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

©Connor Burbridge, Dave Schneider, Tony Kusalik, All rights reserved.
Unless otherwise noted, copyright of the material in this thesis belongs to
the author.

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 Background

Chapter one provides a brief background on *Trichoderma* and gene prediction tools, providing context for this research. Section 1.2 discusses the evolutionary mechanisms that may have led to the increased production of secondary metabolites in *Trichoderma* species, such as horizontal gene transfer, gene duplication, and transposable elements.Section 1.3 discusses the novel *Trichoderma* genomes that were sequenced and assembled as part of this work, and how they relate to the research questions. Section 1.4 discusses the structure of eukaryotic genes how gene finders predict them, while section 1.5 provides an overview of secondary metabolites and their importance in *Trichoderma* species. Finally, section 1.6 provides an overview of comparative studies of gene prediction tools in *Trichoderma* species, and how this work aims to fill the gap in literature.

1.1 *Trichoderma* Species and their Features

Trichoderma are a genus of ascomycete fungi found in a wide variety of soils, and are well-known for their use in biomanufacturing of cellulases and hemicellulases as well as their roles as non-toxic, avirulent opportunistic plant symbionts [19] [10]. Many of these features *Trichoderma* can be attributed to the large number of secondary metabolites produced by these fungi, which are used to interact with plants and other organisms in the soil [14]. *Trichoderma* species differ in the number of secondary metabolites they produce, with some species producing more than others [14]. This variation in secondary metabolite production is of great interest to researchers, as it can provide insight into the evolutionary processes that have shaped *Trichoderma* species over time.

There are many mechanisms organisms may leverage that result in gain or loss of gene function, and ultimately the evolution of a species. One well-studied mechanism is that of horizontal gene transfer (HGT), which is the process of acquiring genetic material from another organism, rather than through inheritance from a parent organism [7]. HGT is a common mechanism in bacteria, but it has also been observed in fungi, including *Trichoderma* species [7]. Another mechanism of interest is the duplication of genes, which can result in the production of more than one copy of a gene, leading to increased expression of that gene [7]. This is particularly relevant in the case of secondary metabolite production, as many *Trichoderma* species have been shown to have multiple copies of genes involved in secondary metabolite production [14]. Lastly, transoposable elements (TEs) are another mechanism of interest, as they can insert themselves into the genome and disrupt or enhance the expression of genes [7]. TEs have been shown to play a role in the

evolution of *Trichoderma* species, particularly in the case of secondary metabolite production [7].

Interestingly, both HGT and TEs tend to present themselves in regions where GC content differs from the rest of the genome, [7]. Abnormal GC content is also associated with other genomic features, such as centromeres [15] and repetitive regions [18], both of which may have an effect on the production of secondary metabolites in *Trichoderma* species. As a result, it is important to consider the GC content of a genome when studying *Trichoderma* species, which forms the basis of the Research Question **??**. Some studies have shown that *Trichoderma* species contain very few if any transposable elements, which is unusual for fungi [9]. It is hypothesized that TEs were lost in *Trichoderma* due to repeat induced point mutations (RIP), which is a process that occurs in fungi where TEs are silenced and eventually lost from the genome [9]. Genes responsible for RIP mutations also tend to target CA dinucleotides, converting them to TA nucleotides, which explains the low GC content observed in some regions of *Trichoderma* genomes [7].

Given the proximity of *Trichoderma* species to other fungi, bacteria, and plants in soils, it is possible that *Trichoderma* species acquired genes which are involved in response to antagonistic and sympathetic intercellular interactions from other organisms. While difficult to prove, studies have shown that HGT events have occurred in a number of fungal species [6]. It is also important to note that some *Trichoderma* species have higher numbers of genes involved in secondary metabolite production than others, indicating an evolutionary divergence, making comparative analysis of *Trichoderma* species and strains a useful tool for understanding the mechanisms behind secondary metabolite production [14].

1.2 Evolutionary Landscape of *Trichoderma*

The evolutionary landscape of *Trichoderma* species is an important aspect of this work. DC1 and Tsth20 have not yet been classified into a specific *Trichoderma* species, and further understanding of their evolutionary relationships to other *Trichoderma* species may help explain how DC1 and Tsth20 originated. Figure 1.1 shows the phylogenetic relationships among selected *Trichoderma* species as well as the outgroup *Fusarium graminearum*. The tree was constructed using cursory alignments of the Rpb2 gene, which is a commonly used gene for phylogenetic analysis in fungi [1]. Representative Rpb2 proteins from the various fungal species were extracted from the NCBI resource for fungal orthologs, while representative Rpb2 proteins from DC1 and Tsth20 were identified from the results generated later on in this work. *Trichoderma* species separate into a number of distinct clades, and these clades can vary between studies as evidenced by the low confidence in placement of *T. atroviride* and *T. gamsii*. Lack of confidence in evolutionary relationships is not uncommon in *Trichoderma* species, and is likely due to the high levels of horizontal gene transfer and gene duplication that have occurred in these species [7]. This makes *Trichoderma* species interesting to study but difficult to classify.

Phylogeny is also important when selecting genomes for comparative analysis of bioinformatics tools. Some tools may vary in their performance depending on the nature of the genome they are applied to, making

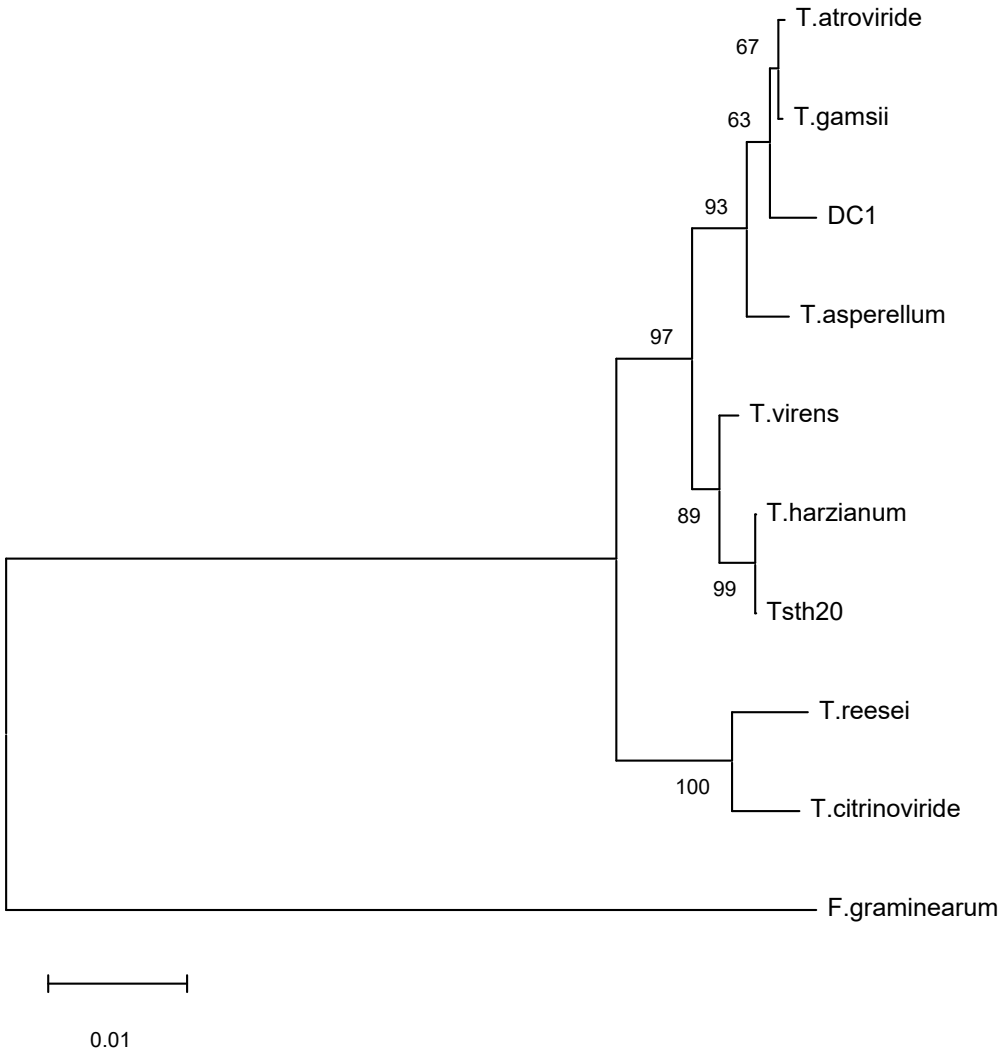

Figure 1.1: Phylogenetic tree illustrating relationships among selected *Trichoderma* species based on multiple sequence alignments of Rpb2. Evolutionary history was inferred using the Neighbour-Joining method [16]. The optimal tree with the sum of branch length = 0.146 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches [5]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree.

a heterogeneous selection of genomes important for understanding the performance of the tools in different contexts. In this work, we selected three well-studied *Trichoderma* species, *T. reesei, T. harzianum*, and *T. virens* for future comparisons, which are all commonly used in research and industry, but are relatively distant from each other evolutionarily according to more detailed studies [1].

1.3 Novel *Trichoderma* Genomes

Recently, two strains of *Trichoderma* have been identified in the prairie regions of Alberta and Saskatchewan. These two strains, named Tsth20 and DC1, have been found to have beneficial properties when used as an inoculant for plants in the soils. Tsth20 has been classified into the *Trichoderma harzianum* group, and may act as a bioremediation tool in soils contaminated with hydrocarbon content. DC1, while not yet classified, is believed to confer salt and drought tolerance to plants in dry, salty soils. While, bioremediation and resistance to drought tolerance has been investigated in other strains of *Trichoderma* [17], the study of new strains is still useful in the effort of discovering unique mechanisms used by *Trichoderma*. Little is known about the mechanisms at work in these strains, so DC1 and Tsth20 were sequenced at the Global Institute for Food Security in an initial attempt to better understand the mechanisms at work in these genomes. While this research does not explicitly identify genomic elements related the beneficial properties of these genomes, it may serve as a foundation for future research of *Trichoderma*. The assembly of these genomes come as a result of Research Question **??**.

1.4 Gene Predictions and Complementing Similarity Searches

Gene prediction methods are generally based on the idea that genes can be identified by searching for patterns in a sequence, which match an expected gene structure or model. These gene structures begin from a $5'$ start codon, continue through a series of exons and introns, and end with a $3'$ stop codon [11]. Flanking the gene structure are promoter regions, which are typically found upstream of the start codon, and untranslated regions (UTRs), which are found both upstream of the $5'$ start codon and downstream of the $3'$ stop codon. Promoter regions are important for the regulation of gene expression, while UTRs are important for the stability and translation of the mRNA transcribed by the gene [11]. Sequences are translated from the $5'$ start codon through to the $3'$ stop codon, splicing together exons and ignoring introns. An example of a eukaryotic gene structure is shown in Figure 1.2.

Gene prediction methods vary in their approach, but they generally fall into two categories: *ab initio* methods and evidence-based methods [4]. *Ab initio* methods rely on the identification of patterns in the sequence that match an expected gene structure, while evidence-based methods use prior information such as RNAseq data, expressed sequence tags (ESTs), and expressed protein sequences to identify genes within a new genome [4]. Gene prediction methods can also vary in their ability to identify different components of a gene, such as the promoter region, UTRs, and introns. Some methods may only identify the coding

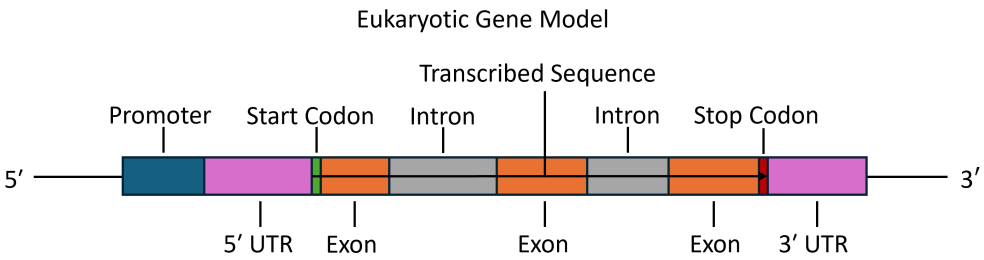

Figure 1.2: Example of a eukaryotic gene structure, showing a promoter region (blue), the 5′ and 3′ UTRs (purple), start codon (green), exons (orange), and introns (grey). The evrected arrow indicates the direction of transcription.

sequence (CDS) of a gene, while others may also identify the promoter region and UTRs [4]. Making things more complicated, different gene prediction tools may predict different combinations of exons and introns for a gene, and some tools may even predict multiple combinations of exons and introns, leading to multiple transcripts for the same gene. As a result, different transcripts may produce different proteins, or the same protein with different post-translational modifications, which can significantly impact the function of the resulting protein [4]. Given the already complex nature of secondary metabolites, it is important to consider the impact of variability in gene prediction methods, especially in the case of *Trichoderma* species and their secondary metabolite production. This consideration is one of the main motivators behind this research, and forms the basis of Research Question **??**.

While the basic structure of a gene may be present in a sequence, that does not mean that the gene codes for a functional protein, or in the case of secondary metabolites, a component that helps produce them. To assess potential function of a gene product, gene prediction methods are often used in conjunction with similarity searches, which compare the predicted genes to known genes in other organisms to identify potential functions, and serve as a form of validation for the predicted genes [11]. Similarity searches can come in several different forms, such as blast searches, which compare the predicted genes to a database of known proteins, or InterProScan, which compares the predicted genes to a database of known protein domains and binding motifs [11]. These similarity searches can provide additional information about the predicted genes, such as potential functions, and can also help to validate the predicted genes by comparing them to known genes in other organisms. Similarity searches can also be used to evaluate the completeness of the predicted genes. One example would be comparing predicted genes to a set of conserved single-copy orthologs expected in fungal genomes, such as those provided by BUSCO [12]. Together, these methods can provide a more complete picture of the gene predictions and their potential functions, and can help to identify potential targets for further research. This forms the basis of Research Questions **??**, **??**, and **??**.

1.5 Secondary Metabolites

While cellular products essential to an organism's viability are of great interest, there are a vast number of cellular products that while not essential, still provide great benefit to the organism and may be necessary for survival in some situations [3] [14]. These other cellular products are known as secondary metabolites, and they are involved in several roles ranging from cellular signalling to antibiotic activity, making them a frequent subject of study in pharmaceutical research. Enzymes that produce secondary metabolites are comprised of non-ribosomal peptide synthetases (NRPS) and polyketide synthetases (PKS), which synthesize amino acids and other basic enzymatic building blocks into proteins [8]. Genes encoding the NRPSs and PKSs responsible for production of secondary metabolites are often found in clusters within a genome, but their products remain unknown [14].

The evudy of these products is difficult, as the genes encoding the modules that make up NRPSs and PKSs are not expressed under normal laboratory conditions [14]. The evPSs and PKSs that have been studied have been shown to be large enzymatic structures containing several functional modules, each responsible for a specific step in the synthesis of a protein [14]. Studying these complicated mechanisms and their genes requires as close to a complete set of gene predictions as possible, as the absence of one gene encoding a module could affect the entire protein complex. This concept of completeness of gene predictions is explored in Research Question **??**. In addition, gene predictions can be processed further with tools such as antiSMASH, which can identify secondary metabolite biosynthetic gene clusters (BGCs) in a genome [2]. Comparison of predicted genes with known secondary metabolite BGCs can provide insight into completeness of the predicted genes with a specific focus on secondary metabolite production. This work is not covered in this thesis, but is worth noting for context as it is a common approach in the field of secondary metabolite research.

Another interesting feature of NRPSs and PKSs is the length of the genes encoding them, which can be quite long, often exceeding 10,000 base pairs in length [8]. Examining the outputs of gene prediction tools can provide insight into the lengths of predicted genes, and whether or not the tools are able to capture the full length of these large genes. This topic serves as the basis for Research Question **??**.

1.6 Gene Prediction Tools in Fungi

With the biological context laid out, we can now turn our attention to the computational aspect of this work. Many *Trichoderma* species have been assembled and annotated in the past, mostly for the purpose of understanding the mechanisms behind secondary metabolite production [14]. Gene predictions in this context are typically compared against sequences from other organisms to derive functional information, and to validate the predicted genes [11]. In contrast, very few studies have been conducted comparing gene prediction tools in fungal genomes, and even fewer have been conducted in *Trichoderma* species. This is surprising, given the large number of gene prediction tools available, and the fact that many of them are

used in the field of fungal genomics [4]. Comparisons of gene prediction tools are common when new tools are released, but these comparisons are often limited to a small number of tools, and tend to be focused on achieving higher counts of interesting genes, rather than comparing the results in a biological context [13]. Complicating the matter further, the quality of genome sequences available may vary, making evaluation of different gene prediction tools difficult when working with multiple genomes and lesser-studied fungal species with low quality assemblies. Given the gap in literature regarding comparative analysis of gene prediction tools in *Trichoderma* species, and the availability of the novel DC1 and Tsth20 genomes, this work aims to fill that gap by comparing the outputs of several gene prediction tools on these two genomes along with three other frequently studied *Tichoderma* species.

References

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

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

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

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

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

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

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

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

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

[10] 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.

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

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

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

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

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

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

[17] 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.

[18] 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.

[19] 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.