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, 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 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	Megabases
Kb	Kilobase
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 Results and Discussion

1.1 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 1.2. Following this, is Section 1.3, 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 1.4 followed by more detailed analysis of predicted CDS lengths in Section 1.5. Gene predictions from each gene finder and assembly combination are supplied to the BUSCO metric process in Section 1.6. Results from the initial spatial comparison of gene predictions using the region identification process are then presented in Section 1.7. The region identification process is then applied with additional information from tblastn alignments, InterProScan annotation, and GC sequence composition in Sections 1.8, 1.9 and 1.10, respectively. A final comparison of all results and selection of an appropriate gene finder are then discussed in the Conclusion **??**.

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.

1.2 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 1.1, 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 [4] and nextPolish [4]. 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 [5], 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 1.1: 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 [5].

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 Figure1.1. 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 1.1. 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 1.1, 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 1.10, 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.

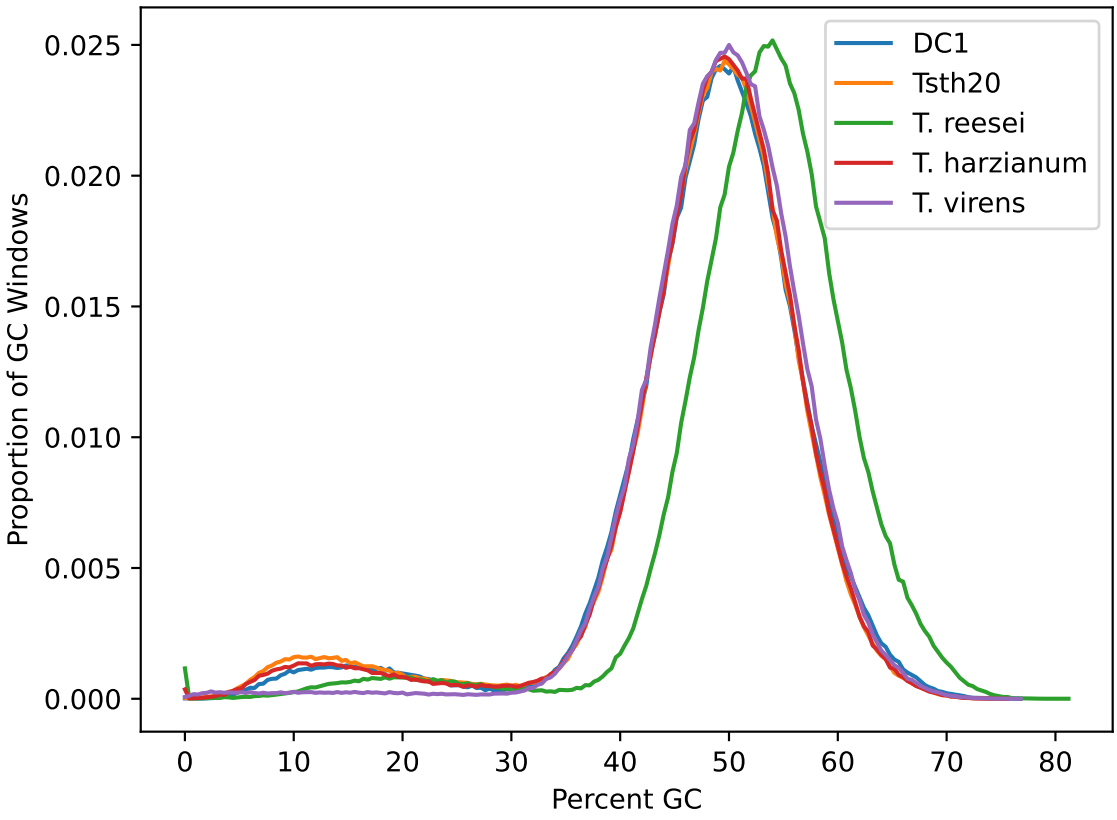


Figure 1.1: Plots showing proportions of sliding windows of GC content for each *Trichoderma* genome assembly.

1.3 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 **??**. 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 [2] 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 [3] 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 [9], BamTools [1], BedTools [8], GeneMark [2], StringTie [7], GFFRead [6] 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.

1.4 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 Table 1.2. 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 GeneMark

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

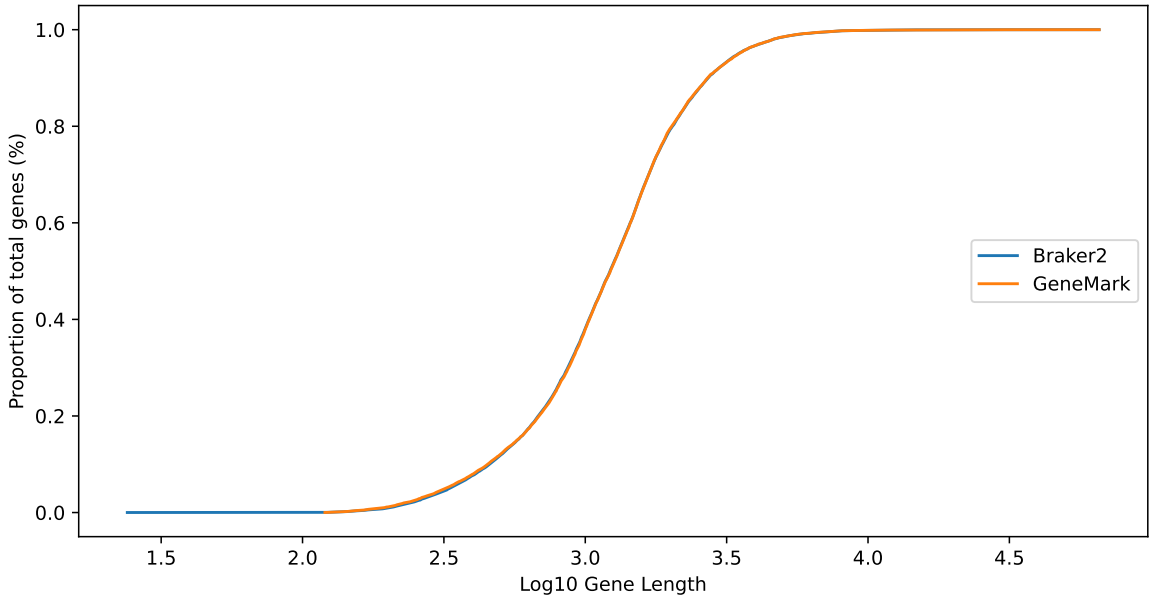
Table 1.2: Number of genes and coding sequences (isoforms) predicted by each gene finder for each *Trichoderma* genome.

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.

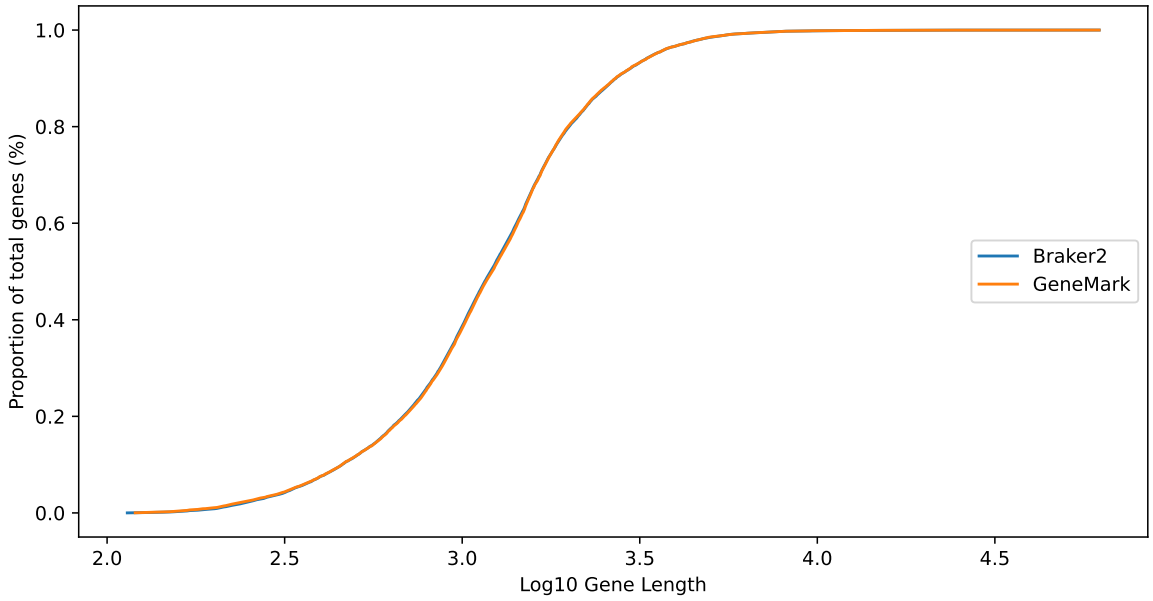
1.5 Distribution of Predicted Coding Sequence Lengths

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 1.2, 1.3 and 1.4. 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, 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 1.3. In the cases of DC1 and Tsth20, we see that in agreement with Figure 1.2, Braker2 and GeneMark do not produce statistically different lengths of genes, which is interesting considering that the Braker2 includes experimental evidence for

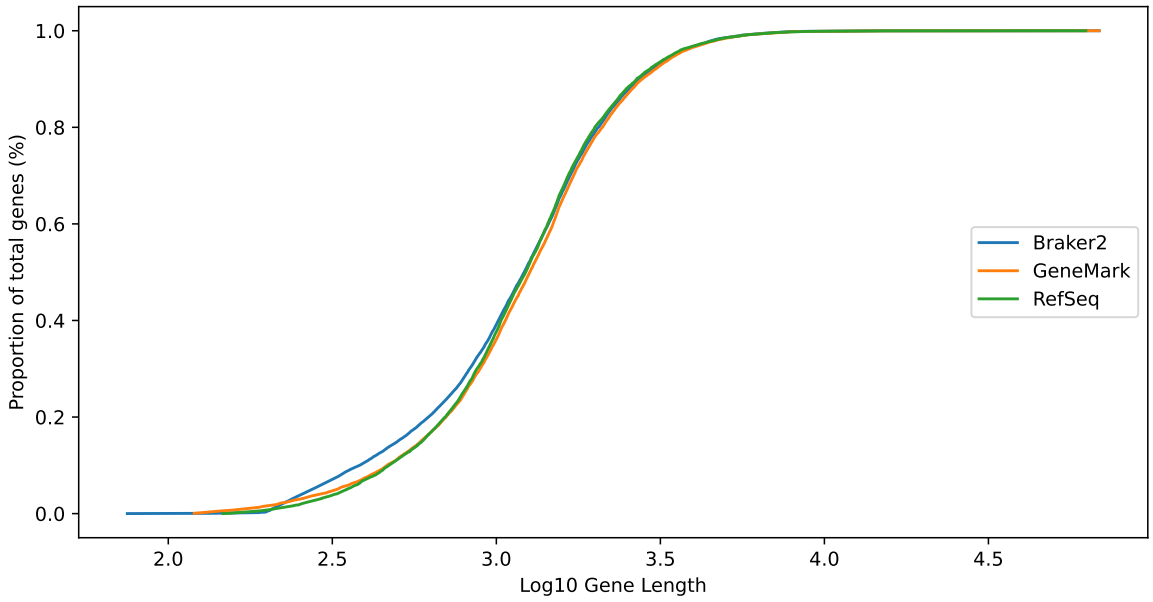


(a) DC1

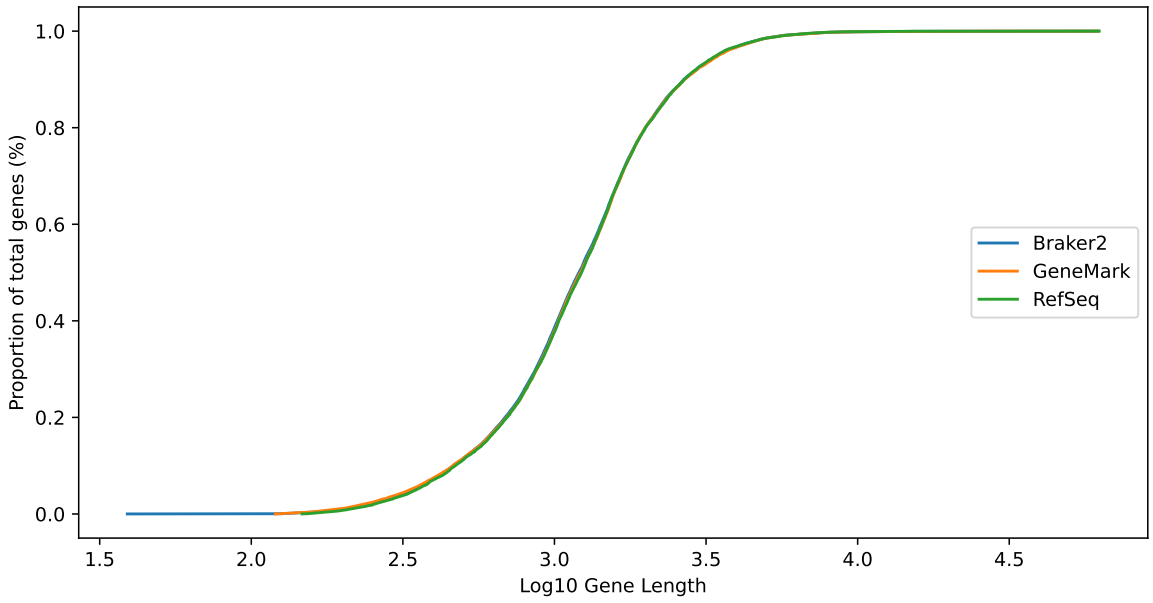


(b) Tsth20

Figure 1.2: Plots of the cumulative distribution function for CDS lengths produced be each gene finding tool applied to DC1 and Tsth20

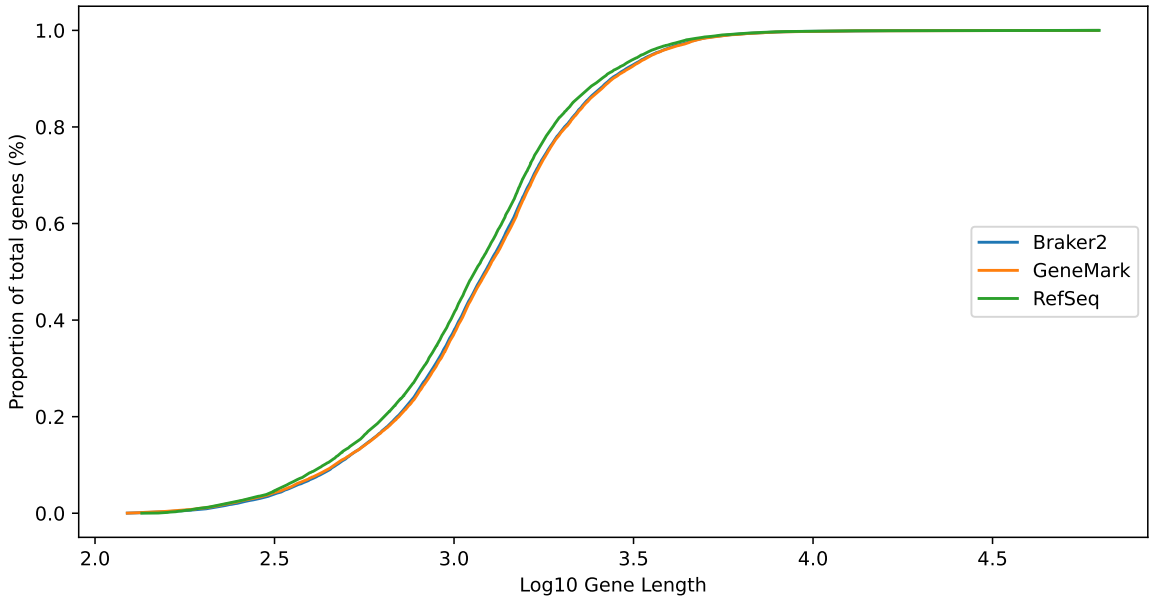


(a) *T. reesei*



(b) *T. harzianum*

Figure 1.3: 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 1.4: 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 1.3: Table of P-values from two-sided two-sample Kolmogorov-Smirnov tests between gene finding tools.

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 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.

1.6 BUSCO Results

Results of BUSCO analysis using the sordariomycetes_Odb12 dataset provided by BUSCO are presented in Table 1.4. The results indicate that all gene sets considered in this analysis have a BUSCO completeness of 94.9% or higher, with a maximum completeness of 99.9% 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 far 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.

In general, all gene finders perform well in regards to BUSCO performance. 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 without matching gene predictions. Table 1.4, shows breakdowns of genes missed by each gene finder across the *Trichoderma* assemblies.

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

1.7 Region Identification

Visual breakdowns of the types of regions identified and their counts for each *Trichoderma* assembly are shown in Figure 1.5. 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

Strain	Complete	Single	Duplicated	Fragmented	Missing
DC1	4319	3505	814	2	2
Tsth20	4309	3575	734	6	8
T. reesei	4269	3616	653	27	27
T. harzianum	4309	3562	747	6	8
T. virens	4309	3504	805	7	7

(a) Braker2

Strain	Complete	Single	Duplicated	Fragmented	Missing
DC1	4313	4311	2	7	3
Tsth20	4304	4291	13	10	9
T. reesei	4297	4297	0	11	15
T. harzianum	4301	4291	10	11	11
T. virens	4301	4291	10	11	11

(b) GeneMark

Strain	Complete	Single	Duplicated	Fragmented	Missing
T. reesei	4102	4101	1	159	62
T. harzianum	4244	4234	10	52	27
T. virens	4174	4164	10	116	33

(c) RefSeq

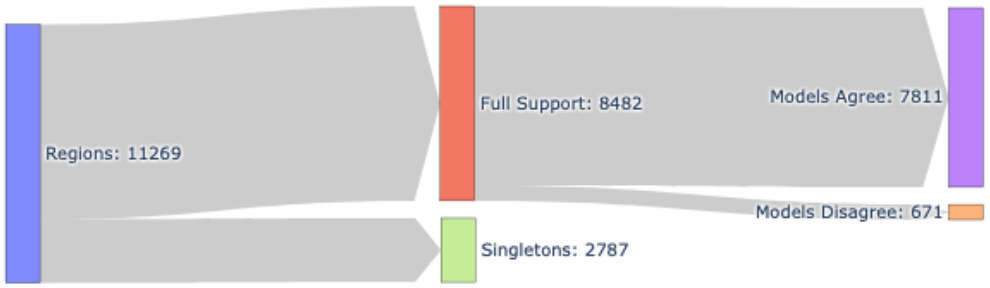
Table 1.4: Results from BUSCO using the fungal analysis option organized by gene finding tool. The selected BUSCO dataset contains 4323 markers. For more information on the categories assigned by BUSCO, please refer to the documentation.

predictions agree on the start and/or stop positions of the gene. Regions with support from only one gene finder are labelled as singletons.

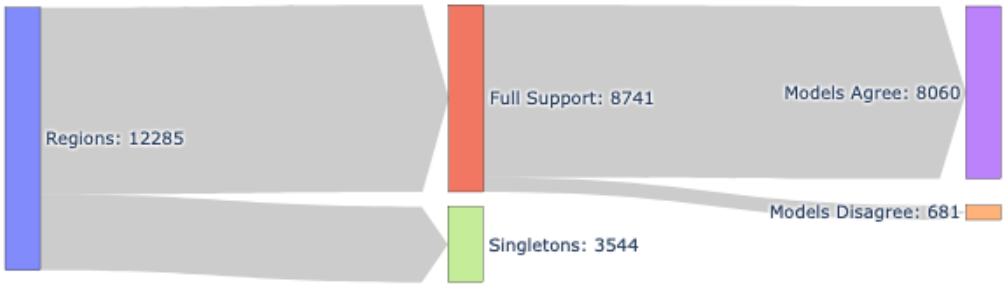
Looking at DC1 and Tsth20 in Figure 1.5, it appears that Braker2 and GeneMark predict genes in the same regions in the 69% and 65% of cases, respectively. For regions with full support in DC1 and Tsth20, the models also tend to agree on the start and stop positions of the gene in that region. This is 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 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 1.6 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.

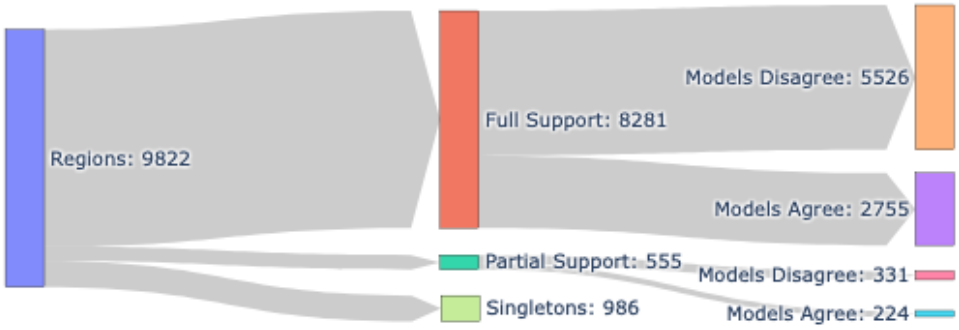
Another interesting set of criteria to investigate is whether or not gene finders always predict genes on the same strand within a region. We observe that regions containing predictions on different strands do exist, and are observed in predictions from all assemblies included in this analysis. Figure 1.7 shows an example



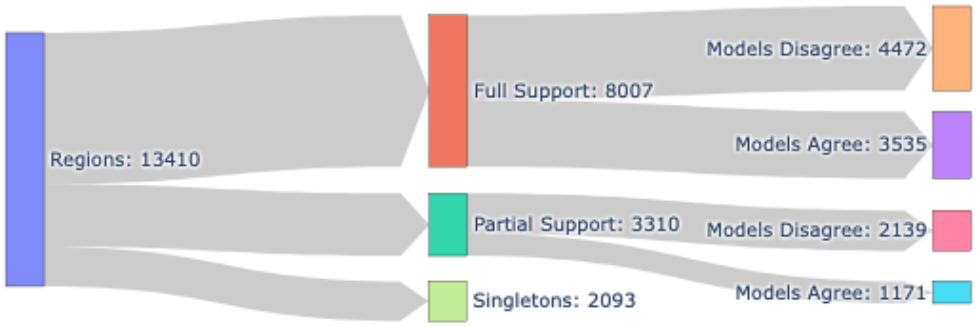
(a) DC1



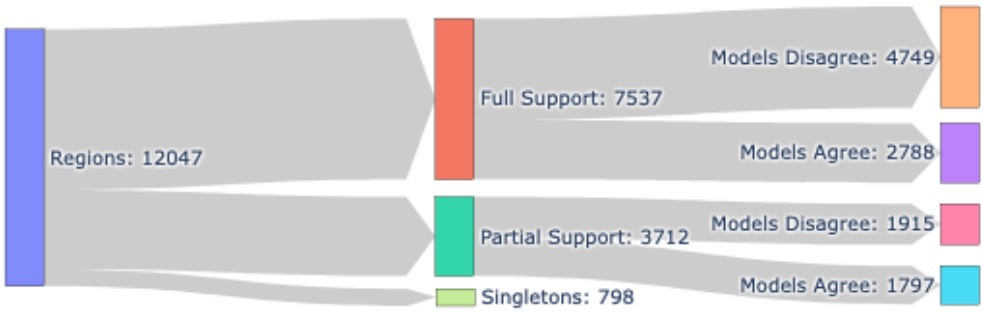
(b) Tsth20



(c) *T. reesei*



(d) *T. harzianum*



(e) *T. virens*

Figure 1.5: 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).

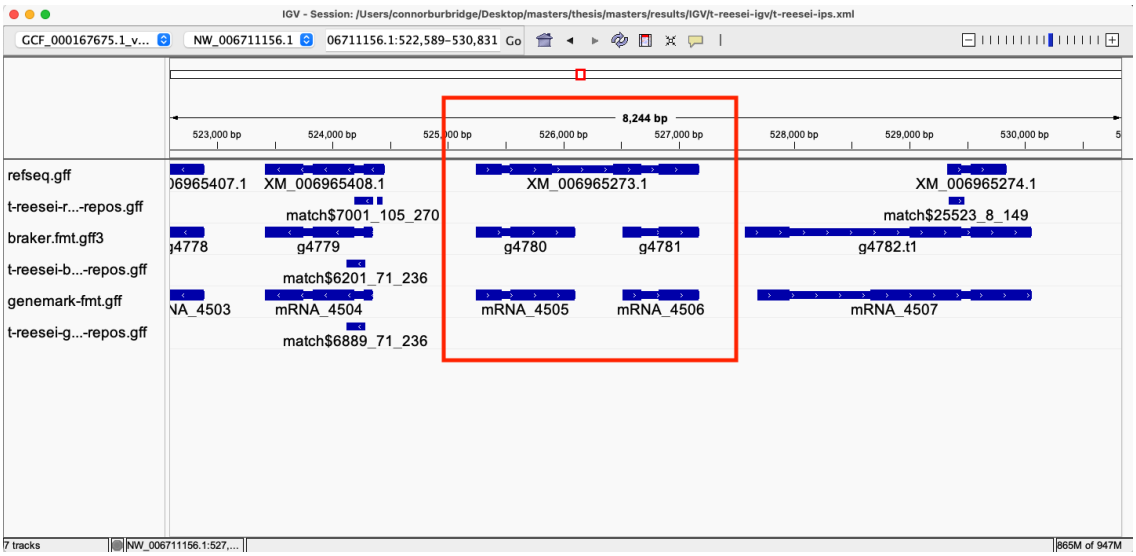


Figure 1.6: An IGV screenshot from *T. reesei* showing a region containing five gene predictions in which one of the gene finders disagrees with the other two, highlighted in the red box.

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.

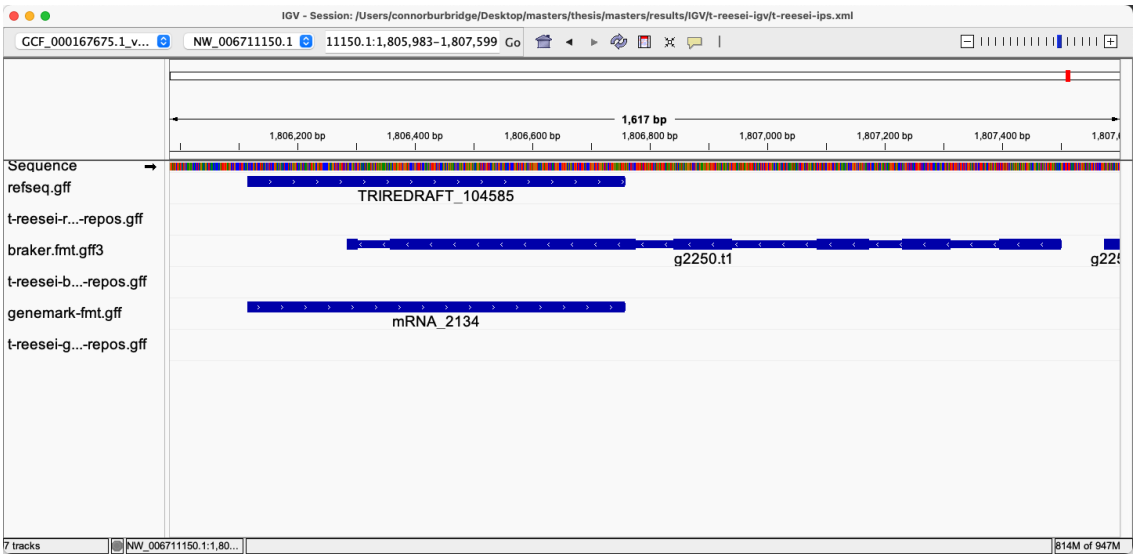


Figure 1.7: An IGV screenshot of *T. reesei* showing a region which contains gene predictions on opposite strands.

1.8 InterProScan as Supporting Evidence for Predicted Genes

Pfam hits from InterProScan analysis are presented in Table 1.5, 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 1.8. 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.

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

Assembly	Braker2	GeneMark	RefSeq
DC1	$\frac{10676}{14479}$	$\frac{8416}{11354}$	N/A
Tsth20	$\frac{11389}{15546}$	$\frac{9168}{12373}$	N/A
T. reesei	$\frac{8471}{11704}$	$\frac{6990}{9196}$	$\frac{6964}{9111}$
T. harzianum	$\frac{11370}{15408}$	$\frac{9061}{12164}$	$\frac{9293}{14065}$
T. virens	$\frac{11249}{15062}$	$\frac{8871}{11866}$	$\frac{9062}{12383}$

Table 1.5: Table showing the fractions of predicted genes with Pfam annotations from InterProScan as a percent

gene models in the region do not agree. There are even cases such as the region shown in Figure 1.9, 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 1.10, 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.

1.9 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 Table 1.6. 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 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

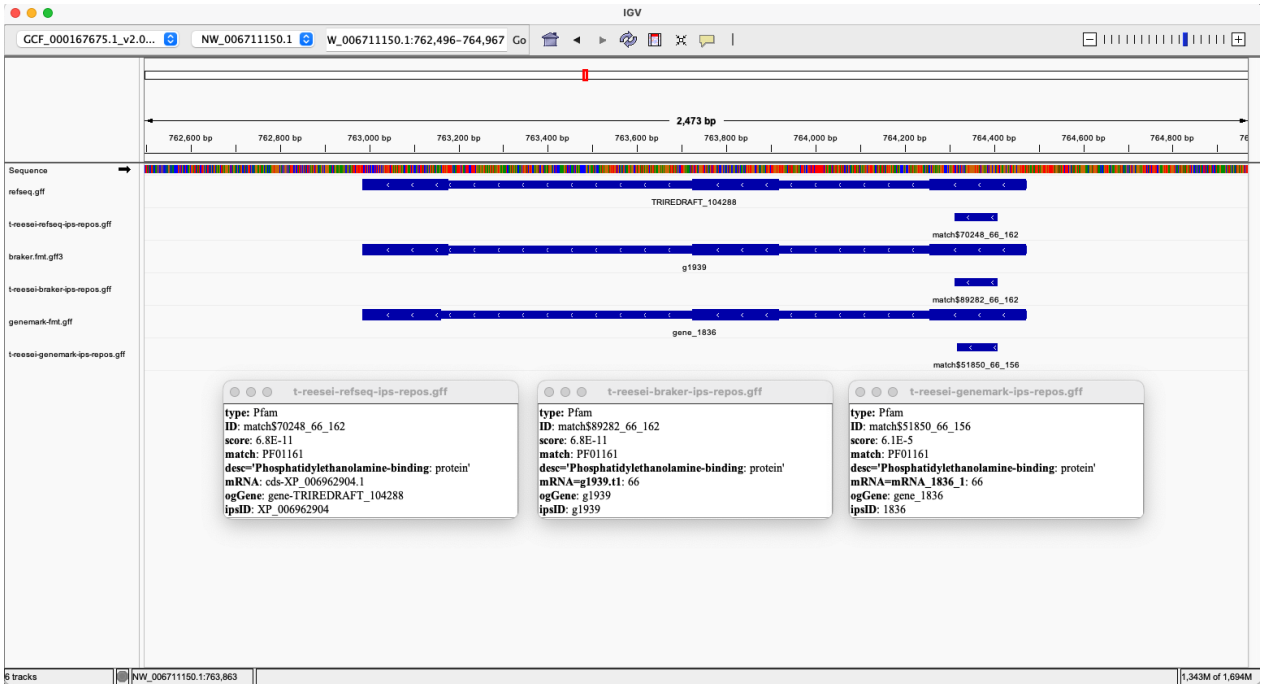


Figure 1.8: An IGV capture showing complete agreement between gene finders for both gene model and protein Pfam hits. The longer segments are predicted genes and the smaller segments are annotated Pfam matches, all with agreeing start and stop positions. The **desc** in each sub-window indicate agreeing 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 1.6: 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.

shown in Table 1.7. In a similar trend to Table 1.6, 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*.

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

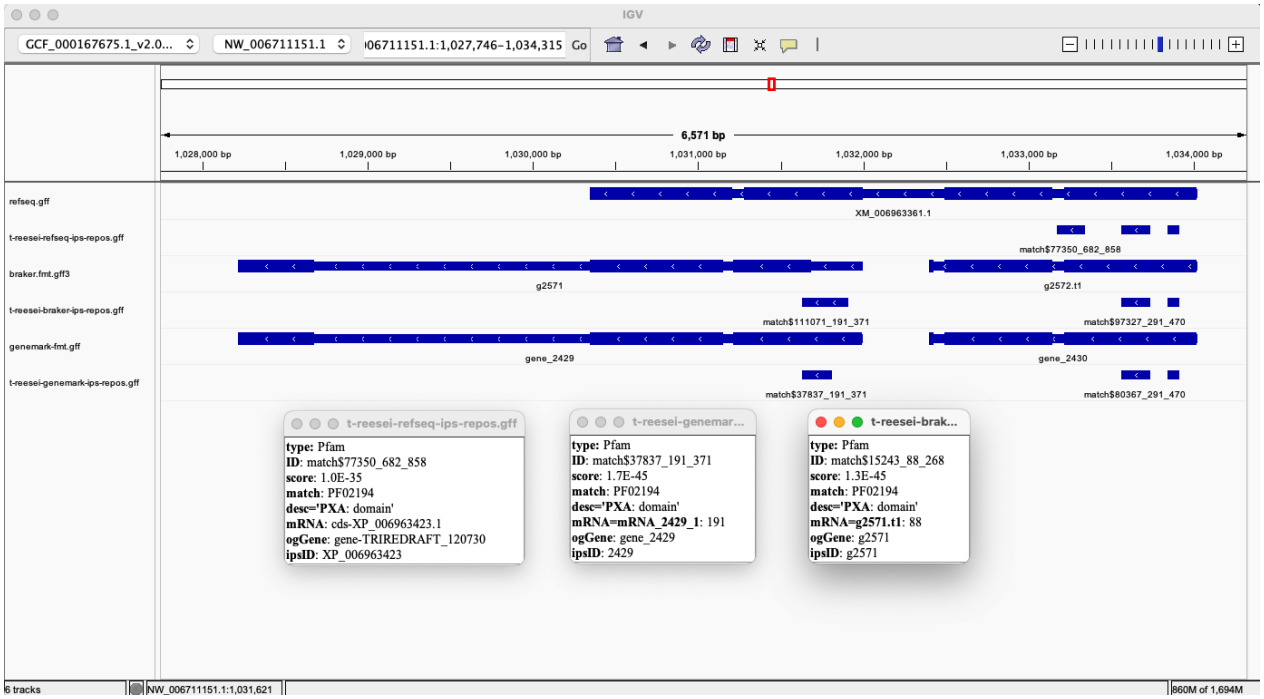


Figure 1.9: An IGV capture showing Braker2 and GeneMark reporting two genes and their resulting proteins and Pfam hits as separate, while RefSeq reports one gene, one protein and three Pfam matches.

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.

1.10 Genes in Regions of Anomalous GC Content

Table 1.8 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 GC composition less than 28% as determined in Section 1.2. In comparison to results from Section 1.7, 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 1.7, there are no regions with partial support as only two gene finding tools were applied to those assemblies.

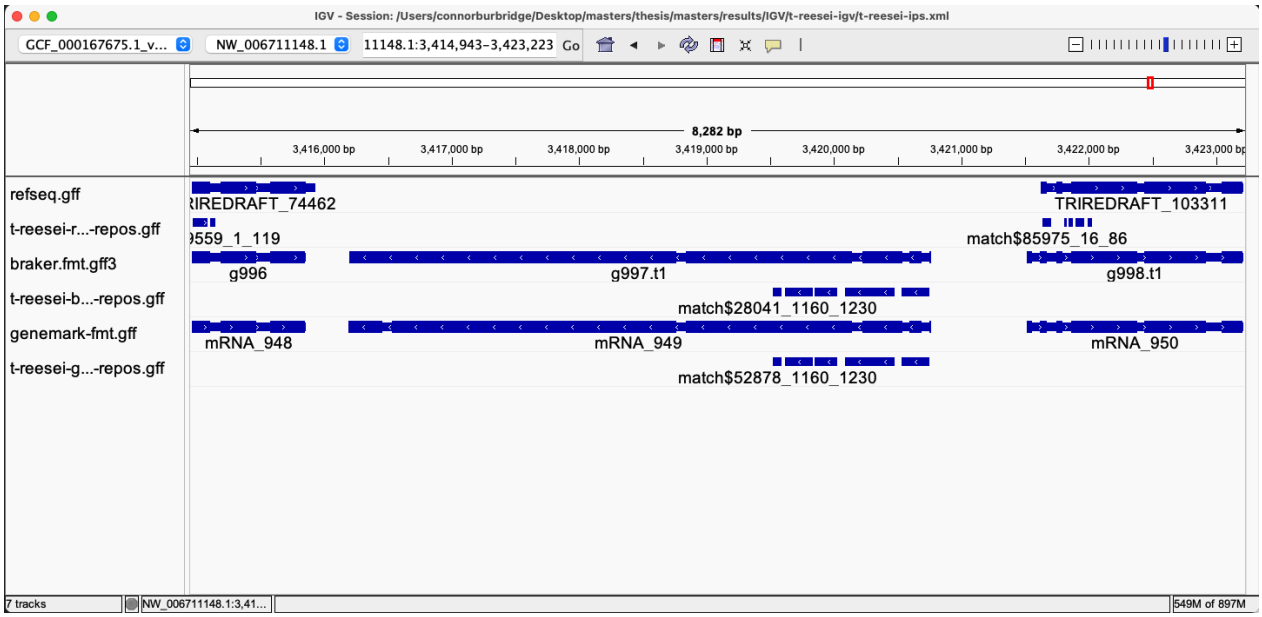


Figure 1.10: An IGV capture showing a scenario where GeneMark and Braker2 agree on a gene model with supporting Pfam evidence and RefSeq does not report any gene.

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

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 1.8: Total number of regions identified in each assembly followed by counts of regions (both agreement and singletons) in regions of 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.

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 1.7.

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. Table 1.9 shows the number of genes predicted by each gene finding tool in regions of AT-rich genomic sequence. GeneMark 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 1.10. 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

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 1.9: Number of genes predicted by Braker2, GeneMark and RefSeq in AT-rich genomic sequence from each assembly. It is important to note that in the cases of DC1 and Tsth20, full support indicates supporting gene predictions from both GeneMark and Braker2.

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

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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.