to as reads after sequencing, are the foundation for problems ranging from taxonomical classification to the understanding or complex biological functions like signaling pathways.

2.2 Whole Genome Shotgun Sequencing

Whole Genome Shotgun sequencing (WGS), is a method to produce a large number of genomic sequences from a sample of interest. This is form of sequencing is quite common as it is has a wide variety of applications in research. WGS involves slicing up genomic DNA into smaller segments. These small segments are then processed further resulting in a set of physical molecules that can be supplied to a compatible sequencing platform of which there are a variety. Modern equencing platforms are typically broken up into next generation sequencing (NGS) and 3rd generation sequencing.

2.3 Next Generation Sequencing - Illumina

Illumina sequencing is one of the most popular NGS platforms currently available. Illumina sequencing produces a very large number of high quality short reads, typically between 75 and 250 base pairs in length. Sequencing libraries can be prepared to produce reads solely from one end of a sequence fragment (single-end) or both ends (paired-end). Advantages of paired end sequences are the additional context provided by the paired sequence on the opposite end of the fragment. This context is leveraged by read processing tools to identify features such as repetitive regions and genomic rearrangments, which can be significant in downstream analyses. Illumina sequence librairies are generated by first fragmenting the DNA samples, amplifying them via PCR, ligating adapters that allow the sequence to bind to the sequencing plate, and finally identifying each fragment's sequence of nucleotides using fluorescently-labeled nucleotides that bind to the fragments.

2.4 3rd Generation Sequencing - Nanopore

2.5 Trichoderma

Crop resistance to environmental stressors is a necessity for crop health and overall crop yields. Current popular methods for crop protection involve the use of pesticides and genetically modified organisms, which can be expensive and potentially politically dividing in the case of GMOs[7]. In addition, crops suffer when soils are not sufficient for crop growth and health. Soil insufficiencies can result in drought stress as well as nutrient stress, leading to poor overall yields.

Trichoderma is a fungi that can both communicate with and colonize the roots of plants in a non-toxic, non-lethal, opportunistic symbiotic relationship[10]. Many strains of *Trichoderma* have been shown to provide resistance to pathogenic bacteria and other fungi in soils through the use of polyketides, non-ribosomal peptide synthetases and other antibiotic products[10]. 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 mentioned before. In addition to these beneficial properties, the two strains mentioned previously provide even further protection for plants in dry, salty soils and one strain also has potential for use as a bioremediation tool in soils contaminated with hydrocarbon content. Bioremediation and resistance to drought tolerance has also been investigated in other strains of *Trichoderma* as well[5]. However, little is known about the mechanisms at work in these strains, so DC1 and Tsth20 were sequenced by the Global Institute for Food Security (no publication yet) in an initial attempt to better understand the details of these genomes. While this research does not directly identify genomic elements related to the secretome of these genomes, it may serve as a foundation for future research of *Trichoderma*.

2.6 Genome Assembly

Sequence assembly has been a long-standing problem in the field of bioinformatics[4]. Determining the correct order and combination of smaller subsequences into an accurate complete sequence assembly is computationally difficult in terms of compute resources such as memory, CPU cycles and storage required for input sequences[4]. In addition to these difficulties, there can be other issues encountered during asssembly due to the nature of the data or genomes themselves, such as low quality base calls for long read data, which is not necessarily the case today, or the inherent content of genomes themselves using repetitive regions as an example. Insufficient data used in an assembly may result in short, fragmented assemblies, depending on the size of the genomes, while sequence data that is not long enough can fail to fully capture repetitive regions in an assembly. To solve this problem, a wide range of assembly tools have been developed with their own unique approaches to the genome assembly problem, so it is important to use an appropriate assembler for the task at hand, and also important to evaluate the assembly thoroughly.

Genome assembly tools generally approach the assembly problem using a graph-based approach. The most common graph-based approach is the de Bruijn graph assembly. A graph in this context, is set of nodes (k-mers from sequences) connected by edges (overlaps between k-mers). Traversing through this graph results in longer subsequences that ultimately result in a set of consensus sequences and final assembly. In the early years of long read sequence data, sequencing platforms encountered difficulties producing consistently high scores for base calls when seuqncing. To combat this, some assembly workflows may also include a polishing or correction step once the initial assembly is completed in which high quality short read sequences are supplied as supplemental information to correct low quality base calls in the assembly. These low quality base calls are typically not present in modern long read sequencing approaches as the methodology and quality of calls have improved drastically. While the polishing step is arguably unnecessary in modern assemblies, the polishing programs remain available should researchers be interested in applying additional reads for polishing.

One approach to aid in the previously mentioned issue of assembly correctness is to use a combination of long and short reads in what is known as a hybrid assembly. Combining both highly accurate short reads with deep coverage along with less accurate but much longer reads can produce high quality genome assemblies that capture long repetitive regions. Hybrid assembly approaches have been shown to produce high quality assemblies in a wide variety of organisms as the combine long read data with short data to produce assemblies that properly represent long repetetive regions with additionaly high quality Illumina sequences for correction. Once assembled, the sequences must also be evaluated with measures such as N50, L50, coverage, average contig length and total assembled length to ensure that the genomes are well assembled, at least based on these metrics[4]. Following appropriate assembly protocols is essential to the further success of a project as downstream processing such as annotation depends on a high-quality assembly.

2.7 Identification of AT-rich Genomic regions

One important aspect of interest when assembling any form of sequence is GC content or percent GC of the assembled sequence. Large regions of anomolous GC content may be of interest to researchers as they may contain repetitive regions and unique features responsible for traits specific to the organism in question.

2.8 Repeat Identification and Masking

Repeat identification within assembled genomes is a problem that needs to be considered during the genome annotation process. Regions with long repeats can have a significant impact on genome assembly as well as gene finding due to the limitation of short reads used in some assemblies[8]. Short reads may be unable to bridge or cover entire repeat regions within a genome, so it is important to consider the use of long reads from technologies such as Nanopore or PacBio to provide a complete picture of these regions when pursuing a new genome assembly project. It is also possible for repetitive regions to contain genes as well, making for an interesting investigation in regards to *Trichoderma*, as fungal genomes have been shown to contain many repeat regions with a high concentration of A and T nucleotides[9]. Once these repetitive regions have been identified, the genome could be masked to exlude these regions in downstream processing if desired, as these regions may be poorly assembled and may result in found genes that do not truly exist in those regions. However, this may not be as common today, as repetetive regions have been shown to contain genes as well[6]. This may affect the gene finding process described later and may be an interesting topic to look into considering the large number of available gene finding programs.

2.9 Gene Finding Methods

Gene finding (or gene annotation) has been a long standing computational problem in bioinformatics, which concerns itself with identifying potential genes within assemblies based on patterns or pre-existing experimental evidence evidence considered by the gene finding program. This process is critical for unraveling and understanding the complex processes occurring in all forms of life with applications in medical science, agriculture, biomanufacturing, environmental studies and many others. In a general sense, gene finding programs operate by searching for patters or indicators showing that a gene of feature may be present. The most basic indicators being start and stop codons, with introns and exons in between should the sequence match the applied model. The results produced by gene finding tools can vary considerably for a number of reasons, including quality of the assembly, the intrinsic model used by the gene finder, filtering criteria, and even the nature of the organism and assembly itself. Given the broad applications, choice of gene finding tools, and the variability of assemblies being considered, it is important that we gain a deeper understanding of these tools prior to putting them to use.

There are two common methods for gene finding, those methods being *ab initio* methods, where programs search for patterns and gene structures, and similarity or evidence-based searches, which use prior information such as RNAseq data, expressed sequence tags and expressd protein sequences to identify genes within a new genome[2]. Complicating the process more is the introduction of introns and alternative splicing in eukaryotes, making it possible for one gene to have several possible transcripts at the same locus. An example of an *ab initio* method would be GeneMark-ES[3], while an evidence based tool would be Braker2[1]. *Ab initio* gene finders typically predict genes using a Hidden Markov Model (HMM)[2]. These predictions are based on 'signals' or features associated with a gene, such as the usual start, stop, exon and intron portions of a gene as well as upstream promoter sequences and more. In this case, these signals would be considered states in the terminology associated with HMMs. Gene finders wish to predict these states based on observations, or sequences presented to the model. HMMs in gene finding tools are trained beforehand and then applied to a sequence. This means that a gene finding program may not be trained in the context of any assembly provided to it, and thus may miss genes that are unique to the assembly in question. On the other hand, while still relying on HMMs for a 'base' set of predictions, evidence-based gene finding tools leverage new evidence that may be outside the scope of the pre-existing model[**?**]. As an example, an evidence-based model would be useful in a situation where you are interested in annotating a new assembly for a non-model organism. The addition of experimental data provides context specific to your assembly of interest while still retaining the predictions from existing HMM models.

There are also other aspects of gene finding tools that are important to consider. These include features such as whether or not the gene finders find non-coding RNAs, annotation of 5' and 3' UTR regions, and in the case of ab-initio methods, the assumptions made by the underlying models used for gene finding. These features and others can influence a user's decision on which gene finding tool to consider and will complicate comparative analysis of multiple gene finding tools. (citation needed somewhere in here)

2.10 File Formats

2.10.1 FASTA

Sequences used in the asembly and gene finding process

3 Data and Methodology

3.1 Methodology

3.2 Data and Processing Overview

3.2.1 Data

Comparing the performance and features of gene finding tools, both qualitative and quantitative, in the context of any set of genomes is important for those interested in selecting a specific gene finding tool. To accent(?) the processing for genomes of interest, those being DC1 and Tsht20, we should include other previously assembled *Trichoderma* assemblies. Currently selected genomes include *Trichoderma reesei*, *Trichoderma harzianum*, and *Trichoderma virens*, with *Trichoderma reesei* being the 'reference' in this case, as it is well studied and there are several patents involving it's use a organsim for production of compounds such as antibiotics in industrial applications.

The general methodology for this work is described in figure 1. Each portion of this figure is discussed in detail in this section.

3.3 Assembly and Annotation

In an attempt to produce high quality assemblies of DC1 and Tsth20, We decided on a set of tools named NextDenovo and NextPolish as they have produced excellent assemblies based on previous experience. (should find a citation to confirm this)

(Might be better for discussion or omitted since it is specific to our setup) Initial attempts to run the example dataset resulted in permissions errors due to the management of the storage system being used, which were encountered with other tools in the past. To remedy this, the software installation was copied to RSMI's scratch space on Copernicus. Once the approriate permissions were given to run nextDenovo, the example dataset was run without issue.

Following assembly using nextDenovo, Illumina sequence data from DC1 and Tsth20 was used to polish each respective genome using nextPolish. Default parameters were used from assembly except for modification of the parallel option to reduce processing times.

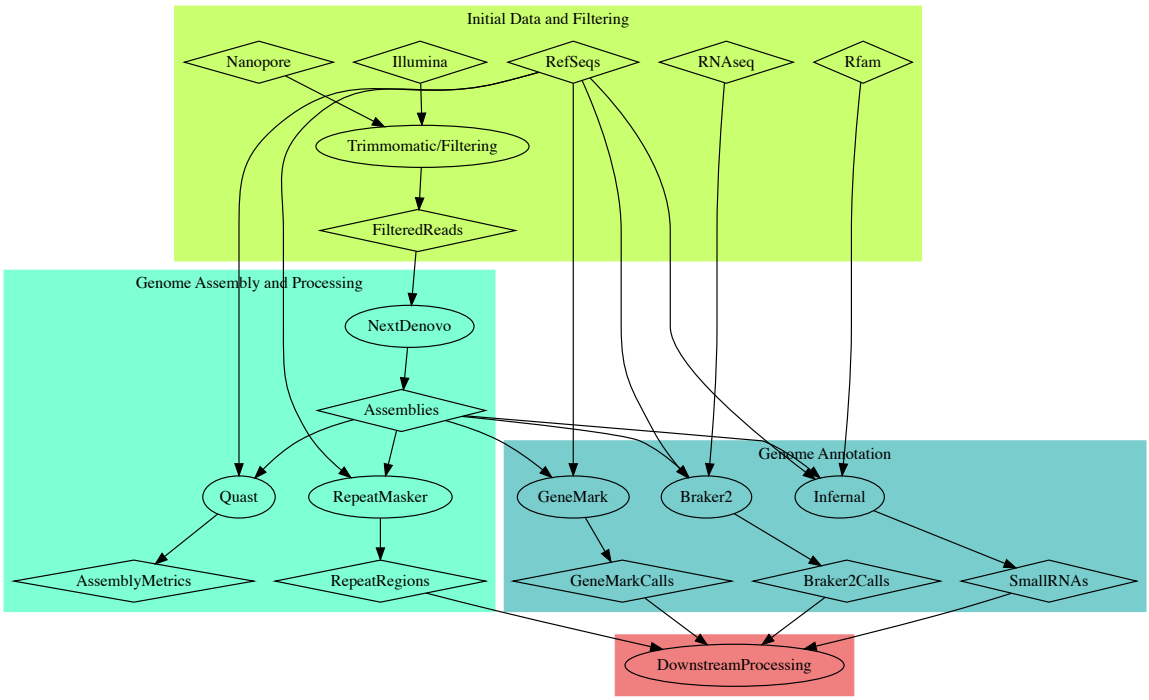


Figure 3.1: A flowchart of the methodology followed for this research. Sections are separated based the general process they are associated with (i.e. input data, assembly, gene finding and downstream analysis).

3.3.1 Repeat Masking

In order to evaluate the performance of gene finding tools in repetitive or low complexity regions in the context of *Trichoderma* genomes, we must first identify said regions in the genomes considered. To do this, RepeatMasker has been selected as a tool to identify repeat regions based on a fungal subset of the Dfam database by specifying the fungi species tag to RepeatMasker when running the program. The program was configured with options to produce several output formats for each genome considered, which will allow for more informative downstream analysis of results. All commands for repeat masking are located withing the processing directory for each strain/genome.

(probably more suited for additional materials at the end?) General command for running RepeatMasker:

/datastore/Roots/Connor/masters/software/repeatmasker/RepeatMasker/RepeatMasker -pa 10 -a -small -species fungi -html -gff -dir ./ path-to-genome/genome.fasta

3.3.2 GeneMark-ES

To begin, GeneMark-ES was run as it requires no prior information or alignments in order to run. In this case GeneMark-ES has an option specifically for fungal genomes, which was used in this case. Apart from the fungal option, the only additional options supplied were for output format of GFF3 and number of cores for reduced processing time.

General command structure for GeneMark-ES:

gmes_petap.pl –ES –fungus –format gff3 –cores 48 –sequence /path/to/sequence

3.3.3 Braker2

As mentioned previously, *Trichoderma reesei* was selected as the ŕeferencegénomeÓ for this work. With this in mind, several short read archives (SRAs) from *T. reesei* were selected for Augustus training. Following Augustus training, the model for *T reesei* was applied to all genomes considered. Settings and procedures from running Braker2 are described below.

The variables that need to be set are AUGUSTUS_CONFIG_PATH and TSEBRA_PATH. Augustus, by defuault, tries to write species information to the location where the software is installed. In this case, we don'thave write permissions to the compute canada software stack hosted byt Research Computing, so the AUGUSTUS_CONFIG_PATH variable must be set in order to create a writeable directory. As long as that path has a directory within it called braker, and a species directory within the braker directory, things should go smoothly. TSEBRA is a set of scripts also made by the creators of Braker and is required to merge results from the various gene prediction tools involved in the Braker2 pipeline. The TSEBRA_PATH simply points to the directory where TSEBRA is located Both Braker2 and TSEBRA can be cloned directly from GitHub (links to come)

3.4 Identification of Overlapping Features and Regions

Feature Identification: To first undertand how gene prediction tools perform in comparison to other gene prediction tools, we must identify features. This identification of features will help us descirbe the similarities, and differences between gene finding tools. A feature, in this context, is any feature stated within a Genomic Feature Format file (GFF) provided to the program, in which mutliple GFF files can be provided. The definition of a feature, for this application, is an object that contains a contig ID, a start position, an end position and a strand property. In the context of features on different strands, start and stop positions of features are sorted based on left and right positions of the feature in respect to the reference sequence.

Region Identification: In addition to feature creation, we will also identify regions of overlapping features based on the precitions from each gene finding tool. These regions will help identify the agreements, or disagreements, between different gene-finding tools. A region, in this context, is a set of overlapping features, all of which overlap at least one other feature in the region. With each overlap, there will be an overlap type. These types can be defined based on Allen's Interval Calculus (reference), with the exception of features that start beyond the end point of the current region.

======= Example command for braker2:

/scratch/p2irc/p2irc_rsmi/cbe453/masters/software/braker2/BRAKER/scripts/braker.pl –gff3 –threads 60 –TSEBRA_PATH=/scratch/p2irc/p2irc_rsmi/cbe453/masters/software/braker2/tsebra/TSEBRA/bin/ –genome /path/to/sequence –species=TreeseiFungal –fungus –useexisting

3.5 Analysis of Results

After completion of the processing portion of this work, the results must be processed in a useful way, which includes both the biological implications of the gene calls as well as the computational, or gene finding features, of the the selected programs. To better understand how gene finders perform in these two classes, we must define an appropriate plan for analysis of the results produced so far. Currently, downstream analysis plan has been broken down into several sections.

3.5.1 Basic Analysis

Basic analysis of gene finding results is an important part of this research. Total gene, transcript and protein counts will be identified for each genome and gene finding tool combination. Comparing the general outputs of these programs will provide an idea of their performance in different *Trichoderma* genomes In addition to these basic outputs, analysis will also be performed for the following: distribution of gene lengths, intersection of gene calls, smallRNAs and repetetive regions, shared gene content with a close fungal relative. Analysis for these results can be performed through simple shell scripting with grep and other unix tools, although processing through Python might provide results that are easier to reproduce with proper

programming. Having one script with several modules that can be rerun at will would be easier to handle than multiple shell scripts. This thinking for processing will be applied to subsequent sections of this as well.

3.5.2 Distribution of Gene Lengths

One important aspect of gene finding tools to consider is the distribution of gene lengths predicted by each individual tool. Certain tools, such as GeneMark are based on pre-defined models, which may limit the length of predicted genes, while tools such as Braker2, which incorporate RNAseq data, may predict a wider distribution of gene lengths depending on the input dataset used. Regardless, the ability of a gene finding tool to predict a wider range of gene lengths can be usefull if users are looking for short or larger genes. To help determine whether or not these tools find shorter genes, or small RNAs, the genomes of interest have been annotated using Infernal along with the Rfam database to identify small RNAs as a ground truth. These annotation results will also be included with results from other annotation processes further down the line. Again, these results can be produced with a Python script. The resulting data could then be used as input to violin plots for each genome and set of tools considered in this analysis process. Violin plots should provide a good visualization of gene lengths as well as the number of genes found with specific lengths. Means could also be compared staatistically for genomes and the mutliple tools considered as well.

Analysis of gene lengths was performed using a Python script. Combined predicted CDS sequences for each predicted gene were used as input for the total gene length. CDS sequences predicted by Braker2, were directly available in the output directories when the program was run. CDS sequences from GeneMark required extraction of the CDS sequences from the genome FASTA files. This process was performed using the gffread tool from the Cufflinks package. Predicted CDS sequences were loaded into Python using Biopython's SeqIO package. Sequence lengths were then placed in a list and analyzed using a combination of pandas and numpy. A log10 transformation was applied to the sequence lengths as the original distribution was heavily skewed due to long outlier CDS sequences. After transformation, the CDS length distribution appears as a normal distribution, although there are interesting troughs that occur in several of the peaks for several of the genomes considered. These troughs did not appear in a comparable Yeast reference dataset, although the log transformed data still appears to be a normal distribution.

3.5.3 Intersection of Gene Calls, smallRNAs and Repetitive Regions

Annotation of all three features in the title are important in assessing the ability of gene finding tools. Even more important, is the potential for for overlap between gene calls and small RNAS as well as repetitive regions the genomes. As discussed in the previous subsection, the distribution of gene lengths predicted by a gene finding tool can be an important metric for users. Overlapping predicted genes from tools alongside the output from Infernal and the Rfam database may provide insight into whether or not these gene finding tools are able to predict RNAs of very short length. In addition to small RNAs, repetitive regions in *Trichoderma* genomes hold potential for recombination and gene content, although the inherent nature of these repetitive

regions (low nucleotide diveristy) suggests that gene content should be low, based on the nucleotides required for start and stop codons. Analysis of these intersections can be perfomed via bedtools or through biopython (I believe). Again, having all processing steps included in one script as separate functions that can be called at whim will make further processing easier if changes need to be made.

3.5.4 Methodology for Indetifying Overlapping Features

To analyze the results from multiple gene-finding tools, we must first define two conceptual topics.

Definition of a Feature: First, a feature, in the context of this research, is any item contained within a Genomic Feature Formated file (GFF). Each feature contains a contig ID, a start position, end position, and a feature ID based on the information from the GFF file. These features will be used in the process of identifying regions, or overlapping features from prediction tools.

Definition of a Region: Secondly, a region is defined as any overlap between features relative to the reference sequence being considered. To identify regions, every feature from each GFF file being considered, is sorted by start position. For clarification, the start position is based on the left most position in the GFF file, regardless of the strand that the feature is predicted on. Once the features have been sorted by left position, the sorted features are iterated over to identify regions, as long as the left position of the next feature is consistent with the left position , within the left and right position, or equal to the right position of the current region. One drawback to this approach is that ideally, identification of overlap types based on Allen's interval algebra would be performed at this point. However, the methodology of initially sorting features based on left position somewhat prevents this process from happening. This implementation requires further processing of identified regions late on the process, simplifying the all to all comparison that would occurr if all features were considered at once.

3.5.5 Shared Gene Content with Yeast

While considering novel gene calls can be useful, comparing those calls to a well-studied close relative can provide a rudimentary validation of the calls as a ground truth. In this case, a comparison to Yeast will be made. The agreed upon number for successful gene-finding as compared to Yeast is roughly 80-85 percent of gene content. This will confirm that at least most of a closely related fungal genome's content is predicted and shared by the gene calls for *Trichoderma*. Results for this processing can be produced with a simple BLAST search and apropriate cutoff values (i.e. query coverage, percent nucelotide identity, E-score, etc.). Other tools for evaluation will certainly be considered, although I will need to look into this process further.

3.5.6 Comparative Genomics

With the data produced by this research, it is possible to perform som commparative genomics (time permitted), mostly related to the assemblies generated during this work along with the RefSeq genomes

included from NCBI. Mummer is a potential tool to use for all to all genome alignments, although there may be difficulty in the ordering of contigs/scaffolds/chromosomes when performing thes alignments. This work is not necessarily required but would be interesting from a biological perspective to identify rearrangements, inversions and such.

4 Results

4.1 Assemblies of DC1 and Tsth20

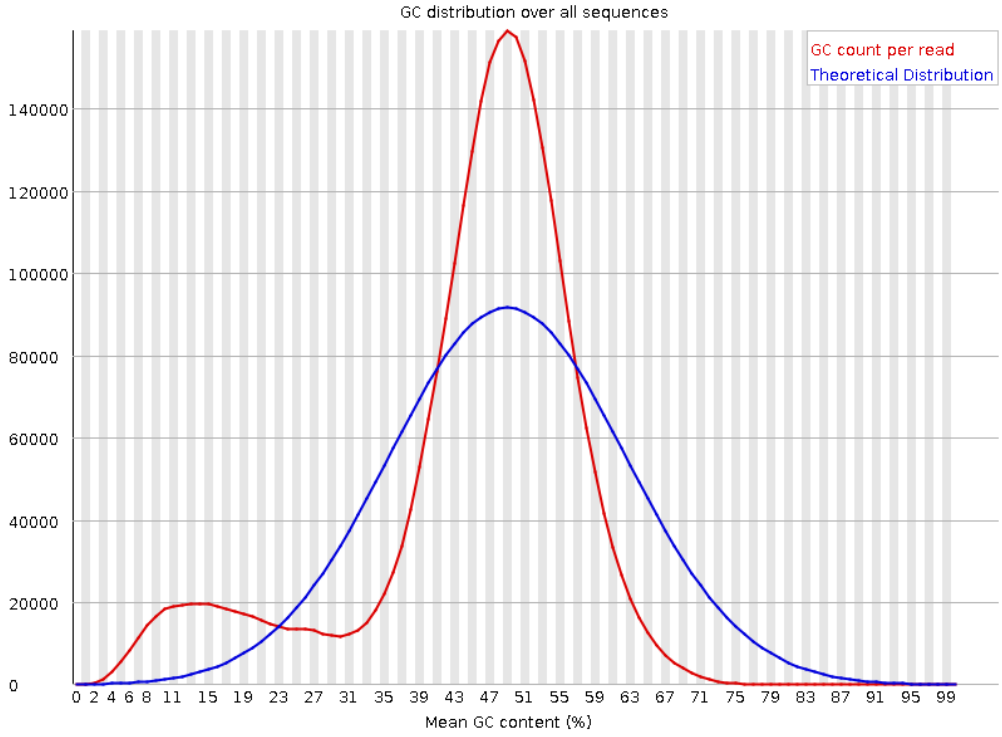
Prior to assembly of DC1 and Tsth20 sequences, the tool FastQC was used to evaluate the quality of the Illumina sequences provided for this project. After trimming the Illumina sequences for low quality reads, one FAIL flag was raised for both samples. This was the per-sequence GC content flag, indicating that the GC content of high-quality did not meet expectations. Plots of GC content for the trimmed R1 sequences of both DC1 and Tsth20 are shown in figure 4.1. In these plots, there is a considerable increase in the number of sequences with GC content between 2 and 30%, which is important to note for later results and discussion.

A sliding window analysis was also performed on all final assemblies in order to identify regions of anomalous GC content. The results of this analysis are shown in figure 4.2. Of the included assemblies, anomalous GC content was identified in DC1, Tsth20, *T. reesei* and *T. harzianum*, with *T. virens* showing no anomalous GC content. In addition to the confirmation of anomalous GC content, it appears that the distribution of GC content in *T. reesei* differs from the other assemblies.

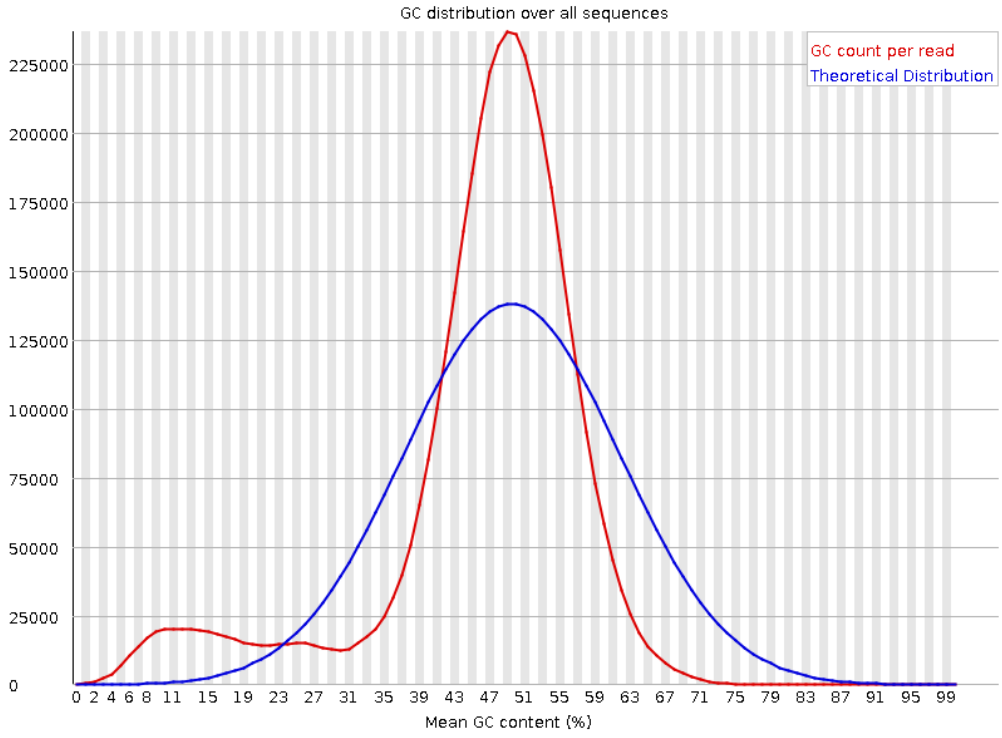
For other general assembly metrics, the QUAST tool was used. Results from QUAST are shown in figure 4.1, from which we can make several observations. The most obvious observation to start with is total contig counts for each assembly. For DC1 and Tsth20, the total contig counts are an order of magnitude smaller when compared to the other NCBI RefSeq assemblies, inidicating highly contiguous assemblies from nextDenovo and nextPolish. This is likely due to the use of long-read sequencing used in the assemblies of DC1 and Tsth20. The total assembly lengths are similar, hovering around the 38-42Mb range, except in the case of *T. reesei*, which is known to have a significantly smaller genome length (ref) at roughly 33Mb. The largest contig size for each assembly vary 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 at minimum three times larger than others assemblies.

4.2 Initial Gene Finding Results

In figure 4.3, we see a summary of total genes and mRNAs predicted by each of the selected tools. An immediate trend can be seen in this data. The total predicted features for Braker2 are significantly lower than those from GeneMark in all assemblies except for *T. reesei*. This may be due to Braker2 using RNAseq data



(a) DC1



(b) Tsth20

Figure 4.1: Plots showing the distribution of GC content in Illumina sequences.

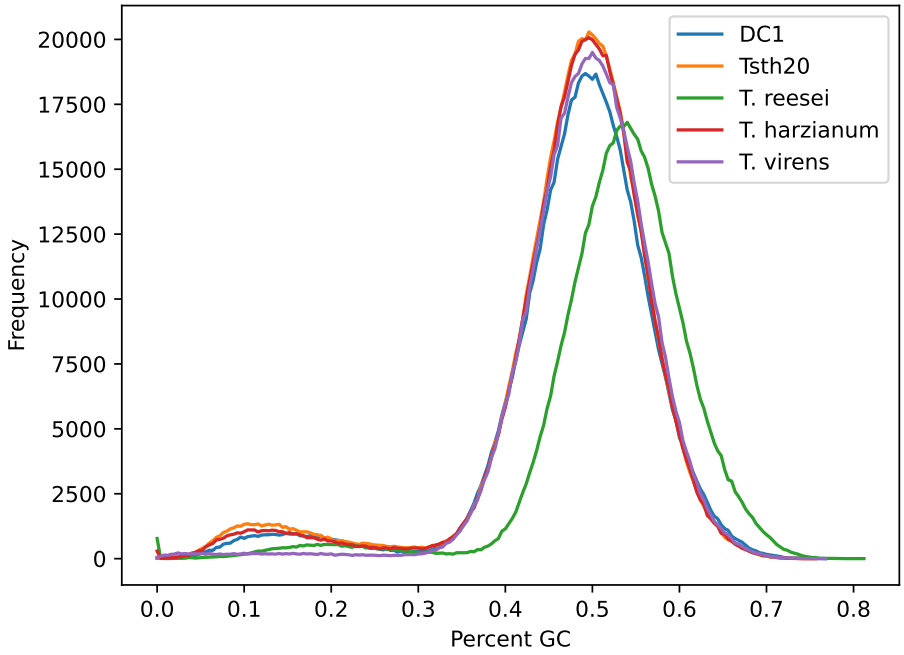


Figure 4.2: Plots showing the frequency of GC values calculated from sliding windows for each assembly.

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 4.1: General assembly metrics produced by QUAST (a genome quality assement tool).

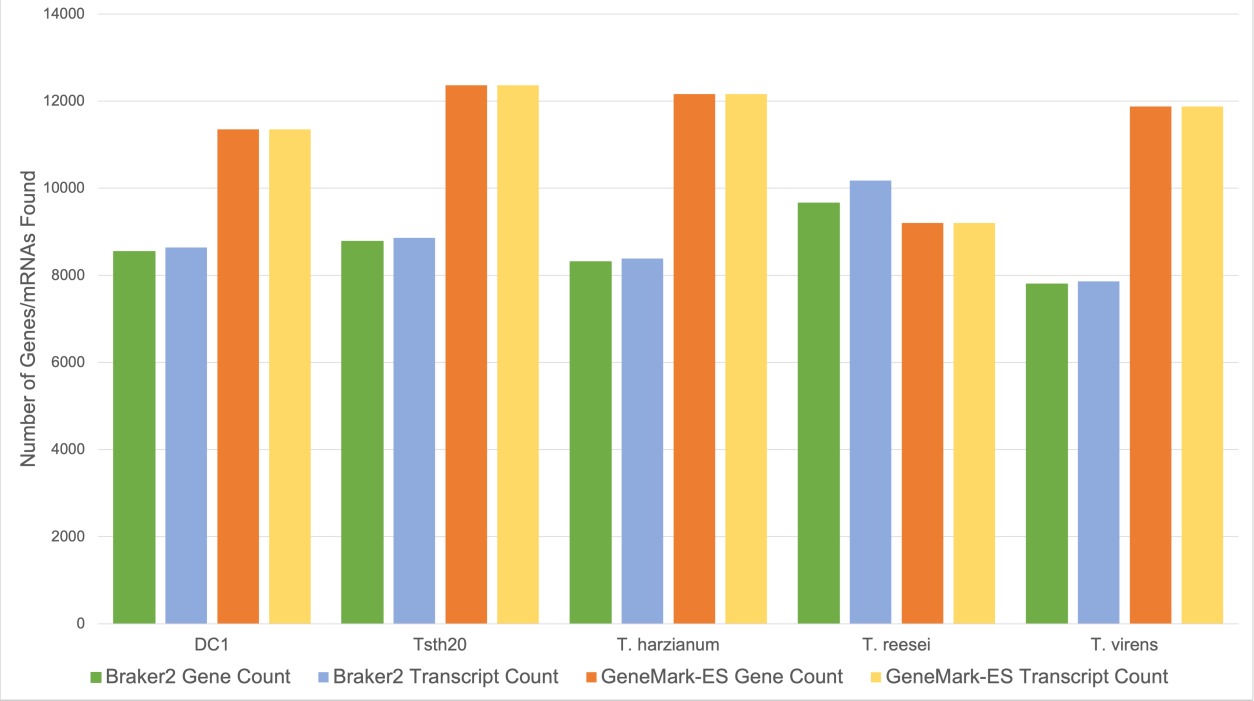


Figure 4.3: Shows the counts of genes and mRNAs found by each gene finding tool for each genome assembly considered.

from *T. reesei* during its training process, which we will cover in the discussion section. Another observation can be seen when comparing the predicted genes and predicted RNAs for the same tool when applied to the assemblies. GeneMark does not appear to identify any additional isoforms, only reporting the entire gene structure. Braker2 does identify isoforms, although very few of them. This may be related to the training set provided to the training process, although this is not yet confirmed. Overall, it appears that GeneMark regularly predicts a higher number of features when applied to these assemblies.

One important ascpect to consider when looking at the output of different gene prediction tools is the distribution of sequence lengths predicted by any given tool. Lengths of possible sequences can vary widely, ranging from small non-coding RNAs, which can be less than 200 nucleotides in length, up to the largest genes which cover more than two kilobases. Due to this wide variation in possible sequence lengths, it is possible that different prediction tools could produce different distributions of predicted sequence lengths. This is important if researchers are interested in small non-coding RNAs or atypically large genes. This section will investigate the coding sequence lengths predicted by Braker and GeneMark for DC1 and Tsth20 while the RefSeq assemblies will also include the RefSeq annotation.

4.3 Dsitribution of Predicted Gene Lengths

In, gene finding tools will predict genes of different lengths within a genome. To better understand the output produced by each gene finder, we inspected the distributions of CDS lengths predicted by each gene

Assembly	Regions (total)	RefSeq	Braker2	GeneMark
DC1	11268	NA	8485	11265
Tsth20	12285	NA	8747	12279
T. harzianum	13377	13122	8051	11529
T. virens	12045	11864	7554	11413
T. reesei	9818	8836	9202	8897

Table 4.2: Counts of regions identified in total and total number of regions where a prediction from each individual tool was found.

finding tool for each assembly. To better understand the distribution of coding sequence (CDS) lengths produced by each gene finding tool, running sums of CDS lengths proportional to the total sum of CDS lengths were plotted and are visible in figure **??**

4.4 Region Identification

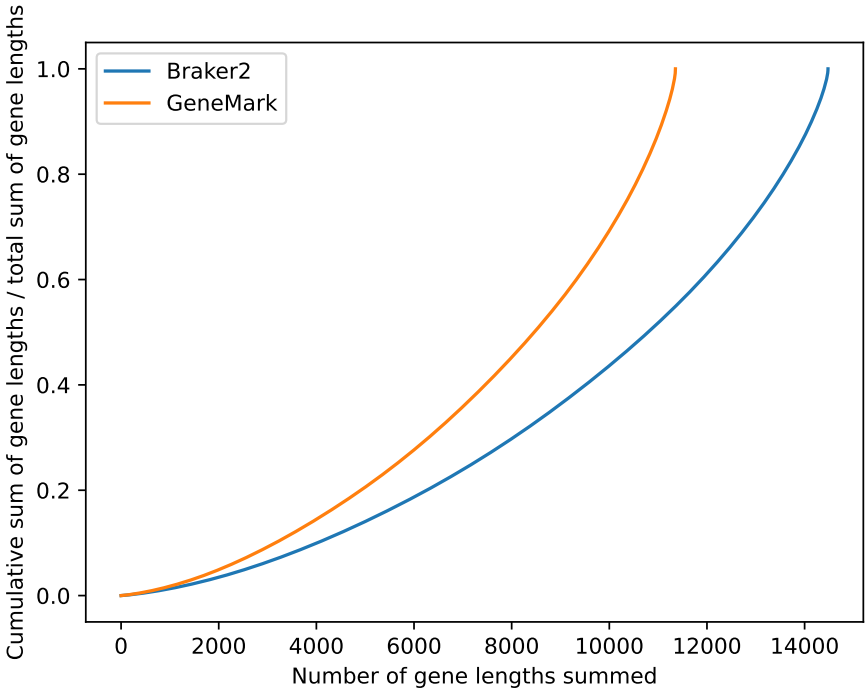
Based on the methodology described earlier, we have identified sections of contigs with one or more overlapping predicted sequences (features). The result of this processing is a GFF formatted file with each entry containing information about a region on a specific contig. Also contained in each entry, is a list of all features included in that region with start/stop position and strand, the tools used to predict the features, and a count of how many prediction tools that were included in said region. Initial analysis of the output is discussed in this section.

With our results, the first interesting area to look at is simply the total number of regions identified for any given assembly as well as number of regions that each tool was a member(weird wording) of. In figure 4.2, we see that Braker2 deviates from both the RefSeq and GeneMark predictions greatly, except in the case of *T.reesei*. As speculated earlier, this may be due to the training dataset supplied to Braker2 as it happens to be present in a similar number of regions when compared to RefSeq and GeneMark in *T. reesei*. In general, it appears that the RefSeq predictions are present in more regions than both Braker2 and GeneMark.

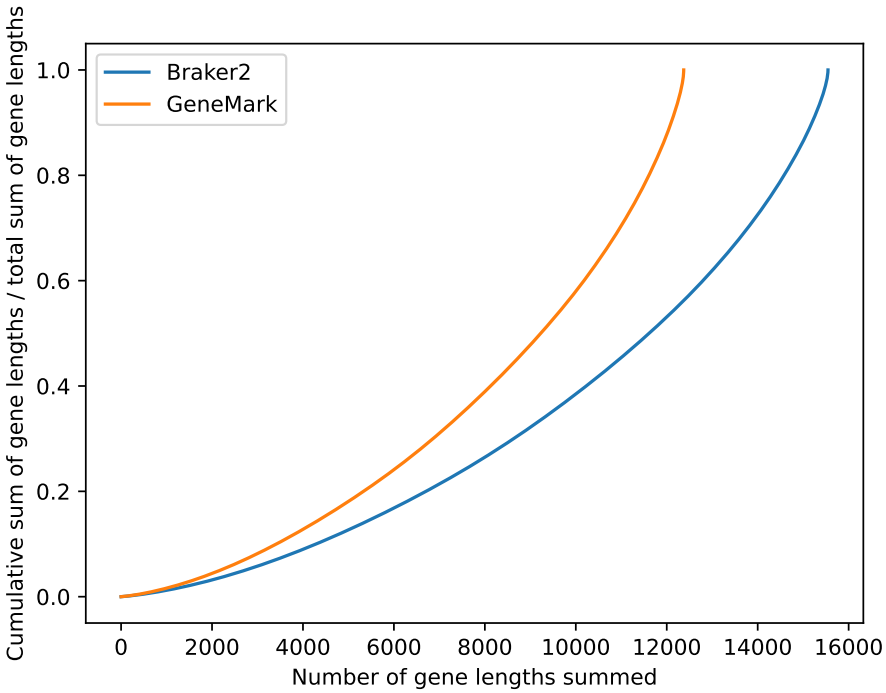
Figure 4.2 demonstrates that the composition of regions can differ significantly based on the tools used. To illustrate agreement of tools in regards to identified regions, Venn diagrams were generated showing agreement of predictions in a region for each tool.

4.5 Genes in Regions of Anomalous GC Content

Evaluating gene finder performance in regions of anomalous GC content is one of the key topics of this research. One simple way to evaluate performance is whether or not gene finding tools predict genes uniformly



(a) DC1

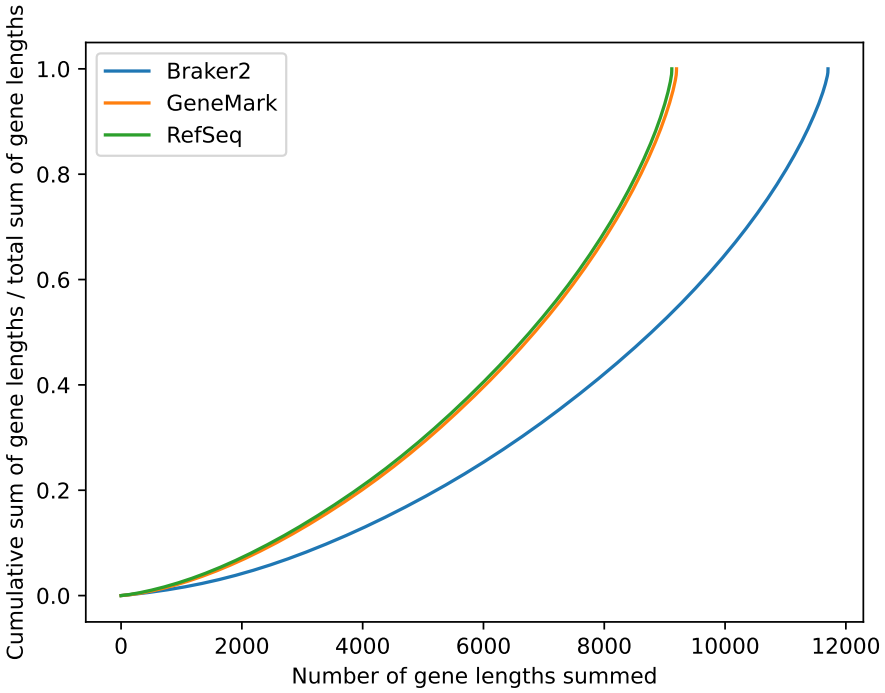


(b) Tsth20

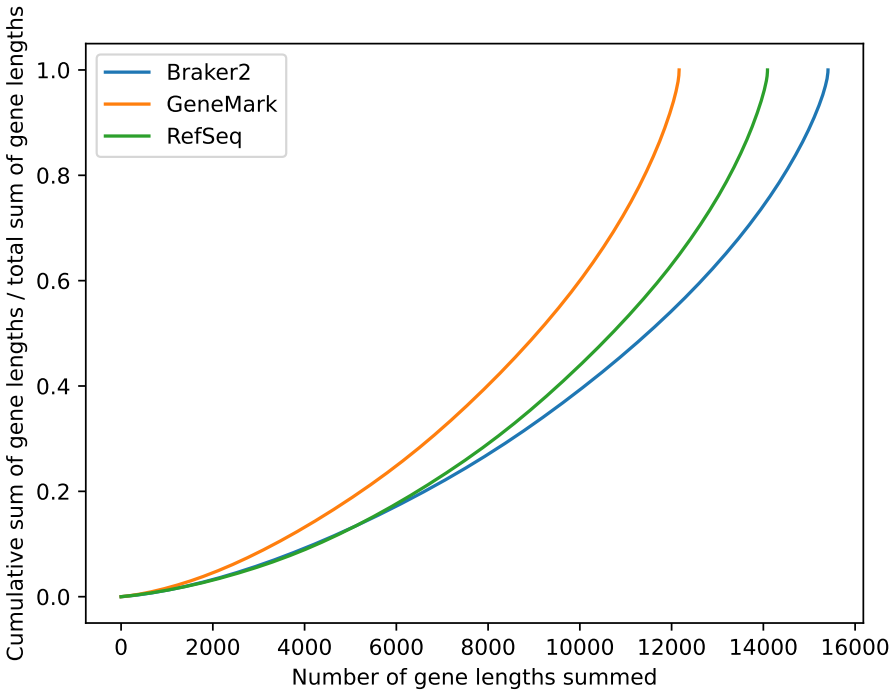
Tool	DC1	Tsth20	*T. reesei*	*T. harzianum*	*T. virens*
Braker2	$9.56^{-}181$	$1.14^{-}259$	$2.68^{-}96$	$4.05^{-}140$	$1.35^{-}35$
GeneMark	$5.12^{-}216$	0.0	$5.66^{-}49$	$5.37^{-}219$	$5.31^{-}35$
RefSeq	N/A	N/A	$1.29^{-}49$	$2.44^{-}205$	$7.40^{-}33$

Table 4.3: p values produced from a two-sided binomial test for each combination of tool and assembly.

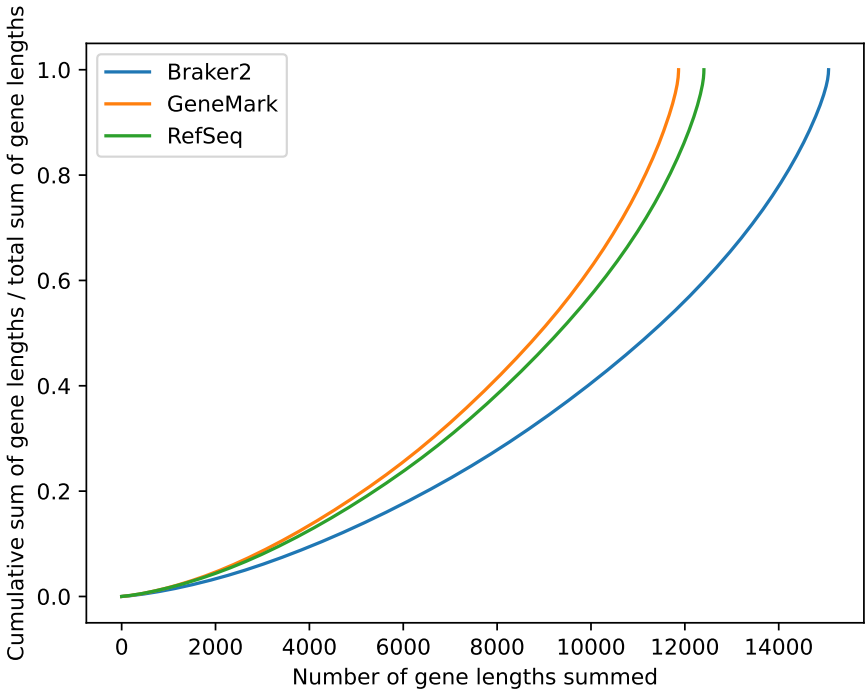
throughout a given sequence. Biologically, we know that regions of anomalous nucleotide composition are less likely to contain coding sequences than typical genomic regions, leading us to the problem of first identifying predicted genes in standard and anomalous regions. After identifying low GC segments within each assembly, we can include them in the region identification method. From this result, we classified predicted genes into two classes; genes in regions with normal GC content, and genes in regions with anomalous content. In this case, anomalous content is defined as a window of genomic sequence containing a percent GC composition of 28% or lower. This number was chosen based on the plots of GC content presented in the assembly section of the results. After classifying predicted genes, two-sided binomial tests were performed with the null hypothesis being that predicted genes are distributed uniformly throughout an assembly. Framed differently, we expect the sum of genes predicted in both regular and irregular regions to be proportional to the sum of lengths of those regions, respectively. This is not the case as demonstrated in table 4.3.



(c) *T. reesei*



(d) *T. harzianum*



(e) *T. virens*

Figure 4.4: Running sums of CDS lengths proportional to total CDS length for each gene finding tool and assembly.

4.6 Discussion

4.7 Conclusions

References

[1] 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, mar 2021.

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

[3] Alexandre Lomsadze, Vardges Ter-Hovhannisyan, Yury O Chernoff, and Mark Borodovsky. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 33(20):6494–6506, 2005.

[4] Niranjan Nagarajan and Mihai Pop. Sequence assembly demystified. *Nature Reviews Genetics*, 14(3):157–167, 2013.

[5] 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.

[6] R Keith Slotkin. The case for not masking away repetitive DNA. *Mobile DNA*, 9(1):15, 2018.

[7] Ioannis S Arvanitoyannis Theodoros H. Varzakas and Haralambos Baltas. The Politics and Science Behind GMO Acceptance. *Critical Reviews in Food Science and Nutrition*, 47(4):335–361, 2007.

[8] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics*, 13(1):36–46, nov 2011.

[9] 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.

[10] 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.