COMPARATIVE GENOMICS OF PATHOGENIC FUNGI THROUGH SEQUENCE HOMOLOGY AND PHYLOGENETIC SIMILARITIES

KENNETH TAN LEE SHEAN

FACULTY OF SCIENCE
UNIVERSITI MALAYA
KUALA LUMPUR

2023

COMPARATIVE GENOMICS OF PATHOGENIC FUNGI THROUGH SEQUENCE HOMOLOGY AND PHYLOGENETIC SIMILARITIES

KENNETH TAN LEE SHEAN

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (EXCEPT MATHEMATICS & SCIENCE PHILOSOPHY)

FACULTY OF SCIENCE
UNIVERSITI MALAYA
KUALA LUMPUR

2023

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Kenneth Tan Lee Shean

Matric No: 17026811/2

Name of Degree: Doctor of Philosophy (Physics/Biology/Chemistry/Geology)

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

COMPARATIVE GENOMICS OF PATHOGENIC FUNGI THROUGH SEQUENCE HOMOLOGY AND PHYLOGENETIC SIMILARITIES

Field of Study: Bioinformatics

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;

(2) This Work is original;

(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;

(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;

(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;

(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 23rd May 2023

Subscribed and solemnly declared before,

Witness's Signature Date: 12th May 2023

Name:

Designation:

COMPARATIVE GENOMICS OF PATHOGENIC FUNGI THROUGH SEQUENCE HOMOLOGY AND PHYLOGENETIC SIMILARITIES

ABSTRAK

Patogenesis kulat adalah salah satu isu ekologi dan perubatan yang paling kuat yang dihadapi oleh banyak saintis. Kemunculan penjujukan DNA telah membolehkan projek-projek penjujukan genom secara besar-besaran dari banyak kulat patogenik yang penting dan paling maut di dunia, gandingan dengan analisis bioinformatik hulu yang merangkumi pemasangan genom dan anotasi genom yang menghasilkan data tersedia secara terbuka untuk penyelidikan bioinformatik gunaan. Kajian ini melibatkan pembinaan pangkalan data gen yang berkaitan dengan Fungal Pathogenicity dengan 5,183 urutan protein dari pangkalan PHI, 921,174 urutan protein dari Database EnzYme Carbohydrate-Active, dan 2,058 urutan protein dari Database Factors Virulence di Fungal Pathogens. Pangkalan data tempatan dicipta menggunakan makeblastdb dalam aplikasi NCBI-BLAST + dan pencarian homologi urutan protein 86 spesies jamur telah dijalankan dengan BLASTP mengakibatkan pengenalpastian potensi gen patogenik yang sama antara kulat dalam kajian, ianya juga untuk yang berpotensi dan memahami hubungan filogenetik. Pangkalan data boleh digunakan sebagai aplikasi agregat untuk anotasi gen patogen kulat yang menyumbang kepada komuniti penyelidikan yang lebih luas.

Kata kunci: Bioinformatik, Perbandingan Genomik, Patogenik, Kulat, Kesamaan

COMPARATIVE GENOMICS OF PATHOGENIC FUNGI THROUGH SEQUENCE HOMOLOGY AND PHYLOGENETIC SIMILARITIES

ABSTRACT

Fungal pathogenicity is one of the most vigorously tackled ecological and medicinal issues facing many scientists. The emergence of DNA sequencing had allowed massive genome sequencing projects of many important and most fatal pathogenic fungi in the world, coupling with upstream bioinformatics analysis which includes genome assembly and genome annotation had resulted in publicly available datasets that can be utilized for applied bioinformatics research. This study involves building of an aggregate Fungal Pathogenicity-related gene database with 5,183 protein sequences from PHI-base, 921,174 protein sequences from Carbohydrate-Active enZYme Database, and 2,058 protein sequences from Database of Virulence Factors in Fungal Pathogens. Local database was created using makeblastdb within NCBI-BLAST+ application and homology search of protein sequences of 86 fungal species was carried out with BLASTP resulting in identification of potential common pathogenic genes between fungus in study, also to identify potential biomarkers and understanding phylogenetic relationships of pathogenic fungi. The database can be utilized as an aggregated application for fungal pathogenic genes annotation that contributes to a wider research community.

Keywords: Bioinformatics, Comparative Genomics, Pathogenic, Fungus, Homology

ACKNOWLEDGEMENT

Special thanks to Professor Madya Dr Saharuddin bin Mohamad for his guidance throughout the entire course of the research project. His inputs were extremely valuable for the success of this study and to the completion of it.

I would also like to take this chance to thank my wonderful wife Lim Lea Yee for supporting my work and allowing me to spend plenty of time on research while she cares for our daughter. Without her support and participation, I would not be able to complete all required work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

BLAST	:	Basic Local Alignment Sequence Tool
CAZy	:	Carbohydrate-Active Enzyme Database
CMS	:	Content Management System
DFVF	:	Database of Fungus Virulence Factors
FGI	:	Fungal Genome Initiative
NCBI	:	National Center for Biotechnology Information
PHI-base	:	Pathogen-Host Interaction Database
SNP	:	Single Nucleotide Polymorphism
CFPG	:	Common Fungal Pathogenic Genes
HGT	:	Horizontal Gene Transfer

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Fungal pathogenicity remains as one of the main challenges in modern science, with more than 300 species of 1.5 million fungal species causing disease across the animal and plant kingdom hence there are still a lot of work to be done to tackle the issue. The two main aims in studying fungal pathogenicity are to tackle two key issues with diseases: diagnosis and treatment. As with any other form of disease early diagnosis will ensure higher success rate of recovery for both animal host or plant host, or in the cases of fungus inflicted plant diseases which are difficult to diagnosed due to the lack of visible symptoms hence early diagnosis will allow development of counter-acting strategy. Once diseases are identified, treatments can then be formulated and applied to animals or plants. Developing fungicides is challenging and needs to hit the right marks with broad-spectrum effectiveness, enhanced bioavailability, and minimal toxicity and side effects (Brauer, et al. 2019). With the genome plasticity that fungus has (Fisher, et al. 2018) it becomes challenging as fungus quickly reproduce and can rapidly develop resistance to antifungal drugs and render the antifungal agents useless.

Recent advancement in genome sequencing technologies, bioinformatics tools and applications, comparative genomics platforms had proven to be a gateway to new research initiatives to blossom. Research initiatives such as the Genome 10K Project (Koepfli, et al. 2015), aimed at sequencing genome of at least one individual from every vertebrate genus which accounts to approximately 10,000 genomes. Rapid development of sequencing technologies that produce more genome sequences at much lower cost more high-throughput sequencing projects targeted at comparative genomics study, examples of such study approach is the fishes of Genome 10K (Bernardi, 2012) and Fungal Genome Initiative (Broad Institute, 2014), each targeting of accelerating research on microbial metabolism, physiology, and functional genomics and studying human/plant pathogens as basis for molecular and cellular biology. Research projects as such targeting to understand how fungus genomic makeup affects its life cycle and in turn plays vital role in the study of fungal pathogenicity which can prove vital in tackling issues around fungal pathogenicity to allow developing effective diagnostic methods and uncovering effective antifungal agents that can be used to treat fungal diseases across human, animal, and plant.

The gold standard of identification for fungal diagnosis is through culture and microscopy observation (Kidd, et al. 2020) but not without its limitations from slow culture growth and highly dependent on the specimen containing viable fungal elements. This will continue to be a challenge to an organism with plastic genome and has rapid evolution life cycle where new disease resistance could arise. Polymerase Chain Reactions (PCR) assays proves to be a great alternative as diagnosis can be made from specimen samples including blood. PCR assays leverages on the specificity of fungal DNA primers that would amplify the target regions and base on that identifying the causative fungus. This is extremely important for early diagnosis of soil fungus that causes diseases to specific host plants as a large number of microbes lives in soil. The identify of fungal species that constitute the soil samples can then be uncovered through designing specific DNA primers and sequencing or through metagenomics sequencing (Donovan, et al. 2018) using shotgun metagenomes.

High throughput sequencers generate massive amount of sequence data, be it nucleotide sequences or protein sequences hence with these data available it now allows researchers to leverage on datasets for applied studies and creating secondary databases that can continue to push the front of developing novel diagnostic and treatment methodology. Given how rapid the fungal species evolved continuous genomics analysis and comparative genomics effort is required to keep up with the pace of fungal genome evolution.

This study aimed to understand underlining genomics commonality between pathogenic fungi through various comparative genomics techniques and applications, generating *in silico* results and datasets that allow identification of candidate common pathogenic genes among pathogenic fungi to propose new candidate regions for pathogenic fungi identification and in turn as a foundation for antifungal agent development.

Comparative genomics is a common and known technique in studying diversity and phylogenetic relationships in the study of fungus diversity. There are a plethora of comparative fungus genomics research study and resources available such as FungiDB (Basenko, et al. 2018) that provides a platform to further annotation of fungus genomics sequences. Usual studies compare different isolates of fungus within a species of fungi and rarely studies fungus across phylum. There is value in understanding inter-phyla relationship to aids understanding of conservation and diversity in the kingdom of fungus in particularly when looking at the topic of fungal pathogenicity.

Availability of genomics sequences in the public domain provides an opportunity to perform applied bioinformatics research to unveil useful insights on the available databases and datasets. According to statistics by GenBank published sequences have now exceeded billions in whole genome sequence data (National Center for Biotechnology Information, 2021), providing an enormous amount of publicly available sequence to be studied. Existing fungal pathogenicity-related databases also allows for applied study leveraging on known experiment results to further dive into the details and understanding of fungal pathogenicity.

1.2 Study Design

1.2.1 Developing a Fungus Comparative Genomics Pipeline specifically for Inter-Phyla Pathogenicity study

Fungus comparative genomics is a commonly known technique to understand fungus diversity and the intricate relationship between fungus diversity and pathogenicity. Throughout the internet there are many resources and web applications that provides user interface for sequence comparison and annotation such as FungiDB (Basenko, et al. 2018) and Carbohydrate-Active enZYme (CAZy) (Lombard, et al. 2014). These platforms are extremely useful for comparative studies of small number of fungal sequences and has limitation is handling specific queries to the database to fulfil a specific study objective. Existing fungal pathogenicity-related database presents repository of fungal pathogenicity-related genes from experimental data like the Pathogen-Host Interaction Database (PHI-base) (Winnenburg, et al. 2006) while others like Database of Fungal Virulence Factor (DFVF) (Lu, et al. 2012) was built using *de novo* text mining methods.

These databases provide avenue for further study to be done based on the data that are available. Due to the characteristic of the fungal genome, pipeline specifically for study of fungal pathogenicity is highly desirable. Existing databases and tools provide tools that are catering for broader comparative genomics effort, example like FungiDB hence there is a knowledge gap to utilize and leverage on existing databases, enabling new findings especially in understanding if diversity plays an important role in fungal pathogenicity.

1.2.2 Discovery of Common Fungal Pathogenicity-related Genes across Fungus Phylum and Species

Through comparative fungus genomics analysis using the established pipeline the study aims to discovery high confidence Common Fungal Pathogenicity-related Genes across fungal species that constitute different phyla within the kingdom of fungus. This will contribute to the continuous understanding of fungus diversity and infer relationships between species, and with the identification of common pathogenicity-genes it would allow the scientific community to infer relationships between fungal species in the context of pathogenicity which could determine a pattern of conserved pathogenicity. Fungus comparative genomics are usually performed between different isolates of the same species or different species in the same genus or phylum as such done on Beetle-Vectored Fungi (Schuelke, et al. 2017) but few had taken a broader view of the subject.

Conservation of pathogenicity across the kingdom of fungus will shed more lights into the conserved mechanisms that lies within the fungus lifecycle and provides a platform for further development of fungal pathogenicity diagnostic methodology that target conserved region within fungal genome and design broad-spectrum antifungal agent that would serve as a treatment for infected hosts.

1.2.3 Creating a Common Fungal Pathogenicity-related Gene Database Portal

This study aims to leverage on available public resources and through comparative genomics analytical methodology to create a Common Fungal Pathogenicity-related Gene Database. The identification of the common pathogenicity-related across multiple fungal phyla and species will be rendered meaningless without providing a public portal to allow access to the data which can enable the scientific community to continue building on the discovery. Data from this study will be made available and accessible to the community through a web portal that allows downloading of the discovered candidate common pathogenicity genes.

A publicly available database portal will also enable further study on the subject by the broader scientific community which allows collaborative effort in understanding and tackling the global issues with fungal pathogenicity be it in human and animal diseases or plant diseases that affects plant of great agricultural importance.

1.3 Objectives

The main aim of this study was to investigate genomic diversity and relationship between pathogenic fungi across the kingdom.

In order to achieve the aim of the study, a number of specific objectives were defined:

1. To evaluate diversity and relationship of pathogenic fungi through comparative genomics.
2. To identify Common Fungal Pathogenicity-related Genes across the Kingdom of Fungi.
3. To develop a comparative genomics pipeline specifically for fungal pathogenicity.
4. To create a portal for public access of data.

1.4 Thesis Organization

This thesis contains six chapters which includes introduction, literature review, methodology, results, discussion, and conclusion. The first chapter introduction described the overview, study design and the objectives of the study and is followed by chapter two that consists of literature review of topics related to the study includes fungal pathogenicity, study of diversity using next generation sequencing and techniques, and various challenges in diagnostic and treatment of fungal pathogenicity. Chapter three describes and explains methodology used in this study, and chapter four presents results from the study and are structured in three parts where the first part contains results from homology searches from protein sequences, part two presents the downstream analysis of the results and finally part three presents the database portal. Chapter five discussed all the findings, and the last chapter summarizes and provide a conclusion for the study.

CHAPTER 2

LITERATURE REVIEW

2.1 History and Background of Fungal Pathogenicity

Fungi are a group of organisms that had appeared in this world from an ancient history and fossil evidence had showed that fungi may had been around the world since 460 to 455 million years ago (Carris, et al. 2012). Further evidence had also proposed important role that fungi played in colonization of the land by earlier plants (Carris, et al. 2012). Fungi play an extremely important role in maintaining ecological balance exhibiting its saprophytic nature allowing it to decompose organic materials – a crucial step in the utmost important carbon cycle.

Fungi had presented itself to be one of the most diverse and largest kingdoms with approximately 1.5-5 million species of fungal species identified to date (Blackwell, 2011). Fungi had showed capability to survive and thrive in the most adverse condition in the world having been found in all temperature zone (Jaejin & Sung-Hou, 2017) making it an extremely robust organism that can colonize any location with very limited resources. Other than being found on hard surfaces fungi can also be found growing on other living organisms and this includes both animals and plants. The ability of a fungus to grow on animals' posts benefits to the ecosystem but bares extreme devastation for the host. One classic example of that is the entomopathogens, which belongs in the Ascomycota genus *Ophiocordyceps* in which these fungi had demonstrated that it can infect and consume insects like caterpillars and ants (Carris, et al. 2012). The extend of infection can causes a change of insect's behaviour, such with the case of the "Zombie-ant" fungi found in Brazil. These fungi can infect the brain of the insects causing drastic change of behaviour directing the victim of infection to climb up to plants and bite into the plant tissue in a manner known as “death grip” (Sanusi, et al. 2016).

About 300 of 1.5 million different species (Hawksworth, 2001) of fungi on earth are known to cause diseases in human (Garcia-Solache & Casadevall, 2010) and in plants agricultural important crops, the effects of fungi inflicted plant diseases cause massive destruction of important crops. Each year fungal infection destroys approximately 125 million tons of world top five food crops: rice, wheat, maize, potatoes, and soybean (Fisher, et al. 2012) and causes loss of billions of dollars in agriculture industry. One example of such devastating impact caused by fungus is the Rice Blast, which is caused by an ascomycete fungus *Magnaporthe oryzae* (Dean, et al. 2005). Study of fungal pathogenicity in plants is vital for eradication of plant fungal infections with then could prevent massive destruction of crops, which is key for the survival of human race.

These pathogenic fungi have been widely studied for their role in diseases and are known to originate from two major phyla in the kingdom of fungi, namely Basidiomycota and Ascomycota. Members of these two major phyla had collectively contributed to numerous plant diseases, infecting wide range of plants including several important staple food stocks for human population such as maize, wheat, rice, potatoes etc. Many causative factors could contribute to pathogenicity in fungi thus discovery and identification of causative factors among different pathogenic fungal species is important.

2.1.1 Fungal Pathogenicity in Plant

Plant pathogenic fungi relies on its life cycle for effective colonization of the host plant (Rodriguez-Moreno, et al. 2018). Fungus pathogens encompasses all types, depending on the different nature of fungus that can include necrotrophic fungus, hemi biotrophic, biotrophic or obligately biotrophic fungus. The differences in life cycle between these fungi, however, are surprisingly negligible when it comes to pathogenicity as pathogenic fungi are known to use well-conserved mechanism in the process of infecting and colonizing the host. This was described in a study to establish a standardized Gene Ontology terms among plant pathogenic fungi (Meng, et al. 2009). One of the most well studied molecular pathway in pathogenic fungi is the cAMP/PKA and MAPK pathways in different fungi. This group of proteins that are related to these pathways is known as the signalling proteins are found to be highly conserved in the fungal pathogenicity evolution and plays an important role the onset of pathogenicity in hosts (Turrà, et al. 2014).

2.1.2 Notable Plant Pathogenic Fungi

Certain pathogenic fungi is known to be extremely damaging to the host plants, which often than not coincides with being important food crops. Diseases caused by such fungus damages not only the environment but also livelihood of people that relies on these food crops either on sales or by the consumption of it. Dean et al. (2005) described a survey that involved fungal pathologist to determine a list of top ten fungal plant pathogens. The Top 10 list of fungal pathogens is listed in Table 2.1.

Table 2.1: Top 10 Plant Fungal Pathogen

Rank	Fungal pathogen	Phylum	Rank	Fungal pathogen	Phylum
1	*Magnaporthe oryzae*	Ascomycota	6	*Blumeria graminis*	Ascomycota
2	*Botrytis cinerea*	Ascomycota	7	*Mycoshaerella graminicola*	Ascomycota
3	*Puccinia spp.*	Basidiomycota	8	*Colletotrichum spp.*	Ascomycota
4	*Fusarium graminearum*	Ascomycota	9	*Usitlago maydis*	Basidiomycota
5	*Fusarium oxysporum*	Ascomycota	10	*Melampsora lini*	Basidiomycota

Plant pathogenic fungi are mostly constituted of members from the phylum Ascomycota and Basidiomycota. From Table 2.1 it showed that of the Top 10 listed plant pathogenic fungi 3 of the fungus in the list is from the Basidiomycota phylum and the rest from the Ascomycota phylum. Most of these fungi causes devastating impact in different plants which causes ripple effects to the economy. *Magnaporthe oryzae* causes rice blast disease which causes damage and losses in rice production around the world (Ou, 1980). Fungus infestation in host plant can be difficult to detect in early stages as certain fungus can remain dormant until triggered by specific environmental cues. *Botrytis cinerea* is known as one of the most destructive fungus due to its broad host ranges, infecting plant species ranging from fruits, vegetables to ornamental flowers (Plesken, et al. 2015). The fungus causes grey mold rot to its host plants at any timepoint of growth of the host plant, from seedling stage to product ripening and will continue to be a threat during transportation.

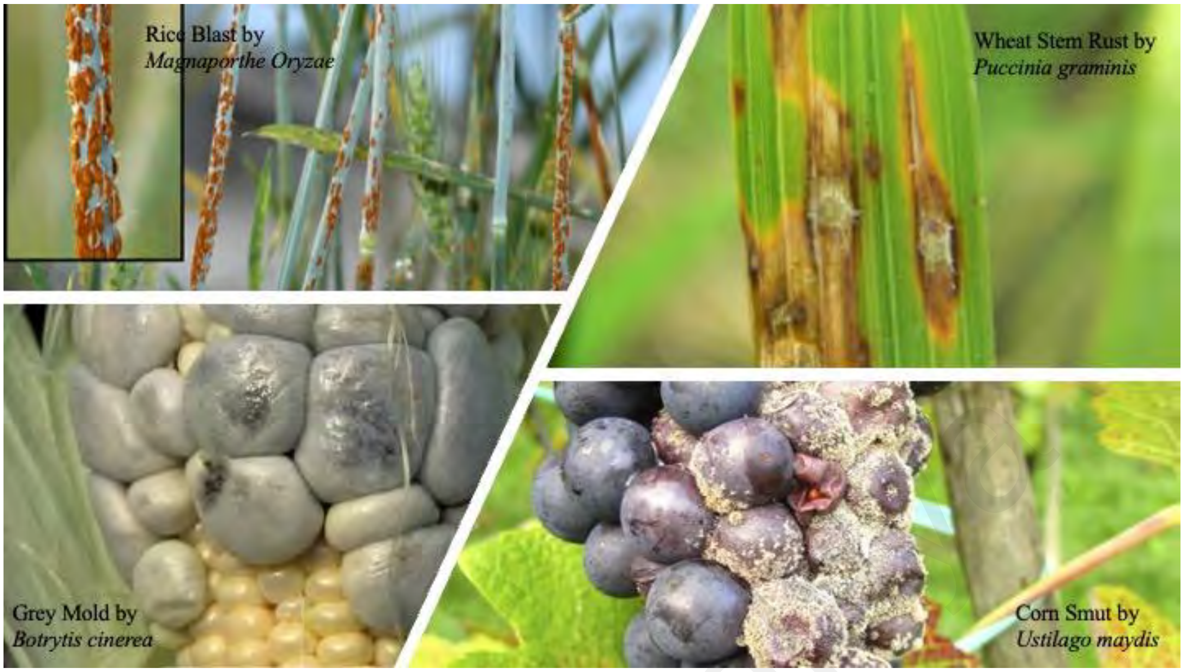


Figure 2.1: Diseases caused by plant pathogenic fungi

Ability to remain dormant often make early detection of fungal diseases in plants extremely difficult, increasing the risk of host plant destruction as often when symptoms for fungal diseases are visible it is usually too late to reverse the impact of the diseases. Classic example is the Basal Stem Rot disease in oil palm (*Elaeis guineensis*) by *Ganoderma boninense*, which causes loss in oil palm production, impacting produces roughly US$500 million every year (Ahmadi, et al. 2017) and widely known as the most destructive disease affecting plantations in Southeast Asia – a region that produces majority of the world palm oil production. Detection of the Basal Stem Rot disease in oil palm is extremely difficult due to the absence of disease symptoms in early stage of infestation, and only showed symptoms of infection at the critical stage of plant growth thus making disease management in oil palm extremely challenging and difficult thus the key in disease control is early detection of the diseases before it is too late.

Animal pathogenic fungi including members from the genus of *Aspergillus*, *Rhizopus*, *Mucor*, *Candida*, *Cryptococcus* and more causes the onset of diseases in human and animals. These fungi causes allergic reactions (Seyedmousavi, et al. 2015) that causes respiratory infection while other member of the genus *Aspergillus* that has the ability to produce mycotoxins has found to cause stonebrood disease in honeybee (Bailey, L. 1963). Pathogenic fungus that attacks animal hosts including human causes serious illnesses and estimated to kill approximately 1.5 million per year (Brown, et al. 2012). This is an alarming figure that often goes under the radar compared to other pathogens such as viruses and bacteria, which gives rise to the questions if there are more knowledge to be uncovered by the scientific community through comparative studies of fungus from a wider spectrum of characteristics, in other words looking at fungal pathogenicity as a whole rather than at a specific fungal species, or phyla. Similar to plant pathogenic fungi, fungus that infects animals and human hosts relies on the life cycle for colonization of the host. In *Candida albicans* for instance Ras/cAMP/PKA (Hogan & Sundstrom, 2009) plays a very important role in its pathogenicity for the involvement in morphogenesis, virulence, and opaque switching (Lin & Chen, 2018). This is a great example of the conservation of pathogenicity mechanism in the kingdom of fungus where similar molecular mechanisms are identified between fungal pathogens that infects different ranges of host.

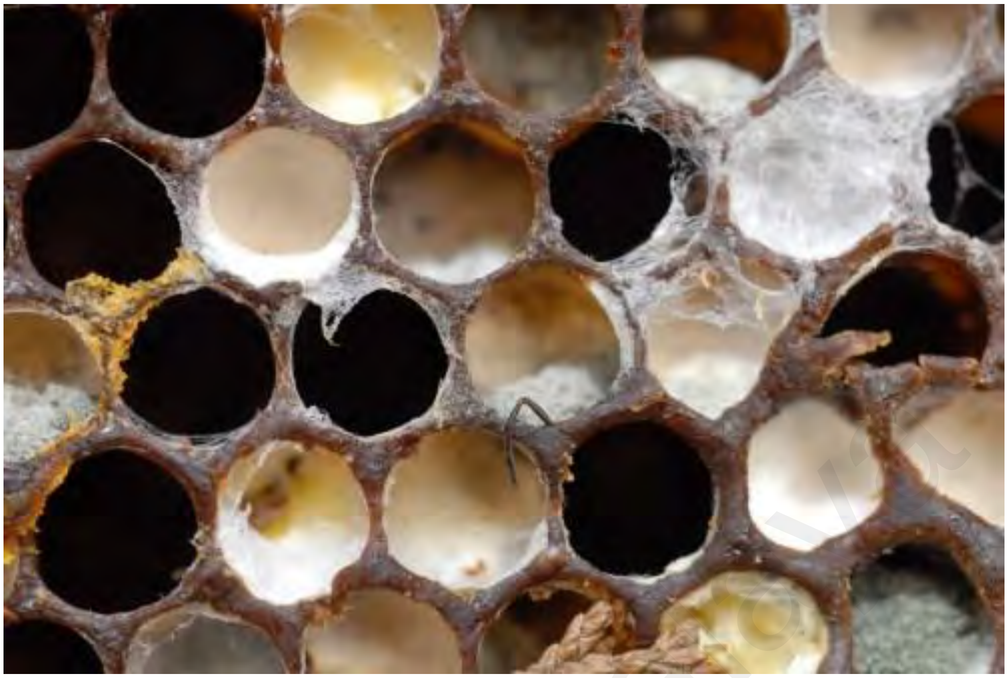


Figure 2.2: Stonebrood disease in honeybee (Scientific Beekeping, 2023)

Advancement in DNA sequencing technology had allowed researchers to develop molecular diagnostic tools that provides more accurate results than conventional diagnostic method. Various methods of detection for *Ganoderma boninense* for instance provides different level of detection of the fungus in oil palm. The earliest molecular detection method was an immunoassay test by using the binding of antibodies to the fungus (Reddy & Ananthanarayanan, 1984) but it was not the most effective method of detection due to the lack of information of taxonomy and inaccurate identification of different species of the genus. In the early 2000s ELISA (Enzyme-Linked Immunosorbent Assay) showed good detection results but has its own flaws as binding of the antibodies was not species specific (Utomo & Niepold, 2001). PCR or known as the Polymerase Chain Reaction is a great tool for amplification of specific regions of a DNA sequence and the development of PCR Primers that will only bind and amplify certain region of genomic sequence becomes a very useful way to identify organisms. Because of the ability to amplify a specific genomic region and leveraging of the general conservation of sequences among species in the family of *Ganoderma* the technique was also utilized to study differences in sequences between pathogenic and non-pathogenic *Ganoderma* spp and had been proven to be more accurate and less prone to contamination (Utomo, et al. 2005). This capability allows development of species-specific Primer sequences which will assist in detection of specific fungal species thus helping early detection and allows disease control to take place much early in the disease timeline.

The key ingredients for PCR experiments are DNA Primer and the DNA template where the primer would bind to and where amplification of DNA sequences take place. Depending on the objective of the PCR experiment different PCR primers that are developed to amplify specific regions, from identifying species of organisms in a sample that contains cocktails of organisms as well as discovering presence absence of protein-coding genes in knockout gene studies. In fungus studies the Internal Transcribed Spacer (ITS) regions of fungal ribosomal DNA is an important region of fungal genomic sequence that are highly conserved yet contains genetic variations that made identification and differentiation of fungal species via PCR experiments (Martin & Rygiewiez, 2005).

PCR analysis has been proven effective in detecting fungal pathogens in human. Study by Ferrer et al. (2001) showed successful application of the technique where the study showed positive identification in all patient cases and control samples as expected were PCR negative. Difference between detection of human fungal pathogens and plant fungal pathogens is the challenges that comes with DNA extraction for fungal plant pathogen studies which more than often has to deal with environmental samples that consists of multitude of organisms, posing challenges for a clean amplification and identification of fungus DNA from those samples. Despite the challenges faced, researchers have been using the technique to detect presence of fungus in environmental samples such as seed. Because of the specificity to the host and the sensitivity to be able to amplify the lowest available amount of DNA it is extremely useful for detection or identification of the organisms in samples (Walcott, 2003).

2.2 Mechanism of Plant Fungus Pathogenesis

Plant pathogenic fungi relies on different mechanisms to assist them in the process of pathogenesis, the initiation of pathogenicity in host plants. Signalling proteins are known as one of the group of proteins that help in this process and one such example are the MAP kinases. MAP kinases had been discovered in several fungal pathogens and play an important role for appressorium formation, invasive hyphal growth, and fungal pathogenesis (Xu, 2002). Study had also confirmed its role in pathogenesis when mutants disrupted of the Slt2 homologues demonstrated and possess weaker cell walls.

2.2.1 Signalling Proteins

Signalling proteins is vital in host-pathogen interaction in the early stages of infection (Tudzynski, et al. 2003) in reception of extracellular signals from the host to pathogens to activate effector proteins for initiation of infection into the host. Example of such gene is the heterotrimeric G proteins where the G proteins activate other effector proteins such as kinases, adenylate cyclases, phospholipases and ion channels (Kronstadt, 1997) and this includes the MAPK gene. Receptor proteins recognize surface protein of the host and initiates infection mechanisms towards the host. GTP-biding proteins is another candidate gene responsible for fungal pathogens' pathogenicity where research had shown that absence of these proteins results in reduced growth rate and morphological changes. Furthermore GTP-binding protein is connected to MAP kinases cascades for cAMP pathway that triggers the development of appressorium formation (Tudzynski, et al. 2001).

Pathogenic fungi develop different infection mechanism depending on the type of host that they are infecting or colonizing, some develop specialized infection structures in order to penetrate the tough protective mechanism of the host organism and in most plants that would be the plant cell well which is made up of large biopolymers cellulose, hemicellulose, lignin and pectin. One example of such specialized structure is the appressoria, which is formed by many pathogenic fungi during pathogenesis to penetrate plant primary defence mechanism to allow infection of the host plant. Peroxisomes are secreted to facilitate virulence proteins, in *Magnaporthe oryzae* (Chen, et al. 2016) peroxisomes proliferate that facilitates β-oxidation which is known to be an important step in pathogenesis.

2.2.2 Carbohydrate-active Enzymes

Carbohydrate-active enzymes, or more famously known as CAZymes are a group of enzymes that are involved in the metabolism of glycoconjugate, oligosaccharides, and polysaccharides (Zerillo, et al. 2013). The presence of this group of enzymes in pathogenic fungi ensure successful penetration through the host plant cell wall as it serves as a catalyst in the process of the degradation of the plant cell wall.

Smut Fungus, or scientifically known as *Ustilago maydis* secretes a set of lignocellulose-degrading enzymes that are capable to breakdown plant cell walls, compromising the plant primary defence before colonizing the host plant.

2.3 Genomics Study of Pathogenic Fungus

Sequencing technologies serves as an enabling platform for various downstream research and development, particularly setting the foundation for bioinformatics research and development. Discovery of different polymorphic markers such as Single Nucleotide Polymorphism, Insertions and Deletions, Copy Number Variations as well as presence of genes is important as each of these polymorphisms plays important roles in causing pathogenicity in fungus which could confers pathogenicity to pathogenic isolates as it is shown in human research.

Various massive sequencing projects around the world provided enormous genomic resources for the study of fungal pathogenicity. From generic resources such as GenBank (Benson, et al. 2018), DDBJ (Fukuda, et al. 2021), and EMBL (Hingamp, et al. 1999) to databases with a focus on such as the Fungal Genome Initiatives by Broad Institute (Broad Institute, 2014), FungiDB (Basenko, et al. 2018), EnsemblFungi (Howe, et al. 2021), to name a few. Most of these fungal genome databases serves as a huge repository for fungal genome databases. GenBank, DDBJ, and EMBL are all universal repository for all types of sequence data including raw sequencing data, whole genome assemblies, gene annotations, protein sequences, variant calls and etc. These data cover all organisms, including various species of fungi across the Kingdom of Fungi. While undertaking bioinformatics analysis of pathogenic fungi this becomes extremely challenging as it requires enormous effort in data clean-up to obtain the datasets of interest for study, which creates a gap to be filled by specialized databases or repository.

FungiDB contains 220 fungal genome sequences for species of fungi that are associated with infectious diseases with mammalian hosts and invertebrate vector of disease (Basenko, et al. 2018). Other than containing fungus sequence data, FungiDB is also an integrated platform for data mining and functional genomics analysis. FungiDB provides online bioinformatics tools to allow homology study using BLAST tools (Camacho, et al. 2009), allowing downstream analysis in comparative genomics effort in various studies such as those performed on *Aspergillus fumigatus* (Guirao-Abad, et al. 2021) and *Cryptococcus* isolates (Yu, et al. 2021). FungiDB Enrichment Analysis in FungiDB allows GO annotations of the studies and contains many other tools that provide convenience for downstream analysis of fungus genomics study. Publicly available fungus genomics data can help accelerate in silico research for bioinformatics community to uncover various insights without needing to perform genome or DNA sequencing projects hence reducing the time to discovery. Fungal Genome Initiative by Broad Institute was launched in November 2000 anchored by a group of fungal geneticists and biologists with the belief that the limitation to speed of discovery in biomedical research was caused by minimal publicly available fungal genome data (Broad Institute, 2014). Since then, the initiatives focused its effort in species of fungi that are important in human health and commercial activities (i.e. agriculture) and its value for fungal diversity and comparative genomics.

Publicly available fungal genomics data is a valuable starting point for downstream analysis, in particular for comparative genomics studies. With available annotation data including genes, proteins, exons, transcripts sequences it allows for secondary databases to be created based on data in primary databases. The Pathogenic Host Interaction Database, PHI-base is a specialized database focuses on catalogues experimentally verified pathogenicity, virulence and effector genes from fungal, oomycete and bacterial pathogens (Urban, et al. 2017). The database is extremely powerful as it provides validated experimental data on genes that participate directly and influence pathogenicity of fungus within a host-pathogen interactions. The database is used extensively in various genomics studies of pathogenic fungi in comparative genomics studies and pathogenic genes annotation and searches through BLAST. The database has been used to annotate pathogenic genes in *Ganoderma boninense* (Ramzi, et al. 2019), allowing identification of genes that participate in virulence of *Ganoderma boninense* in oil palm. It has also been used to predict virulence determinants in draft genomes of *Apophysomyces variabilis* where the species are prevalent causative agents of mucormycosis in India (Prakash, et al. 2021). The most recent PHI-base release 4.11 contains 8,411 genes sequences which are found in 18,190 interactions. These entries are available for public download for local usage of the data which provide opportunity to build fungal pathogenic genes annotation pipeline that can quickly predict presence of candidate pathogenic genes in new genome sequence projects.

Fungal pathogenicity in plants has specific mechanisms to challenge the rigid plant cell wall while undergoing proliferations. Fungus generates enzymes that can penetrate the rigid plant cell wall and this group of enzymes are known as the Carbohydrate-Active Enzymes. CAZy, or known as Carbohydrate-Active enZYmes Database (Lombard, et al. 2014) or known more popularly in its acronym CAZY is a database that contains protein sequences of structurally-related catalytic and carbohydrate-binding modules that are known to have different modes of interactions with glycosidic bonds, a very important linkage and type of covalent bond that joins carbohydrate molecule to another group. Glycosidic bonds are fundamental linkages found in cellular walls (Joseleau & Perez, 2016) thus are prime target of Carbohydrate-Active enzymes and thus Carbohydrate-Active enzymes are considered as candidate fungal pathogenic genes because of the capability to degrades the plant cell wall and these enzymes classes and associated modules are involved in various biological pathway of the host organism.

Massive sequence data and literature published on fungal pathogenicity also allow opportunity to create a database based on these published experimental data. The Database of Virulence Factors in Fungal Pathogens (DFVF) (Lu, et al. 2012) was a project aimed at filling the missing gaps in understanding of fungal pathogenicity by aggregating all known virulence factors also developing an algorithm that allows prediction of potential candidate genes that will be contributing to development of fungal pathogenicity. The database was built by leveraging of text-mining technique pursued by PubMed database and the Internet by looking for fungal disease virulence keywords and in-house tools were developed to allow searching of relevant supporting literatures. With this methodology the database currently contains 2058 protein sequences.

2.3.1 Inter-Phyla Comparison and Host-Independent Comparison

The similarities between pathogenic fungi that attacks plant and animal hosts are unsurprisingly high. Both groups of fungus are similar in the mechanisms of pathogenicity which are all as part of the fungus life cycle from spore germination, invasion via physical openings, colonization and alteration of host, reproduction, and transmission. These similarities in the pathogenicity mechanism prompted the interest in studying these fungi not as a separate group but as a same study group which allow further understanding in pathogenic mechanism in the Kingdom of Fungi.

From a different perspective at looking to compare between pathogenic fungi that infects plant host and animal host genomic identification provides a mean of understanding adaptation of these species of fungi based on host-specificity. Fungal species that infects plant hosts can have broad or narrow host of ranges (Sexton & Howlett, 2006) and specificity is defined by R genes, or known as resistance genes in the host and the virulence factors found in the pathogenic fungi (van der Does & Rep, 2007). The range of host that a fungus can infect does not limit to just plant or animals, some extreme examples like within the genus *Fusarium* causes disease across plant species and animals including human (Sharon & Shlezinger, 2013), which makes understanding the mechanism behind pathogenicity even more peculiar.

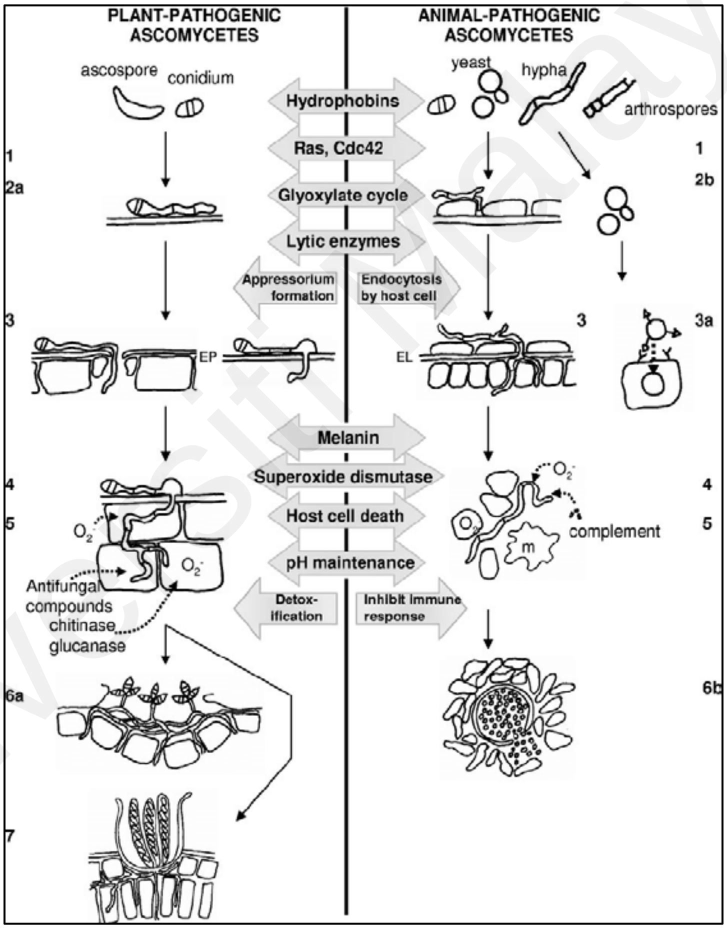


Figure 2.3: Comparison of infection mechanisms by ascomycete pathogens of plants and animals host (Sexton & Howlett, 2006)

Drawing parallels with bacterial pathogens, study has found that *Pseudomonas aeruginosa* which causes pneumonia, infections in blood (CDC, 2021) in human shows high degree of conservation in the virulence mechanism used to infect both human and plants. The pathogen also causes infection on the roots of *Arabidopsis* and sweet basil as its form a layer of biofilm under specific physiological conditions (Walker, et al. 2004). Evidence also showed that the bacterial pathogen used a common subset of virulence factors for pathogenesis in both plants and animals (Walker, et al. 2004) which further demonstrated that pathogens that infects range of host uses a common pathogenesis mechanism. Understanding the common mechanism behind the range of potential hosts for infection can shed lights and gives rise to better understanding of host specificity and mechanism of pathogenicity in the kingdom of fungus.

2.3.2 Development of Genomics Markers through Comparative Genomics

The emergence of sequencing technologies had increased the resolution of research into molecular causative factors in molecular plant pathology. Through genome sequencing of plant pathogens like *Magnaporthe oryzae* (Dean et al. 2005), *Botrytis cinerea* (Amselem et al. 2011)*, Ustilago maydis* (Kamper et al. 2006), and *Puccinia graminis* (Duplessis et al. 2011) coupling with improving bioinformatics methodology genome assembly, genome annotation, comparative genomics enabling pathologist to identify genomics features in fungal pathogens that plays important role in fungal pathogenicity, on top of that allowing further understanding of those genomic features will allow scientists to pursue and develop faster and more accurate diagnostic tool for fungus-related diseases.

Whole genome sequencing of plant fungal pathogens allows high quality genome assembly to identify reveal-underlying sequences of the fungus. Genome annotation of the assembled genome then predicts gene models based on *ab initio* prediction as well as homology searches (Yandell & Ence, 2012) to known nucleotide or protein sequences. Availability of an annotated genome allows downstream bioinformatics analysis such as polymorphic markers identification through genome mapping (Davey, et al. 2011) and comparative genomics (Wei, et al. 2002). Recent genomic studies, coupling with the advance application of bioinformatics tools had shed lights on fungal pathogenicity. A study on *Verticillium dahliae* proposed the possibility of horizontal gene transfer (HGT) from bacteria origins in which directly contributed to the pathogenicity of the fungus – known to be a plan pathogen that inflicts hundreds of plant species and causing huge economic losses annually (Shi-Kunne, et al. 2019).

Same effort was applied to the comparative genomics of human pathogenic fungi as well. Most prevalent fungal species that causes significant health implications in human are the *Candida* and *Aspergillus* (Moran, et al. 2011) hence understanding the sequences in the genomic level is extremely important to allow development of effective antifungal therapy and understanding emergence of drug resistance. A study was done to understanding drug resistance of *Candida auris* where genomic data such as epidemiology and evolutionary information were used for the study (Chybowska, et al. 2020). Comparative genomics study had also been done on *Aspergillus* to improve understanding of genome heterogeneity between *Aspergillus fumigatus*, *Aspergillus lentulus*, and *Aspergillus fumigatiaffinis* (Dos Santos, et al. 2020). These three species are extremely similar morphologically to one another hence making it challenging to distinguish one species from another by phenotypic observation (Alastruey-Izquierdo, et al. 2014). This make genomic study extremely important as sequencing and downstream bioinformatics analysis can uncover genomic features that are unique to each species such as Single Nucleotide Polymorphism.

Comparative genomics techniques were applied in studying not only genetic diversity, but also in discovery of important genomic markers such as Short Sequence Repeats (SSR), Short Tandem Repeats (STR), Long Tandem Repeats (LTR), Single Nucleotide Polymorphisms (SNP) etc. Recent study on *Fusarium oxysporum* is an example of such application of comparative genomics in uncovering genomics markers for quicker detection of pathogenic isolates of the species (van Dam, et al. 2017). The study includes candidate effector genes from 88 *Fusarium oxysporum* genomic assemblies for comparative genomics to distinguish the isolates based on the traits where it could differentiate between cucurbit-affecting *formae* speciales from each other and differentiating the pathogenic and non-pathogenic isolates.

General identification of pathogenic and non-pathogenic fungi often investigates genetic features such as the presence of what was considered as pathogenicity related genes and proteins. Presence or absence of pathogenicity-related genes is important in understanding fungal pathogenicity and its viability was demonstrated in a study comparing *Fusarium graminearum* and *Fusarium venenatum* where each is known as a non-pathogenic and a pathogenic species of fungi respectively (King, et al. 2018). The study presented a useful insight to support such a hypothesis as the group of scientists discovered, through comparative genomics that through a comparison of the proteomes of each species there were 15 putative secondary metabolite gene clusters, 109 secreted proteins, and 38 candidate effectors that are not identified in the non-pathogenic subject.

Comparative genomics effort will create a good foundation on using identified pathogenicity-related genes and the molecular markers identified for molecular diagnostic. Fungal infections on human or animal hosts are easier to detect and identified compared to plant disease caused by pathogenic fungi. Fungal nail infections or known technically as "onychomycosis" can be diagnosed easily as the disease symptoms can be observed visually through rotting of nails (Gupta, et al. 2000). The same can be said for many fungi disease caused by different genus of fungi such as *Aspergillosis*, *Candidiasis*, *Mucormycosis*, *Pneumocystis pneumonia* and many more (CDC, 2021). Fungal infection in human and animal jeopardize health and livelihood of the subject hence early diagnosis is crucial. With visible symptoms such as skin rashes or coughing, it is easier for early detection and diagnosis was done through direct microscopic examination of clinical samples, histopathology, culture, and serology of patient clinical samples (Kozel & Wickes, 2014). Fungus diseases in plants however in some cases is hard to detect and symptoms are not visible and could be too late when it is observed. Classic example of that is the basal stem rot (BSR) and upper stem rot (USR) by *Ganoderma boninense* (Hushiarian, R. et al. 2013). As the infection is not visually observable, it will be too late when its symptom is observable as palm trees dies from within 1 or 2 years, to 3 to more years depending on the age of the palm once symptoms is observable (Corley & Thiker, 2003).

In the case of BSR or USR caused by *Ganoderma boninense* traditional diagnostic methods will not be practical as it will be too late. Molecular diagnostic methods using PCR amplifications provides the way forward for early detection of fungal diseases that are not observable. This method requires the presence of unique genome sequence of the target organism, and this is usually a well-conserved region with polymorphic markers identifying different species. A specific primer (Hariharan & Prasannath, 2021) will be designed to amplify the target region of interest. Example target region of the fungal genomes that had been identified for molecular diagnostic includes the highly conserved internal transcribed spacer ITS-region in fungus – known for fungal diversity analysis and important marker for fungal DNA barcoding (Bellemain, et al. 2010), and alternative sequences such as cytochrome b gene which was used as a target region for Loop-Mediated Isothermal Amplification (LAMP) Assay for detection of airborne *Uromyces betae* (Kaczmarek, et al. 2019). Molecular diagnostic provides the possibility of early detection, and application of the methodology is applicable to both fungal diseases in plants or in animal and human.

2.3.3 Bioinformatics Tools and Platforms – Availability

Explosion of biological data produced by different research institutions around the world creates an entirely new challenges to uncover meaningful insights of these generated data and information. Data ranging from genomics, transcriptomics, and proteomics data requires further curation, annotations, and interpretation to facilitate useful and beneficial discovery. Throughout the years since the reduction of sequencing cost had led to development of various bioinformatics tools that enabled scientists to uncover mystery behind large pool of generated data. Fungus specific databases like FungiDB, CAZy, PHI-base, and DFVF are some examples were information and data related to fungal pathogenicity were made available. This provided opportunities for the research community to leverage on these data and using the right tools to uncover more insights into fungal pathogenicity.

Development of bioinformatics tools and software specializing and focusing on different paradigm of study is key to increase spectrum of understanding, enlarging perspective of biological research. These bioinformatics tools are developed to deal with data in various stages of readiness, ranging from tools like FastQC (Andrew, 2010) that enable quality control of DNA/RNA sequences generated by sequencing machine to downstream through that deal with more complex interpretation of data such as Cytoscape (Shannon, et al. 2003), VisANT (Hu, 2014), Pathway Studio (Nikitin, et al. 2003) and Patika (Demir, et al. 2002), which allow scientists to explore biological networks, as a mean to better understand integrative biology, system biology, and integrative bioinformatics.

Standalone tools such as BLAST, a universally common tool utilize for comparison of two or more DNA/RNA/Protein sequences understanding the degree of similarity and identity between sequences which implies degree of conservation of sequences among subject of studies, often utilize to understand relationship between species of organism. ClustalW is another example of such standalone tool that incorporates statistical analyses of subject sequences, building relationship trees of input sequences that allow not only understanding but also visualization of relative relationships between multiple sequences in study. Recent trends in bioinformatics tools development indicate that there are more requirement and necessities within the scientific community to have integrated tools that behave like an "One-Stop Centre" for biological data analysis as it can become cumbersome for scientists that does not have the required skillsets to execute sequence analysis via multiple bioinformatics tools as it requires time invested in understanding the selected bioinformatics tools and as such are a higher barrier to entry for most scientists. In view of such demanding unique scenario increasingly integrated bioinformatics analysis platform are developed for scientists for integrated sequence data analysis.

UGENE (Okonechnikov, et al. 2012) is an example of a bioinformatics tools and provide a platform for development of an integrated pipeline. UGENE provides a friendly user-interface for scientists to develop desired bioinformatics pipeline and workflows for sequence data analyst. With many popular standalone bioinformatics tools within UGENE, it also provides a user-friendly interface for scientists to easily build desired workflow with a drag-and-drop feature that requires minimum computer programming knowledge.

The continuous innovation of Next Generation Sequencing technology sees cost of raw Megabase of DNA sequence steadily dropping from when it was US$10K to less than US$100 in 2019, whereas cost per genome has seen identical trend in reducing from US$100M in 2001 to US$1K aligning with Moore's Law – a theory that states the doubling of compute power every two years and its known that technology improvements that are in trend with Moore's Law is seen as performing well (Wetterstrand, 2020). With the reducing cost in sequencing effort and increase availability of the technology across many areas of research, more sequencing data are generated – with some sequencing platforms like the Illumina NovaSeq generates 2TB-6TB of raw sequencing data for each sequencing runs that are performed (Besser, J. et al. 2018). With so much data generated it requires bioinformatics tools and software to process the datasets to generate useful insights into the massive pool of data.

2.3.4 Trends of Integrated Comparative Genomics Platform Development

Comparative genomic analysis usually involves the comparative analysis of sequence data from multiple sources, some within species and some across multiple species. These analyses usually involve multi-stage data analysis and therefore requires combination of bioinformatics tools and applications to draw meaningful discussions and deduction in quest of answering experimental hypothesis. Most comparative genomics platforms allow comparative analysis of DNA sequences and streamlining the process from data analysis to visualization of results. EDGAR (Dieckmann, et al. 2021) is such example of integrated comparative genomics platform and is one of the most popular platforms for gene based comparative genomics and differential gene content analysis. Venn diagrams or synteny plots can be generated to provide a user-friendly and visually appealing results interpretation.

CHAPTER 3

METHODOLOGY

3.1 Compute Resource and Environment

Google Cloud Compute was utilized to host a virtual machine running Ubuntu 18.10 with 10GB of RAM and 6 cores to run initial database creation and homology searches and a local Hyper-V virtual machine running Ubuntu 18.10 with 16GB of RAM was used to run downstream interpretive analysis.

3.2 Data Source

3.2.1 Fungal Genome Initiative

Genome sequences of 86 fungal species in this study was downloaded from the repository of Fungal Genome Initiative, a collective effort between Broad Institute Harvard and Massachusetts Institute of Technology and a wider fungal community (Broad Institute, 2014). The Fungal Genome Initiative has collected and sequenced fungal species that had portrayed importance of its existence and applications development in medicine, agriculture, and industry (Broad Institute, 2014). The initiative had sequenced more than 100 fungal species, of which includes well known human and plant pathogens like *Magnaporthe oryzae, Botrytis cinerea* and many more. These studies and sequencing projects are immensely important to explore and increase the understanding of fungal pathogenicity, as the sequencing of a fungal species lays important foundation for applied studies in the quest of answering question of fungal pathogenicity on its host, be it human or plant and discovering the answers to diagnose genomic pathological patterns or discovering and enhancing treatments for diseases caused by fungal pathogens.

All 86 fungi sequences as listed in Table 3.1 were downloaded from Fungal Genome Initiative FTP site and these sequences includes assembled supercontigs and contigs sequences, annotated genes and protein sequences, as well as sequences upstream and downstream gene coding regions. Genome annotations were done using pipeline and methodology established by Broad Institute Gene finding Method (Broad Institute, 2014) and it is a multistage genome annotation process the annotation process is described in detailed.

The Fungal Genome Initiative in total consists of both nucleotide and protein sequence data for 247 fungal species and isolates. With the duplication and existence of multiple isolates for some species a representative strain was select randomly for the search of homologous pathogenicity-related sequences and this resulted in the final 86 fungal species and sequences as listed in Table 3.1 for this study.

Table 3.1: List of 86 Fungal species

#	Fungal Species	Phylum	Human/Plant
1	*Arthroderma benhamiae*	Ascomycota	Human
2	*Aspergillus clavatus*	Ascomycota	Animal/Human
3	*Aspergillus flavus*	Ascomycota	Plant
4	*Aspergillus fumigatus*	Ascomycota	Human
5	*Aspergillus nidulans*	Ascomycota	Human
6	*Aspergillus niger*	Ascomycota	Plant
7	*Aspergillus oryzae*	Ascomycota	Human
8	*Aspergillus terreus*	Ascomycota	Human
9	*Blastomyces dermatitidis*	Ascomycota	Human/Animal
10	*Botrytis cinerea*	Ascomycota	Plant
11	*Candida albicans*	Ascomycota	Human
12	*Capronia coronata*	Ascomycota	Human
13	*Capronia epimyces*	Ascomycota	Human
14	*Capronia semiimmersa*	Ascomycota	Human
15	*Cladophialophora bantiana*	Ascomycota	Human
16	*Cladophialophora carrionii*	Ascomycota	Plant
17	*Cladophialophora immunda*	Ascomycota	Human

Table 3.1, continued.

#	Fungal Species	Phylum	Human/Plant
18	*Cladophialophora psammophila*	Ascomycota	Animal/Human
19	*Cladophialophora yegresii*	Ascomycota	Plant
20	*Coccidioides immitis*	Ascomycota	Human
21	*Colletotrichum graminicola*	Ascomycota	Plant
22	*Colletotrichum higginsianum*	Ascomycota	Plant
23	*Coniosporium apollinis*	Ascomycota	Plant
24	*Exophiala aquamarina*	Ascomycota	Animal/Human
25	*Exophiala mesophila*	Ascomycota	Animal/Human
26	*Exophiala oligosperma*	Ascomycota	Animal/Human
27	*Exophiala sideris*	Ascomycota	Animal/Human
28	*Exophiala spinifera*	Ascomycota	Animal/Human
29	*Exophiala xenobiotica*	Ascomycota	Animal/Human
30	*Fonsecaea multimorphosa*	Ascomycota	Animal/Human
31	*Fonsecaea pedrosoi*	Ascomycota	Animal/Human
32	*Fusarium graminearum*	Ascomycota	Plant
33	*Fusarium oxysporum*	Ascomycota	Plant
34	*Fusarium verticillioides*	Ascomycota	Plant
35	*Gaeumannomyces graminis*	Ascomycota	Plant
36	*Geomyces destructans*	Ascomycota	Animal
37	*Histoplasma capsulatum*	Ascomycota	Animal
38	*Magnaporthe oryzae*	Ascomycota	Plant
39	*Magnaporthe poae*	Ascomycota	Plant
40	*Microsporum canis*	Ascomycota	Plant
41	*Microsporum gypseum*	Ascomycota	Human
42	*Neosartorya fischeri*	Ascomycota	Human
43	*Neurospora crassa*	Ascomycota	Human
44	*Ochroconis gallopava*	Ascomycota	Human
45	*Paracoccidioides brasiliensis*	Ascomycota	Human
46	*Paracoccidioides sp*	Ascomycota	Human
47	*Phaeosphaeria nodorum*	Ascomycota	Human
48	*Phialophora europaea*	Ascomycota	Plant
49	*Pneumocystis carinii*	Ascomycota	Human
50	*Pneumocystis jirovecii*	Ascomycota	Human
51	*Pneumocystis murina*	Ascomycota	Human
52	*Pyrenophora tritici-repentis*	Ascomycota	Human
53	*Rhinocladiella mackenziei*	Ascomycota	Plant
54	*Schizosaccharomyces cryophilus*	Ascomycota	Human
55	*Schizosaccharomyces japonicus*	Ascomycota	Plant
56	*Schizosaccharomyces octosporus*	Ascomycota	Human
57	*Schizosaccharomyces pombe*	Ascomycota	Human
58	*Sclerotinia sclerotiorum*	Ascomycota	Human

Table 3.1, continued.

#	Fungal Species	Phylum	Human/Plant
59	*Sporothrix schenckii*	Ascomycota	Human
60	*Trichophyton equinum*	Ascomycota	Human
61	*Trichophyton interdigitale*	Ascomycota	Human
62	*Trichophyton rubrum*	Ascomycota	Animal
63	*Trichophyton tonsurans*	Ascomycota	Human
64	*Trichophyton verrucosum*	Ascomycota	Human
65	*Verticillium alfalfae*	Ascomycota	Human
66	*Verticillium dahliae*	Ascomycota	Plant
67	*Cryptococcus gattii*	Basidiomycota	Plant
68	*Cryptococcus neoformans*	Basidiomycota	Human
69	*Microbotryum violaceum*	Basidiomycota	Human
70	*Puccinia graminis*	Basidiomycota	Plant
71	*Puccinia striiformis*	Basidiomycota	Plant
72	*Puccinia triticina*	Basidiomycota	Plant
73	*Ustilago maydis*	Basidiomycota	Plant
74	*Batrachochytrium dendrobatidis*	Chytridiomycota	Human/Animal
75	*Spizellomyces punctatus*	Chytridiomycota	Unknown
76	*Anncaliia algerae*	Microsporidia	Human
77	*Edhazardia aedis*	Microsporidia	Human
78	*Encephalitozoon cuniculi*	Microsporidia	Animal
79	*Encephalitozoon intestinalis*	Microsporidia	Human
80	*Nematocida parisii*	Microsporidia	Human
81	*Nematocida sp1*	Microsporidia	Human
82	*Nosema ceranae*	Microsporidia	Insect
83	*Vavraia culicis*	Microsporidia	Insect
84	*Vittaforma corneae*	Microsporidia	Human
85	*Mucor circinelloides*	Mucoromycota	Human
86	*Rhizopus delemar*	Mucoromycota	Plant

Of the 86 species of fungi downloaded most of the fungal species resides in the phylum of Ascomycota – comprises of nearly 80% of the datasets. The remaining entries comprises member of fungi from Basidiomycota, Chytridiomycota, Mucormycotina, and Microsporida. All species in this study are pathogenic fungi but infects different hosts ranging from animal, human, and plant.

3.2.2 Data Clean Up

Although all variations of sequences were available for each fungal species downloaded, the study focuses on utilizing protein sequences for comparative analysis between different species as it provides lower level of resolution – less variations than using nucleotide sequences however it has a higher level of sensitivity as it would easily pick up variations in sequences between sequences of organisms from different analysis. This is applicable for all sequence analyses other than the extraction of Single Nucleotide Polymorphism.

3.3 Fungus Pathogenic-related Databases

3.3.1 Pathogen Host Interaction - PHI-base

Pathogen Host Interaction Database (Winnenburg, et al. 2006) or better known as PHI-base is a database that contains collection of experimentally verified fungal, oomycetes and bacterial pathogens that are causative agents for inflicting various diseases in its inhabited host that ranges from animals, plants, other fungal species as well as insects. The database was curated by domain experts coupling with experimental results and through gene disruption and complementation methodology.

Protein sequences were downloaded from the website of PHI-base Release 4.5 which consists of 5,183 genes that displayed either increase / decrease in disease virulence. These PHI-base genes were identified from 264 pathogens, all of which are known to cause over 465 types of diseases. PHI-base Release 4.5 was downloaded and created a local PHI-base by using makeblastdb (version 2.6.0) with the following command:

```
makeblastdb -dbtype prot -in <PHI-base Release 4.5 FASTA> -out <Output DB name>
```

3.3.2 Carbohydrate-Active enZYmes Database – CAZY

Carbohydrate-Active enZYmes Database (Lombard, et al. 2014) or known more popularly in its acronym CAZY is a database that contains protein sequences of structurally-related catalytic and carbohydrate-binding modules that are known to have different modes of interactions with glycosidic bonds – important linkage and type of covalent bond that joins carbohydrate molecule to another group. Glycosidic bonds are fundamental linkages found in cellular walls (Joseleau & Perez, 2016) thus are prime targets of carbohydrate-active enzymes and thus carbohydrate-active enzymes are considered as candidate fungal pathogenic gene products because of the capability to degrade the plant cell wall. These enzymes classes and associated modules are involved in various biological pathway of the host organism as described in Table 3.2.

Table 3.2: Enzyme Classes and Associated Modules

Family	Description
Glyicoside Hydrolases (GHs)	Involves in hydrolysis and/or rearrangement of glycosidic bonds
GlycosylTransferases (GTs)	Involves in formation of glycosidic bonds
Polysaccharide Lyases (PLs)	Involves in non-hydrolytic cleavage of glycosidic bonds
Carbohydrate Esterases (CEs)	Involves in hydrolysis of carbohydrate esters
Auxiliary Activities (AAs)	Involves in redox enzymes that act in conjunction with CAZymes
Carbohydrate-Binding Modules (CBM	Involves in adhesion to carbohydrates

Protein sequences from the Carbohydrate-Active enZYme Database were downloaded from dbCAN2 meta server. dbCAN2 meta server is an automated Carbohydrate-active enzyme ANnotation web server supported by the National Science Foundation of the United States of America (Yin, et al. 2012). A total of 921,174 protein sequences in FASTA format were downloaded from CAZY Database dated 20th July 2017 and a local CAZY database was created using makeblastdb (version 2.6.0) with the following command:

```
makeblastdb – dbtype prot -in <CAZY Database 07202017> -out <Output DB name>
```

3.3.3 Database of Virulence Factors in Fungal Pathogens - DFVF

The Database of Virulence Factors in Fungal Pathogens (DFVF) (Lu, et al. 2012) was a project aimed at filling the missing gaps in understanding of fungal pathogenicity by aggregating all known virulence factors also developing an algorithm that allows prediction of potential candidate genes that will be contributing to development of fungal pathogenicity. The database was built by leveraging of text-mining technique sued by PubMed database and the Internet by looking for fungal disease virulence keywords and in-house tools were developed to allow searching of relevant supporting literatures.

In total there were 2058 protein sequences within the database that were downloaded from the database and a local copy of the database were created by using makeblastdb (version 2.6.0) with the following command:

```
makeblastdb – dbtype prot -in <DVFV Database> -out <Output DB name>
```

3.4 Fungal Pathogenic Gene Comparative Pipeline

Development of the Fungal Pathogenic Gene Comparative Pipeline includes multiple steps to provide annotation and comparison of input protein sequences against aggregated known pathogenic Gene Database which includes the sequences from the PHI-base, the Carbohydrate-Active enZYme Database, as well as the Database of Fungal Virulence Factors. The 86 fungal genome sequences, including nucleotide and protein sequences were downloaded.

The pipeline as visualized in Figure 3.1 incorporates homology searches, multiple sequence alignments, phylogenetic analysis to provide interpretation of relationship between in-query protein sequence. Corresponding gene sequences were identified by aligning nucleotide gene sequences with BLASTX to the identified candidate common protein. Visualization of the data including Multiple Sequence Alignments and Phylogenetic Tree can be done through bioinformatics visualization tools like Unipro UGENE (Okonechnikov, et al. 2012) and Artemis (Carver, 2012) can be used to visualize SNP data that were mined using SNP-Sites (Page, et al. 2016).

The Fungal Pathogenic-Related Gene Comparative Pipeline as visualized in Figure 3.1 was then constructed using multiple bioinformatics tools with substantial shell scripting to allow post-processing of results files from various tools. Source code of all shell scripts are attached in Appendix A.

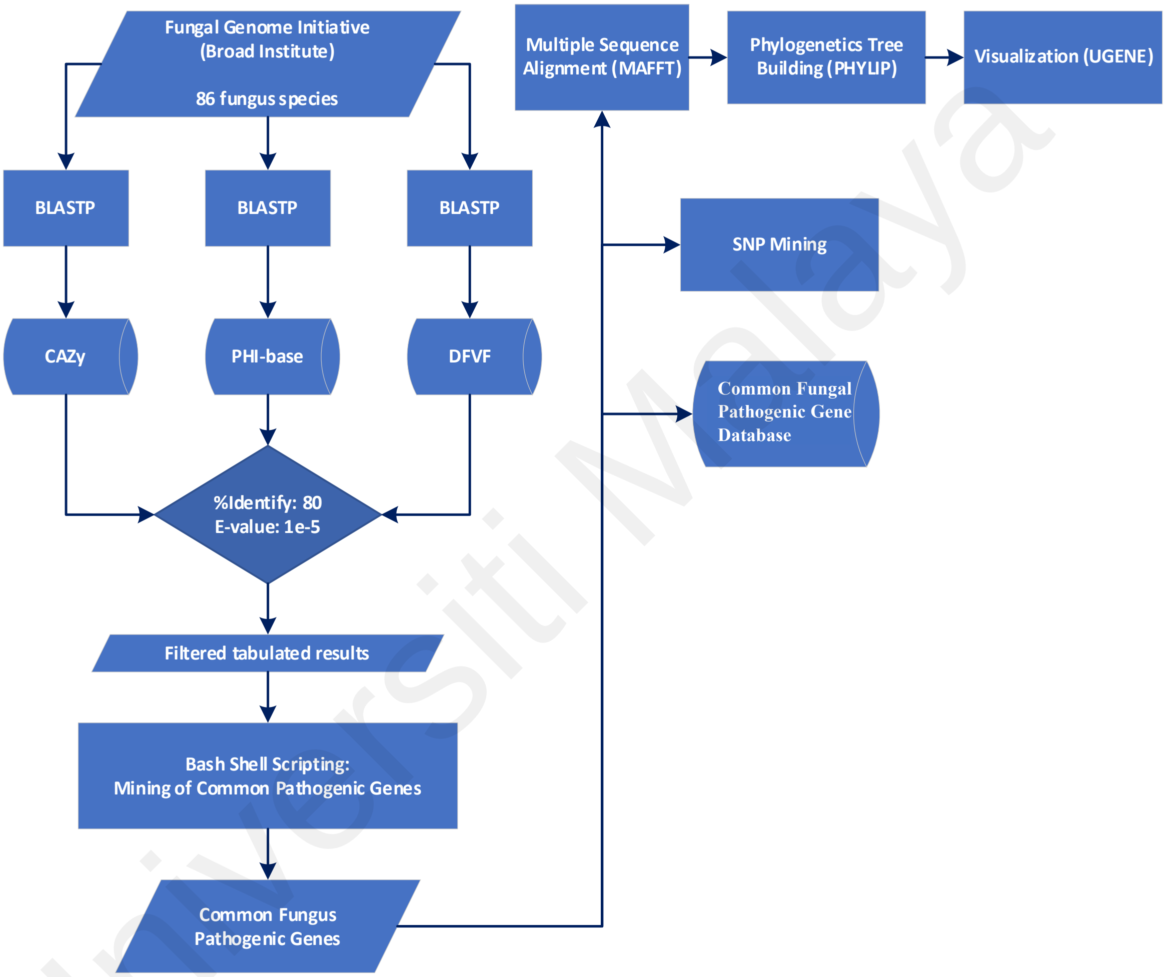


Figure 3.1 Fungal Pathogenic-Related Gene Comparative Pipeline

3.4.1 Identification of Common Pathogenicity-Related Genes

Homology searches using each of the 86 fungal proteome against the three databases namely CAZy, PHI-base, and DFVF (Lu, et al. 2012) yielded results in BLAST tabular format for each of the fungal species. Basic Local Alignment Search Tool, or best known as BLAST (Ye, et al. 2006) is the most widely used bioinformatics sequence alignment tool utilized to search for homology between two given sequences calculating an alignment score based on sequence similarities scores that includes scoring based on mismatches, gap opening and etc. Local NCBI-BLAST+ was utilized for homology searches of annotated translated gene sequences of 86 fungal species in study against local copies of PHI-base, CAZy, and DFVF in Fungal Pathogens. In an effort to improve efficiency in sequence homology search a massive parallel approach was developed using Shell Scripting Language. The protein sequences of the 86 fungi in study is first separated to different portion, each portion is then submitted to run BLASTP analysis on the compute. BLASTP parameters curated includes using an e-value cut-off of 10^{-5} and the results are produced in tabular format.

Although there are no hard rules around cut-off parameters for E-value and %Identity, the values chosen for this study largely based on the understanding of good ranges based on study by Pearson, 2013. E-value < 0.001 is reliable for inferring homology between protein:protein alignments whereas %Identity between 70-80 is useful to infer evolutionary distances. Stringent combination of both parameters will enable identification of high confidence homologs, and that was the approach taken for this study.

The following BLASTP command were utilized to execute homology searches of fungus protein sequences against the three stated pathogenic Gene Database:

```
blastp -query <input sequence FASTA> -db <Database file> -evalue 1e-5 -outfmt 6 -num_threads 2 -out <output file name>
```

Upon completion of homology searches of the 86 fungal protein sequences in tabular output format the results are then processed and filtered based on different percent identity scores starting from 50 with an increment of 10 to percent identity score of 90 and ending with a final cut-off of 95 maximum identity score. Common candidate pathogenic protein is then shortlisted via text mining sorting and filtering the protein sequence identifier.

Genes sequences of the common pathogenic genes are then extracted by aligning genes sequences from the 86 fungi with BLASTX to the protein sequences of the common pathogenic protein sequences. With that the top hit of each fungal sequences with a percent identity score of 80 against the common pathogenic protein sequence:

```
blastx -query <input sequence FASTA> -db <Database file> -evalue 1e-5 -outfmt 6 -num_threads 2 -out <output file name>
```

These common candidate pathogenic genes will then be identified for further analysis with Multiple Sequence Alignment and Phylogenetic Analysis.

3.4.2 Multiple Sequence Alignment of Homologous Pathogenic Genes

Common pathogenic genes across all 86 fungal species that are identified from homology searches are then subject to multiple sequence alignment to produce both sequence alignment files and phylogenetic trees. MAFFT (Katoh, et al. 2002) was utilized to perform Multiple Sequence Alignments of the candidate common genes across most species of fungi and multi-FASTA alignments. Default gap opening penalty of 1.53 was utilized to generate multiple sequence alignments and a phylogenetics tree is then generated by using PHYLIP (Felsenstein, 1989) using F84 Data Matrix neighbour joining method.

3.4.3 Single Nucleotide Polymorphisms (SNP) Mining

Single Nucleotide Polymorphisms mining from multi-FASTA sequence alignment is carried out using SNP-sites (Page, et al. 2016). This is a different approach comparing to conventional SNP mining tool leveraging on deep sequencing data like SAMtools (Li, et al. 2009) as this tool was developed to cater for extracting SNPs from multiple sequence alignment files output from various MSA tools such as MUSCLE (Edgar, 2004), PRANK (Löytynoja, 2014), MAFFT (Katoh & Standley, 2013), or ClustalW (Thompson, et al. 1994). According to Page et al. (2016) SNP-Sites takes only 267 seconds using 59 MB of RAM and 1 CPU core to process multiple sequence alignment files of 8.3 GB file size which approximate to datasets of 1842 taxa with 22618 SNP sites, making it possible and feasible to process large multiple sequence alignment files in a conventional computer.

The output file from MAFFT was fed to SNP-Sites with default settings and three output file types were obtained and there are VCF (Variant Calling Format), aln (Multiple Sequence Alignment file), and a phylogenetic tree file. All phylogenetic tree files were then visualized using Unipro UGENE (Okonechnikov, et al. 2012), a cross-platform bioinformatics software, and VCF files were visualized with Artemis (Carver, 2012), an integrated platform that allows visualization of sequence and its feature data.

3.4.4 Phylogenetics Tree Building

Phylogenetics trees were built using multiple sequence alignment files from MAFFT using PHYLIP Neighbour Joining Algorithm, and Jones-Taylor-Thornton distance matrix model with a Coefficient of variation of substitution rate among sites of 0.50 and Transition/transversion ratio of 2.00. These phylogenetic trees were then visualized using Unipro UGENE.

3.4.5 Common Fungal Pathogenic Gene Database (CFPG)

The building of a web application for the Common Fungal Pathogenic Gene Database to serve as a portal to access data and information found in the study is essential. For the platform of choice, the XAMPP (Apache Friends, 2023) release 7.2.34 Web Server solution was installed to host the database and the web page allowing access to the CFPG Database. The following services are utilised in XAMPP:

- Apache (Web Page)
- Tomcat (Web Server)
- MySQL (Database)
- PHP (Application)

Joomla!, (Rochen, 2017) a Content Management System was used to develop the Front End of the Common Fungal Pathogenic Gene Database along with the Art Table Joomla! extension which enable display of data and allowing user input to search and export data for further study and utilization. The Common Fungal Pathogenic Gene Database is built on the Web Server solution, with standard tables provided by Joomla! CMS template and three custom tables created to store data.

The first table that was created for the CFPG Database was MASTER_FUNGUS. This table contains list of all fungal species that was used for Common Fungal Pathogenicity-related Genes mapping and extraction. The second table that was created for the CFPG Database was MASTER_COMMON_GENE as listed in Table 3.4. This table stores all proteins identified from homology searches against all fungal species listed in Table 3.3 and passed through filtering criteria.

Table 3.3: Master list of All Fungal Species Utilized.

Name	Type	Null	Description
SEQ_NUM	int(11)	No	Auto-incremental unique sequence number.
SPECIES	varchar(254)	No	Full fungal species name.
TAX_ID	int(11)	Yes	NCBI Taxonomy ID related to the species.
PHYLUM	varchar(50)	Yes	Phylum of the species.
CHANGED_ON	datetime	No	Datetime stamp automatically updated when a record is updated.

Table 3.4: Master list of all Common Fungal Pathogenicity-Related Genes

Name	Type	Null	Description
ENTRY_NUM	int(11)	No	Auto-incremental unique sequence number.
NAME	varchar(9)	Yes	CFPG ID, primary key of the table.
UNIPROT_ENTRY_NAME	varchar(50)	Yes	UniProt Entry Name associated to the CFPG entry.
UNIPROT_ENTRY	varchar(50)	Yes	UniProt Entry associated to the CFPG entry.
DB_ENTRY_NAME	varchar(50)	Yes	Source Database Entry Name associated with the CFPG ID.
DB_ENTRY_TYPE	varchar(50)	Yes	Source Database Entry Type associated with the CFPG ID.

Table 3.4: Master list of all Common Fungal Pathogenicity-Related Genes

Name	Type	Null	Description
SOURCE_DB	varchar(50)	Yes	Source Database of associated with the CFPG ID.
FAMILY	varchar(50)	Yes	CAZy Family. Only populated for CFPG ID where Source Database is CAZy.
FAMILY_DESC	varchar(50)	Yes	CAZy Family Description. Only populated for CFPG ID where Source Database is CAZy.
ORGANISM	varchar(254)	Yes	Host organism of the gene based on UniProt.
INTERPRO_ID	varchar(254)	Yes	InterPro ID associated with the CFPG ID.
PROTEIN_NAME	varchar(254)	Yes	Protein Name of the associated CFPG ID based on UniProt.
GENE	varchar(50)	Yes	Gene Name of the associated CFPG ID based on UniProt.
LENGTH	int(11)	Yes	Protein sequence length of the associated CFPG ID based on UniProt.
HOST	varchar(254)	Yes	Known host that are affected by the CFPG ID entry.
RELATED_DISEASE	varchar(254)	Yes	Known diseases that caused by the CFPG ID entry.
CHANGED_ON	datetime	No	Datetime stamp automatically updated when a record is updated.
UNIPROT_LINK	varchar(254)	Yes	Link to UniProt for the CFPG ID.

Third and the last table as listed in Table 3.5 that was created for the CFPG Database was the GENE_SPECIES_MAPPING where this table contains a mapping list between each CFPG Genes and all fungal species where its homologs are found.

Table 3.5: Mapping of CFPG ID to fungal species

Name	Type	Null	Description
SEQ_NUM	int(11)	No	Auto-incremental unique sequence number.
NAME	varchar(9)	No	CFPG ID, primary key of the table.
FUNGUS_SPECIES	int(11)	No	Fungal species associated with the CFPG ID.

Data for each table includes links to primary databases such as UniProt (UniProt Consortium, 2021) and NCBI Taxonomy (Schoch, et al. 2020). Once results are obtained data are compiled and collected in Excel spreadsheets and exported to csv format before uploading to the MySQL. The front-end of the web application was developed using Joomla! CMS, using the default Beez3 template. The relationship between the tables are visualized in Figure 3.2.

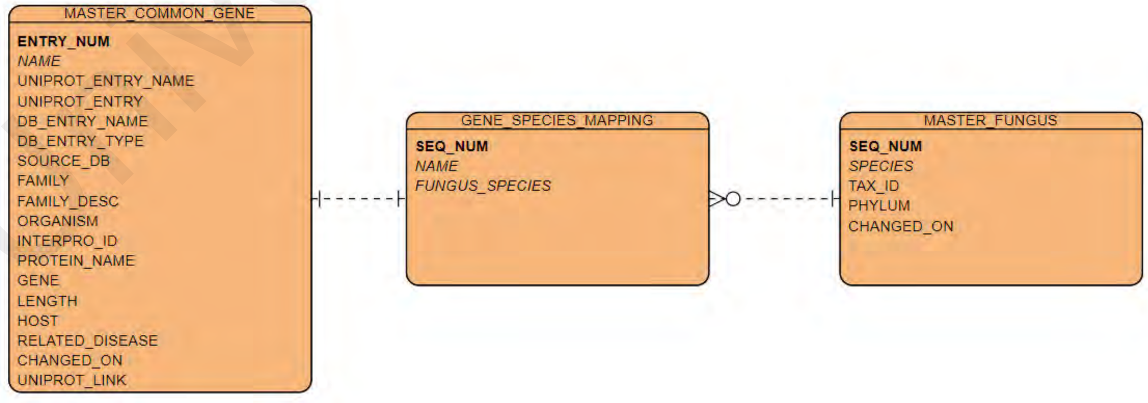


Figure 3.2: Entity-Relationship Diagram of 3 Main tables for CFPG.

CHAPTER 4

RESULTS

4.1 Results from Identification of Common Pathogenic Genes

Homology searches of 86 fungal species against PHI-base, CAZy, and DFVF yielded homologous hits based on different Maximum %Identity cut-off values of 50, 60, 70, 80, 90, and 95 can be seen in Tables 4.1, 4.2, and 4.3 below:

Table 4.1: Homologous Hits of 86 fungal species against PHI-base

Species	All	50	60	70	80	90	95
Anncaliia algerae	8431	5393	3964	2614	1493	690	335
Arthroderma benhamiae	42858	26712	18480	12279	7007	2999	1351
Aspergillus clavatus	54826	33613	22424	14545	8379	3538	1615
Aspergillus flavus	66840	40402	26478	16861	9284	3825	1790
Aspergillus fumigatus	35	24	12	9	4	1	0
Aspergillus nidulans	60701	37195	24364	15690	8785	3691	1674
Aspergillus niger	50284	29871	18966	11975	6612	2812	1266
Aspergillus oryzae	63477	37966	25504	16204	9080	3849	1749
Aspergillus terreus	61026	36952	24327	15430	8596	3572	1599
Batrachochytrium dendrobatidis	45677	30034	21765	14220	7796	3223	1508
Blastomyces dermatitidis	50764	31799	22266	14704	8573	3442	1606
Botrytis cinerea	47457	29489	20057	12898	7263	2954	1356
Candida albicans	33873	21693	15159	10107	6140	2580	1194
Capronia coronata	48730	31173	20548	13517	7810	3302	1522
Capronia epimyces	53912	34339	22177	14522	8468	3584	1668
Capronia semiimmersa	65449	4182	2693	1749	1019	4385	2023

Table 4.1, continued.

Species	**All**	**50**	**60**	**70**	**80**	**90**	**95**
		8	9	5	1		
Cladophialophora bantiana	61415	39491	25367	16133	9168	3816	1716
Cladophialophora carrionii	49753	31711	21026	13694	7919	3348	1514
Cladophialophora immunda	82518	53557	33533	21038	11844	5016	2349
Cladophialophora psammophila	65952	42788	27284	17155	9677	4073	1837
Cladophialophora yegresii	47761	30709	20969	13688	7837	3327	1555
Coccidioides immitis	38927	24110	16522	10993	6364	2749	1285
Colletotrichum graminicola	57633	34932	23180	14942	8471	3550	1752
Colletotrichum higginsianum	67762	39840	26051	16731	9469	3931	1874
Coniosporium apollinis	45931	30125	20360	13351	7795	3231	1425
Cryptococcus gattii	30757	19922	14217	9318	5259	2208	1034
Cryptococcus neoformans	40700	26411	18655	12130	6887	2908	1353
Edhazardia aedis	8771	5846	4262	2876	1554	707	328
Encephalitozoon cuniculi	8870	5795	4244	2882	1659	756	385
Encephalitozoon intestinalis	7585	4899	3515	2397	1350	557	256
Exophiala aquamarina	65459	41643	26571	16790	9489	3939	1793
Exophiala mesophila	65490	42441	27502	17598	9916	4086	1793
Exophiala oligosperma	79958	51658	32803	21092	11746	4853	2244
Exophiala sideris	64877	41599	26367	16854	9714	4221	1970
Exophiala spinifera	62017	39793	25059	16022	9096	3803	1704
Exophiala xenobiotica	73620	47129	29630	18905	10834	4603	2137
Fonsecaea multimorphosa	65044	42308	26776	16761	9457	3980	1835
Fonsecaea pedrosoi	66519	43049	27485	17258	9711	4089	1909
Fusarium graminearum	61421	37272	24613	15648	8954	3619	1634
Fusarium oxysporum	155229	9744	6246	3987	2268	9442	4609

Table 4.1, continued.

Species	All	50	60	70	80	90	95
		4	0	0	1		
Fusarium verticillioides	126510	78427	50438	32548	18662	7604	3717
Gaeumannomyces graminis	54406	34670	23549	15524	8922	3708	1709
Geomyces destructans	41268	26220	18337	12091	7041	2884	1314
Histoplasma capsulatum	42226	27259	18962	12271	6819	2789	1287
Magnaporthe oryzae	55316	33861	22613	14673	8440	3602	1711
Magnaporthe poae	47171	29574	19755	12929	7362	3068	1408
Microbotryum violaceum	39696	25981	18868	12632	7265	3188	1414
Microsporum canis	50122	30972	20757	13588	7827	3292	1473
Microsporum gypseum	46633	28728	19592	12979	7615	3280	1482
Mucor circinelloides	88266	57208	42447	29666	17616	7354	3359
Nematocida parisii	7353	4887	3339	2031	1105	400	194
Nematocida sp1	7767	5132	3478	2058	1071	415	185
Neosartorya fischeri	60456	36598	24622	16089	9186	3893	1710
Neurospora crassa	41052	26320	17920	11834	6709	2889	1313
Nosema ceranae	7942	5342	3964	2694	1480	649	328
Ochroconis gallopava	63980	41258	27526	17876	10220	4421	1906
Paracoccidioides brasiliensis	38046	24235	17140	11594	6726	2924	1367
Paracoccidioides sp.	38242	24324	17251	11535	6689	2877	1329
Phaeosphaeria nodorum	49330	30514	20719	13467	7776	3345	1596
Phialophora europaea	55320	34930	22625	14362	8335	3528	1653
Pneumocystis carinii	25051	16083	12226	8588	5241	2290	1056
Pneumocystis jirovecii	22669	14948	11274	7819	4730	2080	924
Pneumocystis murina	22683	14689	11041	7864	4704	2044	958
Puccinia graminis	38733	2505	1753	1183	6896	3055	1357

Table 4.1, continued.

Species	All	50	60	70	80	90	95
		0	2	8			
Puccinia striiformis	44368	28740	20547	13716	7931	3474	1570
Puccinia triticina	43139	27669	19741	13355	7851	3375	1530
Pyrenophora tritici-repentis	50866	31153	21041	13615	7721	3350	1558
Rhinocladiella mackenzie	59945	38422	25179	16022	9012	3800	1756
Rhizopus delemar	95150	62764	45972	31020	17472	7191	3408
Schizosaccharomyces cryophilus	33997	22052	16819	11753	7311	3215	1447
Schizosaccharomyces japonicus	34509	22134	16728	11806	7149	3086	1438
Schizosaccharomyces octosporus	33558	21752	16431	11533	7025	3097	1424
Schizosaccharomyces pombe	34864	22537	16920	11887	7187	3121	1404
Sclerotinia sclerotiorum	46605	29082	20093	13021	7440	3148	1383
Spizellomyces punctatus	62654	39692	29114	20203	12123	4980	2262
Sporothrix schenckii	47307	30427	20335	13117	7508	3098	1404
Trichophyton equinum	44152	27008	18342	12253	7091	3000	1433
Trichophyton interdigitale	48329	29815	20156	13138	7451	3130	1440
Trichophyton rubrum	59575	37335	25726	17150	9328	3936	1832
Trichophyton tonsurans	44113	27369	18467	12273	7077	3091	1491
Trichophyton verrucosum	845	584	471	359	244	92	32
Ustilago maydis	34765	22448	15584	10307	5829	2439	1068
Vavraia culicis	8317	5489	4062	2649	1487	597	291
Verticillium alfalfae	44783	27957	18692	11934	6429	2616	1224
Verticillium dahliae	48935	31217	20946	13736	7620	3130	1396
Vittaforma corneae	10227	6899	4680	3021	1539	584	255

Table 4.2 Homologous Hits of 86 fungal species against CAZy

Species	All	50	60	70	80	90	95
Anncaliia algerae	11204	4951	3724	2621	1982	1406	516
Arthroderma benhamiae	99214	45017	34948	23981	13385	5123	2285
Aspergillus clavatus	184120	76138	55037	35474	21021	8981	4347
Aspergillus flavus	276908	114914	84414	54237	32042	14592	6721
Aspergillus fumigatus	512	367	361	360	356	301	48
Aspergillus nidulans	241859	99186	72758	47656	27679	12403	5851
Aspergillus niger	177327	74642	55229	35729	20245	8467	4391
Aspergillus oryzae	263005	109191	80191	52631	31905	14321	6543
Aspergillus terreus	246665	102044	76915	49699	28872	12895	6067
Batrachochytrium dendrobatidis	98202	59403	45646	31767	20178	8782	4304
Blastomyces dermatitidis	123653	54854	40483	25185	13961	5344	2159
Botrytis cinerea	227150	97611	73967	48389	28442	12383	5620
Candida albicans	81699	33643	24847	15774	8122	3325	1161
Capronia coronata	125941	52647	38497	24940	14117	6192	2869
Capronia epimyces	122526	54279	40312	25413	14360	5982	2779
Capronia semiimmersa	151183	63342	46434	28696	15366	6241	2802
Cladophialophora bantiana	159733	67374	46529	30709	16744	6744	2641
Cladophialophora carrionii	143187	57474	43035	27671	15102	6417	2688
Cladophialophora immunda	168489	70908	51639	33085	17956	7475	3364
Cladophialophora psammophila	162236	68229	49272	31959	17386	7714	3486
Cladophialophora yegresii	134906	54097	40424	25173	13846	6023	2549
Coccidioides immitis	90222	42998	33198	21998	12404	5158	2313
Colletotrichum graminicola	277256	110666	82593	54295	33175	15478	7820
Colletotrichum higginsianum	0	0	0	0	0	0	0
Coniosporium apollinis	137142	63962	48007	31232	17152	7630	3682
Cryptococcus gattii	82635	36762	26054	15361	8775	4469	2518
Cryptococcus neoformans	100877	46684	34053	21433	12072	5202	2632
Edhazardia aedis	11011	6434	4906	2782	1607	479	124
Encephalitozoon cuniculi	9124	5507	3711	2289	1380	653	274
Encephalitozoon	8485	4949	3313	2131	1325	609	212

Table 4.2, continued.

Species	All	50	60	70	80	90	95
intestinalis							
Exophiala aquamarina	173179	72434	52493	34158	18802	7995	3680
Exophiala mesophila	151959	61434	45527	28550	16961	6607	3165
Exophiala oligosperma	179800	69066	50813	31610	17919	7772	3384
Exophiala sideris	147650	60685	46199	30859	16559	6796	2978
Exophiala spinifera	153570	63640	45990	29369	16214	7045	3546
Exophiala xenobiotica	192088	77590	57353	36411	19765	8857	4209
Fonsecaea multimorphosa	157301	63095	46126	31029	16313	6711	2865
Fonsecaea pedrosoi	148538	63518	46837	31184	17531	7384	3048
Fusarium graminearum	249893	105491	76423	49823	28112	11800	5461
Fusarium oxysporum	426394	186618	135404	84020	50053	21137	9882
Fusarium verticillioides	359496	156996	115279	74420	45129	19729	8657
Gaeumannomyces graminis	249380	105521	78281	52708	31707	14583	6894
Geomyces destructans	131057	63245	47029	31090	17469	7356	3576
Histoplasma capsulatum	95414	40197	30112	19590	12014	4622	2095
Magnaporthe oryzae	254330	106645	78444	50776	30502	14296	6644
Magnaporthe poae	224614	92896	68030	44821	27166	12527	6173
Microbotryum violaceum	108248	57639	45233	31440	18045	7839	3857
Microsporum canis	112850	52847	39529	26064	14430	6399	3186
Microsporum gypseum	104891	47123	36353	24638	13874	5983	2687
Mucor circinelloides	185977	91770	67176	41803	24240	10554	4977
Nematocida parisii	10790	5770	3879	2207	1378	665	275
Nematocida sp1	11541	7796	6151	3770	2206	1226	627
Neosartorya fischeri	103578	74284	46994	26782	11652	5537	
Neurospora crassa	166370	68097	50476	32563	19692	8185	3613
Nosema ceranae	8397	4939	3544	2054	767	254	105
Ochroconis gallopava	153838	62812	46341	28900	16063	7072	3574
Paracoccidioides brasiliensis	86598	39527	29858	18648	10991	4141	1714
Paracoccidioides sp.	89109	40964	30292	18312	10221	4255	1752
Phaeosphaeria nodorum	248720	100718	74753	49644	28024	11821	5517
Phialophora europaea	176451	66132	50499	33097	18464	7578	3666
Pneumocystis carinii	33673	18882	14515	9525	5715	1883	703
Pneumocystis jirovecii	31664	16977	13016	8061	4432	1690	729
Pneumocystis murina	32099	17653	13464	8720	5207	1858	733
Puccinia graminis	159381	73806	56878	39490	25268	11431	5289
Puccinia striiformis	179904	83680	62521	43220	28815	14239	6986

Table 4.2, continued.

Species	All	50	60	70	80	90	95
Puccinia triticina	168368	77761	58944	40566	25131	10465	4994
Pyrenophora tritici-repentis	230016	95158	71236	48037	29143	13050	5952
Rhinocladiella mackenzie	139441	61826	45896	30387	17080	7317	3526
Rhizopus delemar	203171	95598	70379	44975	26986	12068	6328
Schizosaccharomyces cryophilus	65666	25462	18957	12294	7696	3279	1313
Schizosaccharomyces japonicus	65962	27411	20679	13689	8288	3588	1527
Schizosaccharomyces octosporus	64526	25423	19043	12320	7856	3022	1235
Schizosaccharomyces pombe	68726	27991	21043	13868	9007	3783	1931
Sclerotinia sclerotiorum	200652	83233	61660	40234	21791	9294	4384
Spizellomyces punctatus	115870	68652	50308	31568	19378	8479	4299
Sporothrix schenckii	178490	73774	52723	34730	20086	9134	4387
Trichophyton equinum	97460	44261	33895	24191	13903	6077	2720
Trichophyton interdigitale	104025	47214	36221	24794	13883	5889	2809
Trichophyton rubrum	129213	57752	42870	27476	15059	6572	3319
Trichophyton tonsurans	99283	45144	34381	23497	13025	5740	2573
Trichophyton verrucosum	489	435	367	213	79	9	2
Ustilago maydis	105675	42719	31440	21105	12528	5081	2191
Vavraia culicis	9336	4135	3077	1789	1044	473	269
Verticillium alfalfae	256661	104978	77134	52448	30955	13642	6520
Verticillium dahliae	259763	107499	81420	54403	33405	16328	8298
Vittaforma corneae	11048	6869	4579	2535	1575	886	448

Table 4.3 Homologous Hits of 86 fungal species against DFVF

Species	All	50	60	70	80	90	95
Anncaliia algerae	3725	2056	1497	1033	620	303	147
Arthroderma benhamiae	16895	9162	6590	4482	2600	1058	491
Aspergillus clavatus	18185	10213	7256	4906	2855	1211	577
Aspergillus flavus	21519	11909	8290	5478	3052	1235	583
Aspergillus fumigatus	4	2	1	1	1	0	0
Aspergillus nidulans	19022	10534	7552	5102	2921	1232	575
Aspergillus niger	14089	7759	5365	3653	2048	891	416
Aspergillus oryzae	20739	11274	8045	5374	3079	1307	582
Aspergillus terreus	19136	10526	7473	5060	2852	1219	568
Batrachochytrium dendrobatidis	20139	10821	7749	5173	2887	1247	564
Blastomyces dermatitidis	19073	10342	7663	5139	2962	1238	559
Botrytis cinerea	17052	9471	6847	4578	2640	1150	569
Candida albicans	14435	7806	5755	3939	2443	1069	465
Capronia coronata	15890	9066	6530	4454	2688	1189	510
Capronia epimyces	16973	9577	6831	4721	2855	1237	566
Capronia semiimmersa	19962	11350	8042	5446	3310	1473	708
Cladophialophora bantiana	18554	10352	7356	4925	2935	1294	577
Cladophialophora carrionii	16518	9279	6638	4554	2710	1178	584
Cladophialophora immunda	23311	13510	9523	6304	3734	1654	764
Cladophialophora psammophila	19623	11077	7771	5221	3113	1373	599
Cladophialophora yegresii	16053	9115	6676	4619	2721	1190	605
Coccidioides immitis	15995	8132	5808	4006	2333	1035	465
Colletotrichum graminicola	20570	11248	7931	5182	2979	1286	669
Colletotrichum higginsianum	24376	12996	9025	5943	3397	1470	727
Coniosporium apollinis	16742	9598	6744	4558	2699	1198	541
Cryptococcus gattii	12663	7179	5388	3603	2064	846	396
Cryptococcus neoformans	15798	8995	6700	4501	2650	1085	473
Edhazardia aedis	3992	2356	1826	1299	724	354	154
Encephalitozoon cuniculi	4028	2258	1695	1173	718	317	158
Encephalitozoon intestinalis	3494	1938	1398	1002	605	276	129
Exophiala aquamarina	19565	11201	7850	5247	3160	1378	640
Exophiala mesophila	20112	11581	8378	5594	3303	1420	603
Exophiala oligosperma	23044	13594	9710	6700	3895	1702	770
Exophiala sideris	19690	11477	8108	5475	3283	1526	709
Exophiala spinifera	18624	10588	7361	4971	2930	1315	569
Exophiala xenobiotica	22202	12624	8842	5916	3556	1587	733
Fonsecaea multimorphosa	19242	10903	7638	5098	3017	1364	633
Fonsecaea pedrosoi	19423	10910	7684	5149	3029	1372	666

Table 4.3, continued.

Species	All	50	60	70	80	90	95
Fusarium graminearum	21126	11309	7881	5241	3067	1276	597
Fusarium oxysporum	47730	26857	18893	12589	7308	3202	1665
Fusarium verticillioides	40544	22388	15807	10656	6185	2694	1378
Gaeumannomyces graminis	20420	11412	7989	5359	3062	1284	619
Geomyces destructans	16133	8866	6489	4378	2562	1111	496
Histoplasma capsulatum	15693	9031	6588	4344	2434	985	452
Magnaporthe oryzae	21009	11351	7790	5150	2875	1279	596
Magnaporthe poae	18021	9991	6981	4634	2595	1136	522
Microbotryum violaceum	16715	9598	6972	4705	2791	1215	502
Microsporum canis	19136	10381	7334	4907	2826	1186	526
Microsporum gypseum	17969	9761	6994	4772	2767	1149	513
Mucor circinelloides	35097	19944	15093	10853	6616	2748	1294
Nematocida parisii	3390	1949	1388	871	507	194	78
Nematocida sp1	3612	2009	1382	866	488	175	76
Neosartorya fischeri	19647	10881	7856	5400	3146	1331	585
Neurospora crassa	15830	8889	6338	4266	2442	1109	549
Nosema ceranae	3516	2024	1547	1065	633	289	149
Ochroconis gallopava	21308	12252	8869	5993	3496	1618	664
Paracoccidioides brasiliensis	14159	8021	5947	4125	2415	1085	484
Paracoccidioides sp.	14480	8161	6048	4149	2441	1093	502
Phaeosphaeria nodorum	18002	9908	7080	4749	2740	1231	585
Phialophora europaea	17614	9924	7094	4675	2785	1203	545
Pneumocystis carinii	10333	5853	4407	3088	1894	864	379
Pneumocystis jirovecii	9473	5538	4147	2885	1788	840	379
Pneumocystis murina	9542	5442	4055	2874	1749	788	369
Puccinia graminis	17331	9588	6670	4618	2767	1200	526
Puccinia striiformis	19052	10664	7658	5135	3053	1320	574
Puccinia triticina	18987	10429	7501	5193	3136	1290	580
Pyrenophora tritici-repentis	18343	9684	6790	4558	2622	1181	540
Rhinocladiella mackenzie	18234	10428	7503	4978	2920	1317	610
Rhizopus delemar	40475	23477	17336	11871	6951	2934	1410
Schizosaccharomyces cryophilus	13416	7596	5897	4235	2754	1242	546
Schizosaccharomyces japonicus	13779	7714	5899	4237	2618	1138	494
Schizosaccharomyces octosporus	13222	7505	5743	4165	2646	1193	547
Schizosaccharomyces pombe	13686	7679	5878	4285	2652	1144	505
Sclerotinia sclerotiorum	16972	9423	6874	4553	2652	1227	552
Spizellomyces punctatus	24966	13796	10108	7135	4330	1852	789
Sporothrix schenckii	16407	9227	6747	4503	2600	1114	522

Table 4.3, continued.

Species	**All**	**50**	**60**	**70**	**80**	**90**	**95**
Trichophyton equinum	17279	9165	6569	4461	2541	1087	541
Trichophyton interdigitale	19086	10159	7251	4803	2738	1164	527
Trichophyton rubrum	25423	13778	9983	6856	3716	1562	739
Trichophyton tonsurans	17353	9280	6594	4522	2587	1137	553
Trichophyton verrucosum	405	212	179	156	96	31	13
Ustilago maydis	14166	8101	5756	3930	2266	946	438
Vavraia culicis	3650	2123	1641	1172	696	291	153
Verticillium alfalfae	16629	9082	6383	4220	2249	956	459
Verticillium dahliae	17948	9928	7030	4752	2606	1112	521
Vittaforma corneae	5068	3037	2136	1429	761	304	108

Homology searches against pathogenicity-related databases, using protein sequences from 86 species of fungi is the very first step of the identification of candidate common pathogenic genes. This identifies homologous sequences between each species of fungi against the three databases used in this study, establishing datasets of homologous pathogenic protein sequences for each species. These sequences will be pooled and by using the unique identifier for each of the database entries common pathogenic protein sequences can then be identified, across multiple species. The results from BLAST alignments are consistent when performed against all three fungal pathogenicity-related databases as visualized in Figure 4.1, 4.2, and 4.3 where species of fungi from either the phylum of Basidiomycota or Ascomycota show high numbers of BLAST hits against all three different databases. The number of BLAST hits reduces steadily as the percent identity criterion was increased to create a stringent, high confidence dataset to work with for the downstream data.

4.2 Homology Search against Carbohydrate-active Enzymes Database (CAZy)

BLASTP searches against CAZy yielded results as stated in Table 4.1, listing BLASTP hits to CAZy with a filtering criterion of 80% maximum identity. This was visualized in Figure 4.1. By clustering the different fungal species into phyla Ascomycota and Basidiomycota have higher number of homologs compared to fungal species. On average each fungal species has approximately 1196 hits against CAZy at 80% Identity, with the highest count belongs to the species *Fusarium oxyporum* a member of phylum Ascomycota, with a total of 5,292 hits. Species from the phylum of Microsporidia all display low homologous count against the CAZy.

Homologous sequences with 80% Identity from each species against CAZy is then extracted and compared between all 86 species and homologous sequences that are found among 80% of 86 fungal species were identified as a Common Fungal Pathogenicitiy-related Genes.

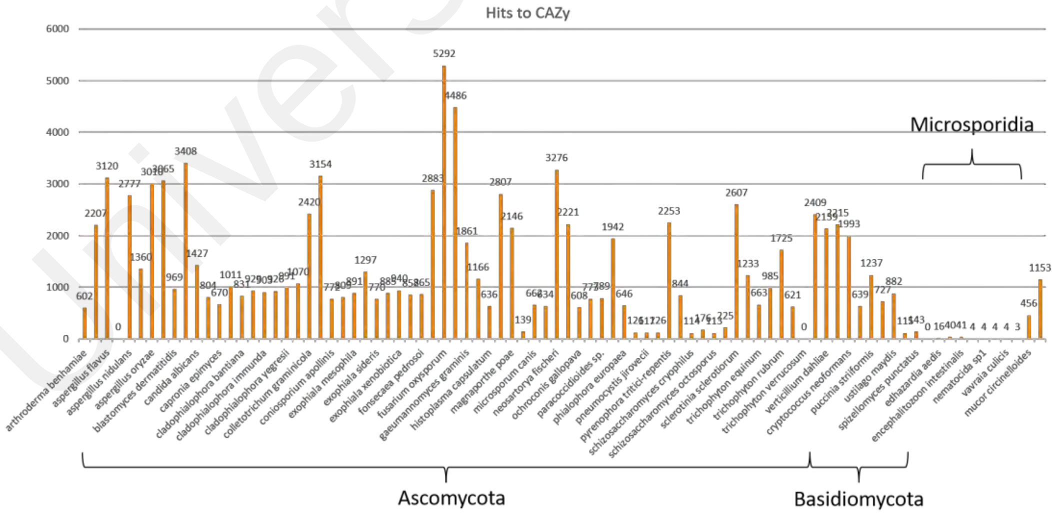


Figure 4.1: Homologous Gene Count against CAZy across Species

4.3 Homology Search against Pathogen-Host Interaction Database (PHI-base)

BLASTP searches against Pathogen-Host Interaction Database yielded results as stated in Table 4.2, listing BLASTP hits to PHI-base with a filtering criterion of 80% max identity. This was visualized in Figure 4.2. By clustering the different fungal species into phyla, Ascomycota and Basidiomycota has higher number of homologs compared to fungal species from other phyla. On average each fungal species has approximately 397 hits against PHI-base at 80% Identity, with the highest count belongs to the species *Fusarium oxyporum* a member of phylum Ascomycota with a total hits of 1,927. Species from the phylum of Microsporidia all display low homologous count against the PHI-base.

Homologous sequences with 80% Identity from each species against PHI-base then extracted and compared between all 86 species and homologous sequences that are found among 80% of 86 fungal species identified as a Common Fungal Pathogenicitiy-related Genes.

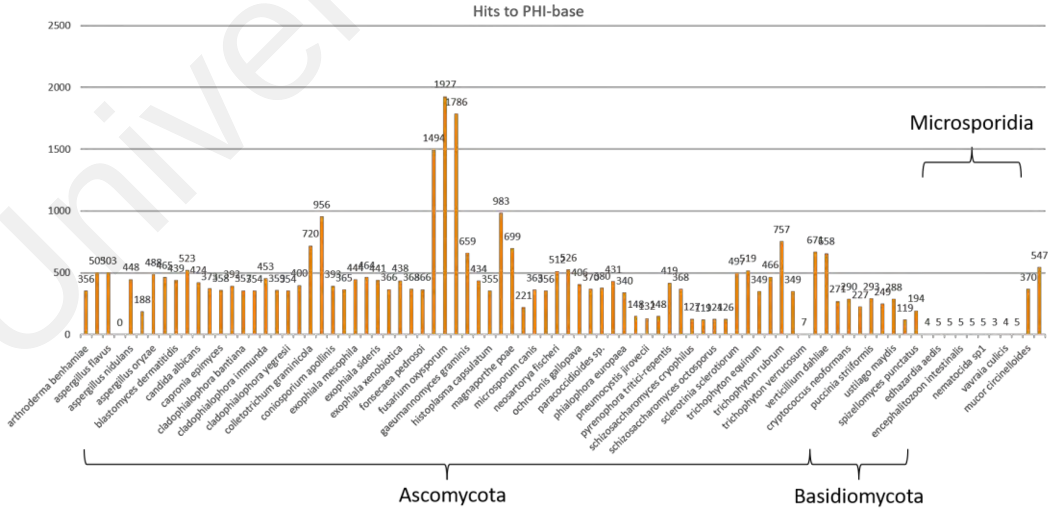


Figure 4.2: Homologous Gene Count against PHI-base across Species

4.4 Homology Search against Database of Fungal Virulence Factor (DFVF)

BLASTP searches against DFVF yielded results as stated in Table 4.3, listing BLASTP hits to DFVF with a filtering criterion of 80% max identity. This was visualized in Figure 4.3. By clustering the different fungal species into phyla, Ascomycota and Basidiomycota has higher number of homologs compared to fungal species from other phyla. On average each fungal species has approximately 329 hits against DFVF at 80% Identity, with the highest count belongs to the species *Trichophyton rubrum* a member of phylum Ascomycota with a total hits of 1,087. Species from the phylum of Microsporidia all display low homologous count against the DFVF.

Homologous sequences with 80% Identity from each species against DFVF is then extracted and compared between all 86 species and homologous sequences that are found among 80% of 86 fungal species identified as a Common Fungal Pathogenicitiy-related Genes.

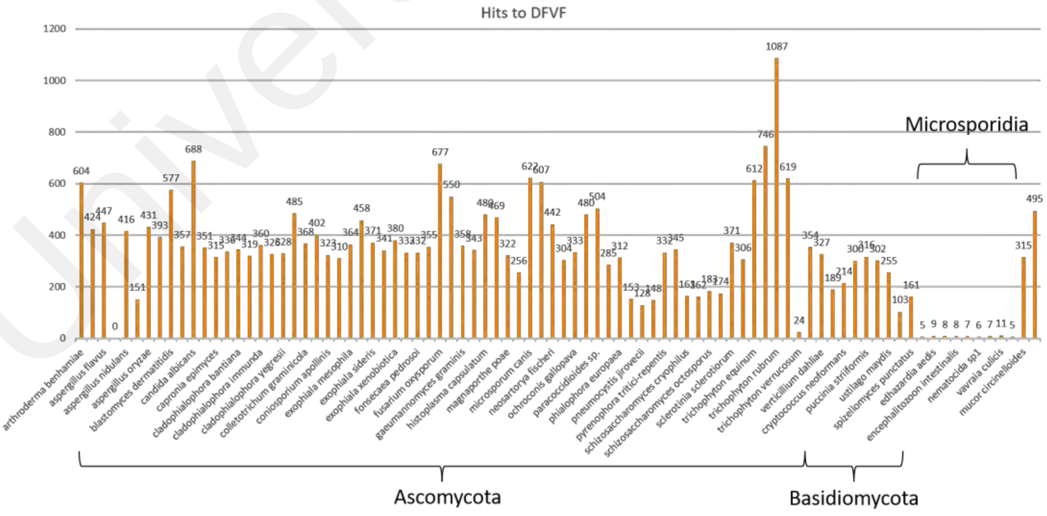
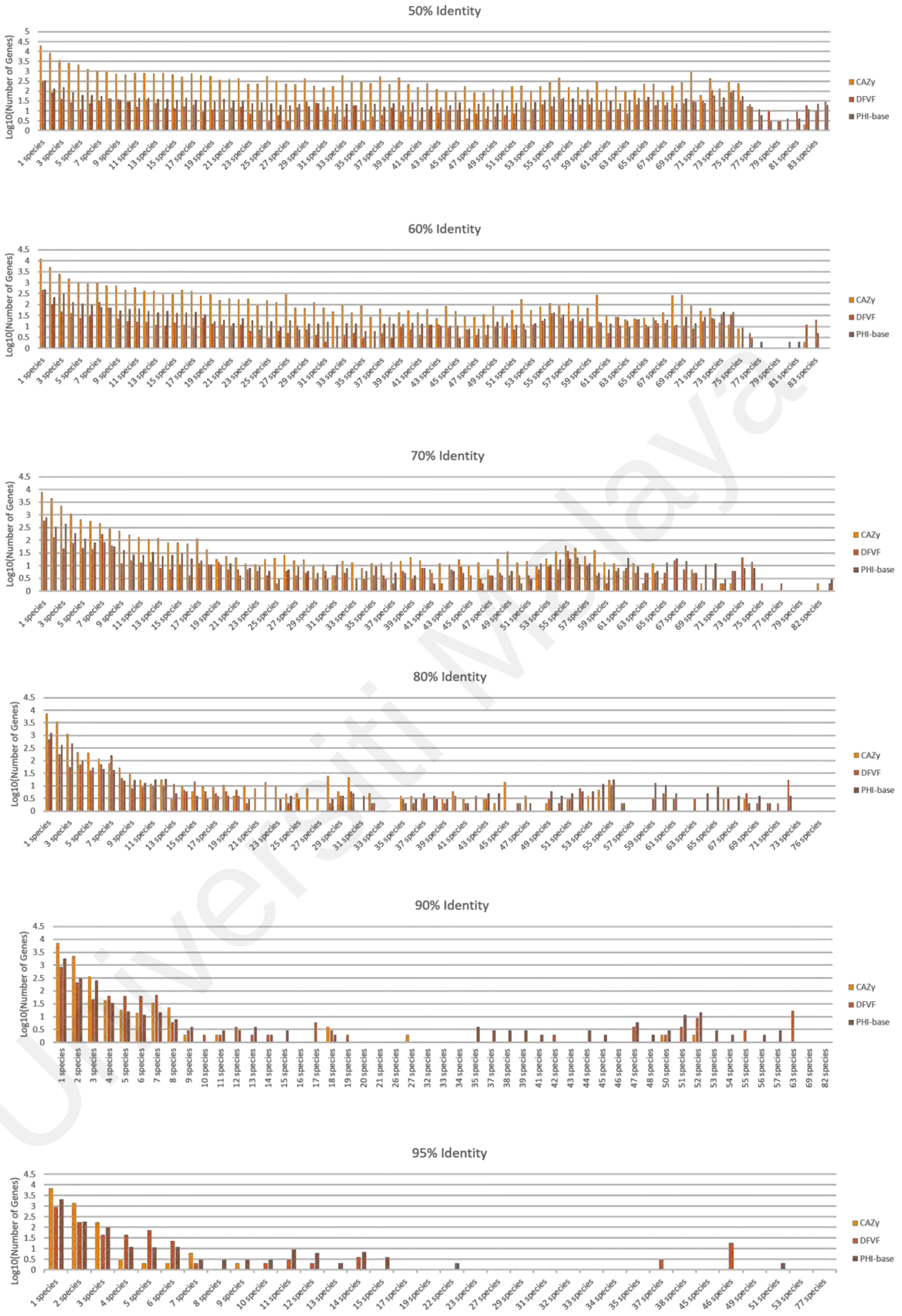


Figure 4.3: Homologous Gene Count against DFVF across Species

4.5 Common pathogenicity-related genes across different species

Sets of homologous protein sequences derived from initial homology searches to CAZy, PHI-base, and DFVF yielded sets of protein sequences that were found in different fungi from various phylum and species. Tabular BLASTP results for each of the protein sequences from the three databases are then processed with developed shell scripts to identify the presence of these protein sequences across the different species via a unique sequence identifier. Firstly, the genes were pooled into genes identified with different % identity, and each set of results was then classified to number of species (n) each protein sequence was found. Six different % identity profiles were classified, starting from 50% identity with an interval of 10% up to 90% and the most stringent criterion of % identity at 95% and these are represented by Figure 4.4.



50% Identity

60% Identity

70% Identity

80% Identity

90% Identity

95% Identity

Figure 4.4: Homologous Pathogenicity-related Protein Sequences with 50%, 60%, 70%, 80%, 90%, 95% Identity across 86 Fungal species

In summary, the 80 % identify criterion (80% sequence identity and 80% coverage across species) was classified and identified as high confidence conserved pathogenic genes in Table 4.6. the number of candidate conserved pathogenic genes cross species mapping to CAZy, PHI-base, DFVF were 8, 20, and 31 respectively and is listed in Table 4.4. These Common Fungal Pathogenicity-related Genes were then further studied by extracting the corresponding sequences from each species of fungi and subjecting the sequences to multiple sequence alignment, phylogenetics analysis and identification of Single Nucleotide Polymorphisms. Identified genes will also be uploaded to the Common Fungal Pathogenicity-related Gene Database Portal for public access to the data.

Table 4.4: Number of High Confidence Conserved Pathogenic Genes across 80% coverage.

Database	CAZy	PHI-base	DFVF
Number of Conserved Pathogenic Genes (based on 80% identity and e-value of 1e-5	8	20	31

These are extremely positive results, confirming the conservation of pathogenicity-related genes across different phyla and multiple species and providing a foundation for further study and understanding of pathogenicity genes conservation in fungus and how this understanding can be utilized to expand methodology in diagnostic and therapy.

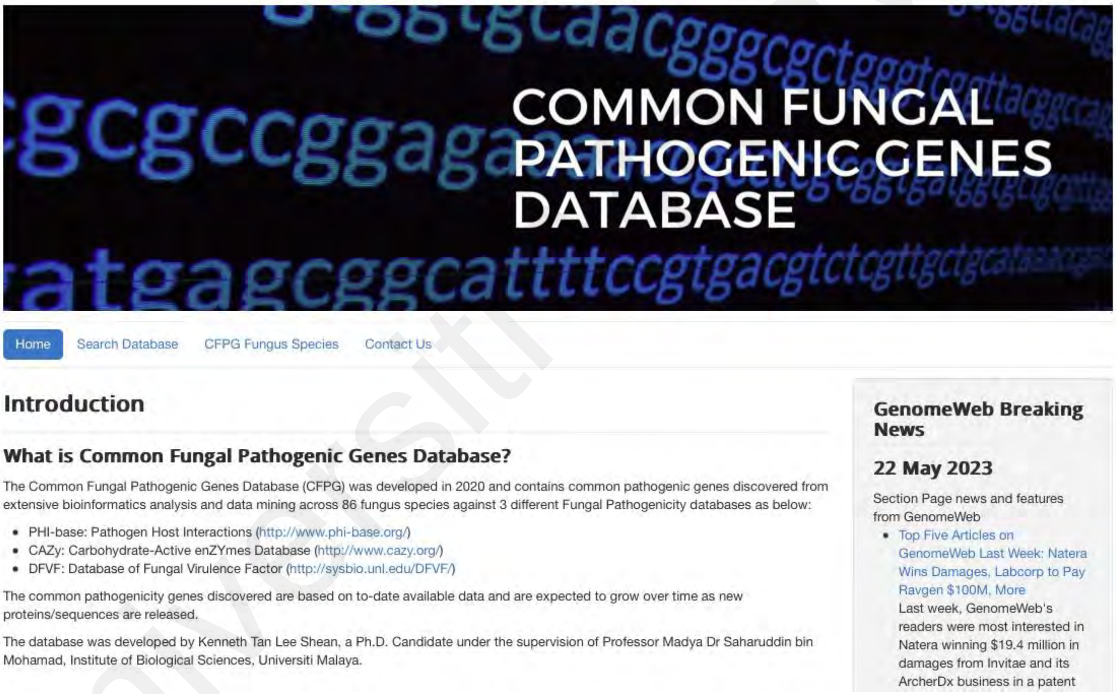
4.6 Common Fungal Pathogenicity-related Gene Database Application

All Homologous Gene extracted from homology searches were assigned with a unique CFPG Identifier and supplemented with additional information before they were uploaded into the MySQL database. Table 4.5 shows the number of records uploaded for each of the tables created. All entries for each table are listed in Appendix C, D, and E.

Table 4.5: Number of Rows Uploaded to CFPG.

Table	# Of Records
MASTER_FUNGUS	86
MASTER_COMMON_GENE	59
GENE_SPECIES_MAPPING	4135

The Home tab in Figure 4.5 display general introduction about the Common Fungal Pathogenicity-related Gene Database with an RSS Feed displayed from GenomeWeb (GenomeWeb, 2021), a reputable genomics news site. The database portal is accessible via https://cfpg.leapomics.com.



Introduction

What is Common Fungal Pathogenic Genes Database?

The Common Fungal Pathogenic Genes Database (CFPG) was developed in 2020 and contains common pathogenic genes discovered from extensive bioinformatics analysis and data mining across 86 fungus species against 3 different Fungal Pathogenicity databases as below:

- PHI-base: Pathogen Host Interactions (http://www.phi-base.org/)
- CAZy: Carbohydrate-Active enZYmes Database (http://www.cazy.org/)
- DFVF: Database of Fungal Virulence Factor (http://sysbio.unl.edu/DFVF/)

The common pathogenicity genes discovered are based on to-date available data and are expected to grow over time as new proteins/sequences are released.

The database was developed by Kenneth Tan Lee Shean, a Ph.D. Candidate under the supervision of Professor Madya Dr Saharuddin bin Mohamad, Institute of Biological Sciences, Universiti Malaya.

GenomeWeb Breaking News

22 May 2023

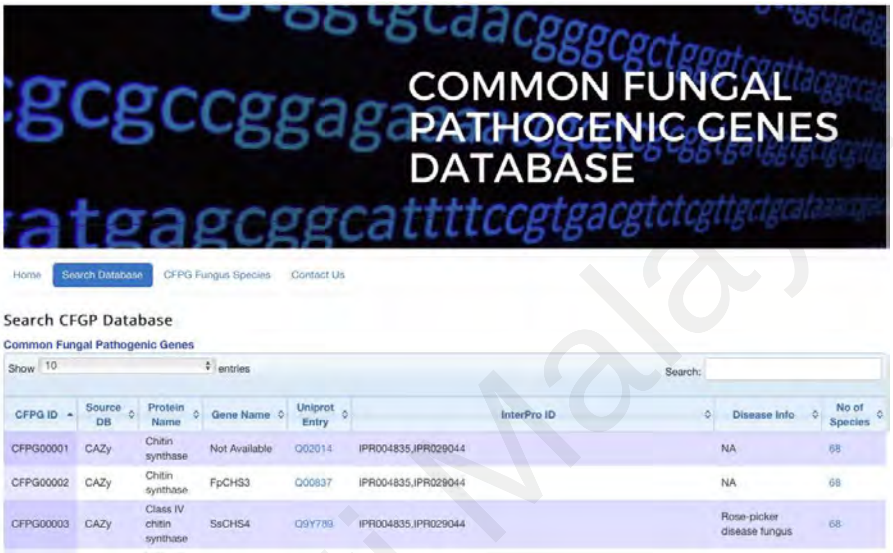
Section Page news and features from GenomeWeb

- Top Five Articles on GenomeWeb Last Week: Natera Wins Damages, Labcorp to Pay Ravgen $100M, More
  Last week, GenomeWeb's readers were most interested in Natera winning $19.4 million in damages from Invitae and its ArcherDx business in a patent

Figure 4.5: Home tab of the Common Fungal Pathogenicity-related Gene Database

Clicking the Search Database tab will display Common Fungal Pathogenicity-related Genes entries in multiple pages. This can be modified by selecting different paging options by selecting Show. To search for specific entries user can enter any search string (i.e. Protein Name/UniProt Entry/InterPro ID and etc) and the datatable will be filtered to display only records that fit entered string-pattern.

The number of species is an indication of how many species of fungi this entry was found in and clicking the hyperlink will lead to the detailed list of fungus for the entry. The hyperlink CFGPDB V1.0.xls allows download of all entries within the database. The different user interface of the CFPGDB can be seen in Figures 4.6, 4.7, 4.8, 4.9, and 4.10.



COMMON FUNGAL PATHOGENIC GENES DATABASE

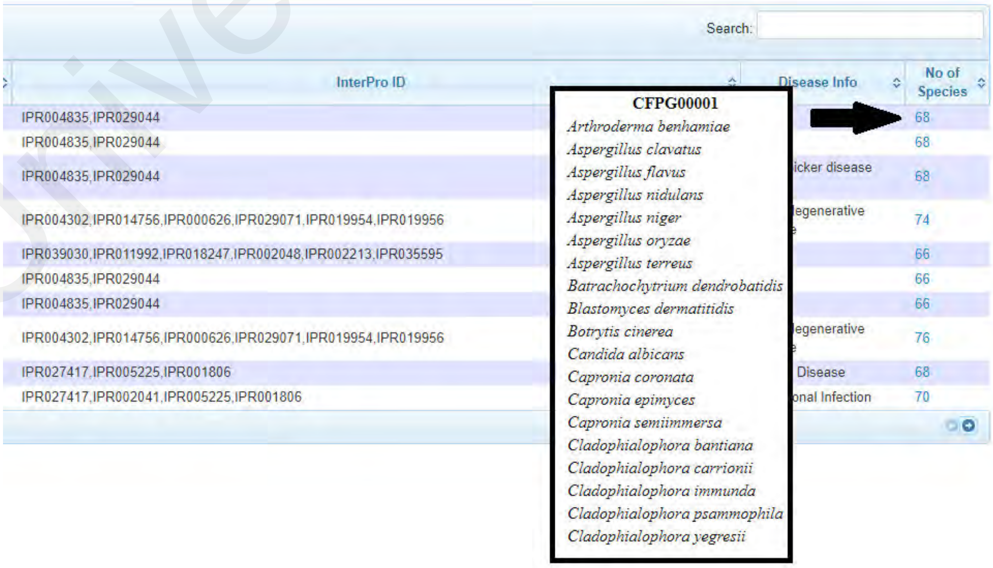
Search CFGP Database

Common Fungal Pathogenic Genes

Show 10 entries Search:

CFPG ID	Source DB	Protein Name	Gene Name	Uniprot Entry	InterPro ID	Disease Info	No of Species
CFPG00001	CAZy	Chitin synthase	Not Available	Q02014	IPR004835,IPR029044	NA	68
CFPG00002	CAZy	Chitin synthase	FpCHS3	Q00837	IPR004835,IPR029044	NA	68
CFPG00003	CAZy	Class IV chitin synthase	SsCHS4	Q9Y789	IPR004835,IPR029044	Rose-picker disease fungus	68

Figure 4.6: Search Database tab of the Common Fungal Pathogenicity-related Gene Database



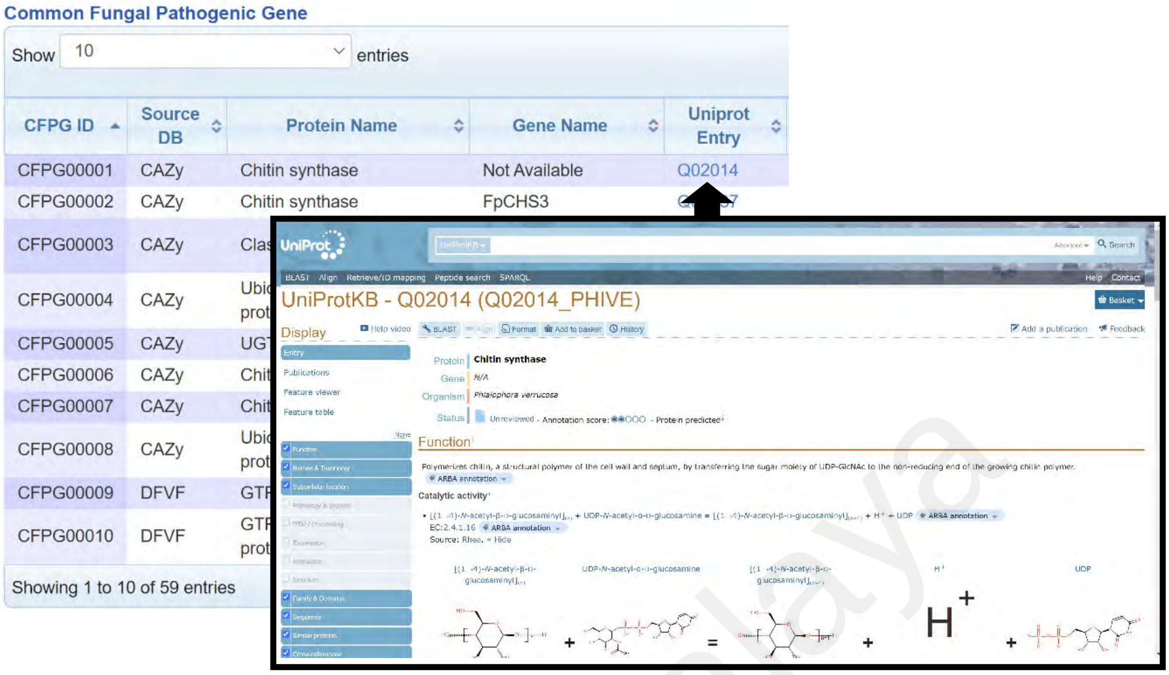
Search:

InterPro ID	Disease Info	No of Species
IPR004835,IPR029044		68
IPR004835,IPR029044		68
IPR004835,IPR029044	icker disease	68
IPR004302,IPR014756,IPR000626,IPR029071,IPR019954,IPR019956	legenerative e	74
IPR039030,IPR011992,IPR018247,IPR002048,IPR002213,IPR035595		66
IPR004835,IPR029044		66
IPR004835,IPR029044		66
IPR004302,IPR014756,IPR000626,IPR029071,IPR019954,IPR019956	legenerative e	76
IPR027417,IPR005225,IPR001806	Disease	68
IPR027417,IPR002041,IPR005225,IPR001806	onal Infection	70

CFPG00001

Arthroderma benhamiae
Aspergillus clavatus
Aspergillus flavus
Aspergillus nidulans
Aspergillus niger
Aspergillus oryzae
Aspergillus terreus
Batrachochytrium dendrobatidis
Blastomyces dermatitidis
Botrytis cinerea
Candida albicans
Capronia coronata
Capronia epimyces
Capronia semiimmersa
Cladophialophora bantiana
Cladophialophora carrionii
Cladophialophora immunda
Cladophialophora psammophila
Cladophialophora yegresii

Figure 4.7: Hyperlink to List of Fungus for a specific entry



Common Fungal Pathogenic Gene

Show 10 entries

CFPG ID	Source DB	Protein Name	Gene Name	Uniprot Entry
CFPG00001	CAZy	Chitin synthase	Not Available	Q02014
CFPG00002	CAZy	Chitin synthase	FpCHS3	Q[illegible]7
CFPG00003	CAZy	Clas		
CFPG00004	CAZy	Ubic prot		
CFPG00005	CAZy	UG		
CFPG00006	CAZy	Chit		
CFPG00007	CAZy	Chit		
CFPG00008	CAZy	Ubic prot		
CFPG00009	DFVF	GTF		
CFPG00010	DFVF	GTF prot		

Showing 1 to 10 of 59 entries

UniProtKB - Q02014 (Q02014_PHIVE)

Protein: Chitin synthase

Gene: N/A

Organism: *Phialophora verrucosa*

Status: Unreviewed - Annotation score: ●●○○○ - Protein predicted

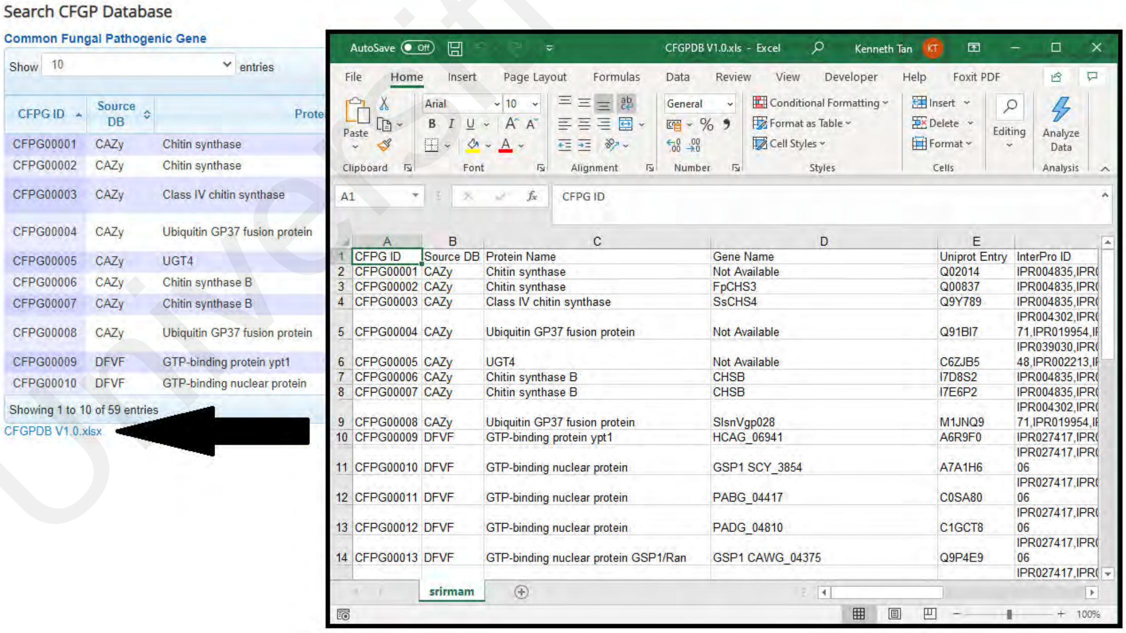
Function

Polymerizes chitin, a structural polymer of the cell wall and septum, by transferring the sugar moiety of UDP-GlcNAc to the non-reducing end of the growing chitin polymer. ARBA annotation

Catalytic activity

- [(1→4)-N-acetyl-β-D-glucosaminyl]$_n$ + UDP-N-acetyl-α-D-glucosamine = [(1→4)-N-acetyl-β-D-glucosaminyl]$_{n+1}$ + H$^+$ + UDP ARBA annotation
  EC:2.4.1.16 ARBA annotation
  Source: Rhea. « Hide

Figure 4.8: Hyperlink to UniProt



Search CFGP Database

Common Fungal Pathogenic Gene

Show 10 entries

CFPG ID	Source DB	Prote
CFPG00001	CAZy	Chitin synthase
CFPG00002	CAZy	Chitin synthase
CFPG00003	CAZy	Class IV chitin synthase
CFPG00004	CAZy	Ubiquitin GP37 fusion protein
CFPG00005	CAZy	UGT4
CFPG00006	CAZy	Chitin synthase B
CFPG00007	CAZy	Chitin synthase B
CFPG00008	CAZy	Ubiquitin GP37 fusion protein
CFPG00009	DFVF	GTP-binding protein ypt1
CFPG00010	DFVF	GTP-binding nuclear protein

Showing 1 to 10 of 59 entries

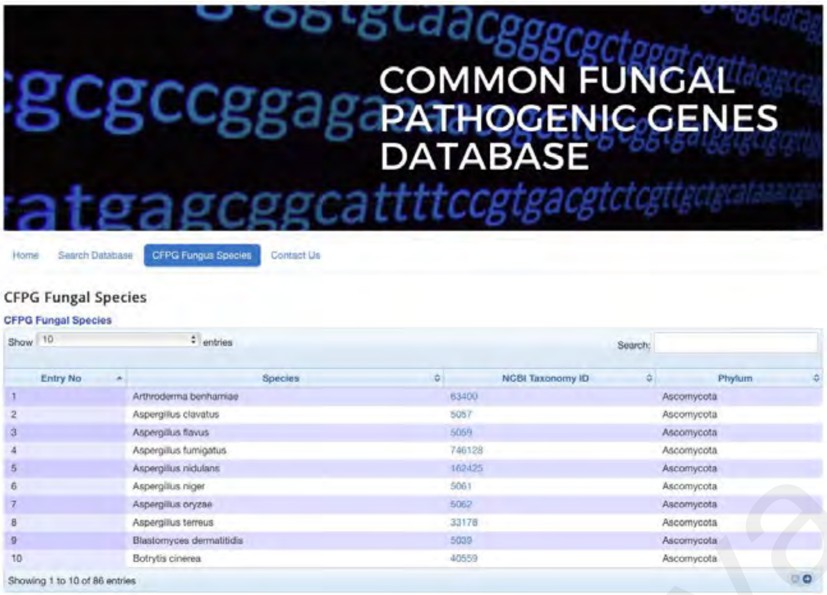
CFGPDB V1.0.xlsx

CFGPDB V1.0.xls - Excel

A1 CFPG ID

	A	B	C	D	E	
1	CFPG ID	Source DB	Protein Name	Gene Name	Uniprot Entry	InterPro ID
2	CFPG00001	CAZy	Chitin synthase	Not Available	Q02014	IPR004835,IPR(
3	CFPG00002	CAZy	Chitin synthase	FpCHS3	Q00837	IPR004835,IPR(
4	CFPG00003	CAZy	Class IV chitin synthase	SsCHS4	Q9Y789	IPR004835,IPR(
5	CFPG00004	CAZy	Ubiquitin GP37 fusion protein	Not Available	Q91BI7	IPR004302,IPR(71,IPR019954,IF
6	CFPG00005	CAZy	UGT4	Not Available	C6ZJB5	IPR039030,IPR(48,IPR002213,IF
7	CFPG00006	CAZy	Chitin synthase B	CHSB	I7D8S2	IPR004835,IPR(
8	CFPG00007	CAZy	Chitin synthase B	CHSB	I7E6P2	IPR004835,IPR(
9	CFPG00008	CAZy	Ubiquitin GP37 fusion protein	SlsnVgp028	M1JNQ9	IPR004302,IPR(71,IPR019954,IF
10	CFPG00009	DFVF	GTP-binding protein ypt1	HCAG_06941	A6R9F0	IPR027417,IPR(
11	CFPG00010	DFVF	GTP-binding nuclear protein	GSP1 SCY_3854	A7A1H6	IPR027417,IPR(06
12	CFPG00011	DFVF	GTP-binding nuclear protein	PABG_04417	C0SA80	IPR027417,IPR(06
13	CFPG00012	DFVF	GTP-binding nuclear protein	PADG_04810	C1GCT8	IPR027417,IPR(06
14	CFPG00013	DFVF	GTP-binding nuclear protein GSP1/Ran	GSP1 CAWG_04375	Q9P4E9	IPR027417,IPR(06
						IPR027417,IPR(

Figure 4.9: Hyperlink to download all entries.



CFPG Fungal Species

CFPG Fungal Species

Show 10 entries Search:

Entry No	Species	NCBI Taxonomy ID	Phylum
1	Arthroderma benhamiae	63400	Ascomycota
2	Aspergillus clavatus	5057	Ascomycota
3	Aspergillus flavus	5059	Ascomycota
4	Aspergillus fumigatus	746128	Ascomycota
5	Aspergillus nidulans	162425	Ascomycota
6	Aspergillus niger	5061	Ascomycota
7	Aspergillus oryzae	5062	Ascomycota
8	Aspergillus terreus	33178	Ascomycota
9	Blastomyces dermatitidis	5039	Ascomycota
10	Botrytis cinerea	40559	Ascomycota

Showing 1 to 10 of 86 entries

Figure 4.10: CFPG Fungal species tab.

4.7 SNP Mining through SNP-Sites

Homologs from 86 fungal species that passes the alignment criteria (i.e. 80% Identity and E-value of 10^{-5}) are extracted and aligned with MAFFT, and the subsequent output sequence alignment files are the used as input files for SNP-Sites. SNP-Sites generates a consensus reference sequence and extract variants from each sequence in the alignment file against the consensus reference sequence. The SNP mining identified large number of SNPs and are listed in Table 4.6 below.

Table 4.6: All 59 Common Fungal Pathogenic-Related Genes, Corresponding UniProt ID and Number of SNP Sites

CFPG ID	UniProt ID	Number of SNP Sites
CFPG00001	Q02014	8678
CFPG00002	Q00837	8534
CFPG00003	Q9Y789	8701
CFPG00004	Q91BI7	3451
CFPG00005	C6ZJB5	3017
CFPG00006	I7D8S2	8671
CFPG00007	I7E6P2	9125
CFPG00008	M1JNQ9	4374
CFPG00009	A6R9F0	6700

Table 4.6, continued.

CFPG ID	UniProt ID	Number of SNP Sites
CFPG00010	A7A1H6	3192
CFPG00011	C0SA80	3113
CFPG00012	C1GCT8	3113
CFPG00013	C4YIU6	3147
CFPG00014	C5GQ05	3194
CFPG00015	C5GS26	6700
CFPG00016	D2JLR3	2882
CFPG00017	D2JLR4	2882
CFPG00018	D2JLR5	2882
CFPG00019	D2JLR6	2882
CFPG00020	D2JLR7	2882
CFPG00021	D2JLR8	2882
CFPG00022	D2JLR9	2882
CFPG00023	D2JLS0	2882
CFPG00024	D2JLS1	2882
CFPG00025	D2JLS2	2882
CFPG00026	D2JLS3	2882
CFPG00027	D2JLS4	2882
CFPG00028	D2JLS5	2882
CFPG00029	D2JLS6	2882
CFPG00030	D2JLS7	2882
CFPG00031	D2JLS8	2882
CFPG00032	D2JLS9	2882
CFPG00033	F2QT01	3074
CFPG00034	HOG1	6827
CFPG00035	Q2PBY8	6317
CFPG00036	Q59P43	3147
CFPG00037	Q5ADS0	8399
CFPG00038	Q7Z7T9	6612
CFPG00039	Q96UM1	6776
CFPG00040	A0A0D2Y8P9	3120
CFPG00041	A1IVT7	6827

Table 4.6, continued.

CFPG ID	UniProt ID	Number of SNP Sites
CFPG00042	G4NC11	6684
CFPG00043	H9B3V9	3363
CFPG00044	I1RN81	6168
CFPG00045	I1S1V9	3612
CFPG00046	P41388	6405
CFPG00047	P53376	6153
CFPG00048	Q0U4L8	6647
CFPG00049	Q1KTF2	6423
CFPG00050	Q2PBY8	6317
CFPG00051	Q4HTT1	2501
CFPG00052	Q4WJS6	5808
CFPG00053	Q4WSF6	7081
CFPG00054	Q51MW4	3197
CFPG00055	Q5AND9	4020
CFPG00056	Q6QIY0	6010
CFPG00057	Q7Z7T9	6612
CFPG00058	Q8NJX2	6662
CFPG00059	T0LLS6	5726

SNPs that were discovered can serve as important biomarkers for diagnosis. By using these biomarkers it will enable identification of pathogenicity before or during the early stages of fuungal disease before it becomes too late for remediation. One observation from the results for CFPG entries from CFPG00016 to CFPG00032 is that they represent consistent number of SNPs discovered and this is due to the reason that the entries from the three different pathogenicity-related database maps to different UniProt entries. By looking at UniProt it was later confirmed that those entries are homologs found in different species of fungi within the genus of *Fusarium* but at the same time is found across different species of fungi when aligned through multiple sequence alignments.

4.8 Phylogenetic Analysis

Homologous nucleotide sequence of 59 Common Fungal Pathogenicity-related Genes were extracted from corresponding fungal species where homologous sequence is found and is further subjected to multiple sequence alignment using MAFFT. Multiple Sequence Alignment output was visualized using Unipro UGENE.

The phylogenetic tree building of the each of the 59 Common Fungal Pathogenicity-related Genes reflected a similar trend. The phylogenetic tree constructed resulted in unsurprising results, where members of the same species and phyla falling in the same clade. This same pattern is observed across all 59 entries in the Common Fungal Pathogenic-Related Gene Database which further confirming the hypothesis that pathogenicity-related genes are well conserved across different species of genus. Top four entries of the findings are further discussed in detailed while all multiple sequence alignments and phylogenetic tree diagram can be viewed in Appendix B.

Figure 4.11 shows the multiple sequence alignment result for CFPG00037 with the gene name UBI4 which codes for the protein Ubiquitin, which is involved in modification of proteins for proteasomal degradation and non-proteolytic functions (Finley, et al. 2012). The multiple sequence alignment of the homologs from across 82 of the 86 species where the high confidence homologs were identified (with 80% Identity) showed that the sequences are highly similar across the species with aligned perfectly. In Figure 4.12 the phylogenetics tree show clustering of the sequences from the parent node and with negligible distance < 1 due to high level for conservation.

Figure 4.11: Multiple Sequence Alignment for CFPG00037

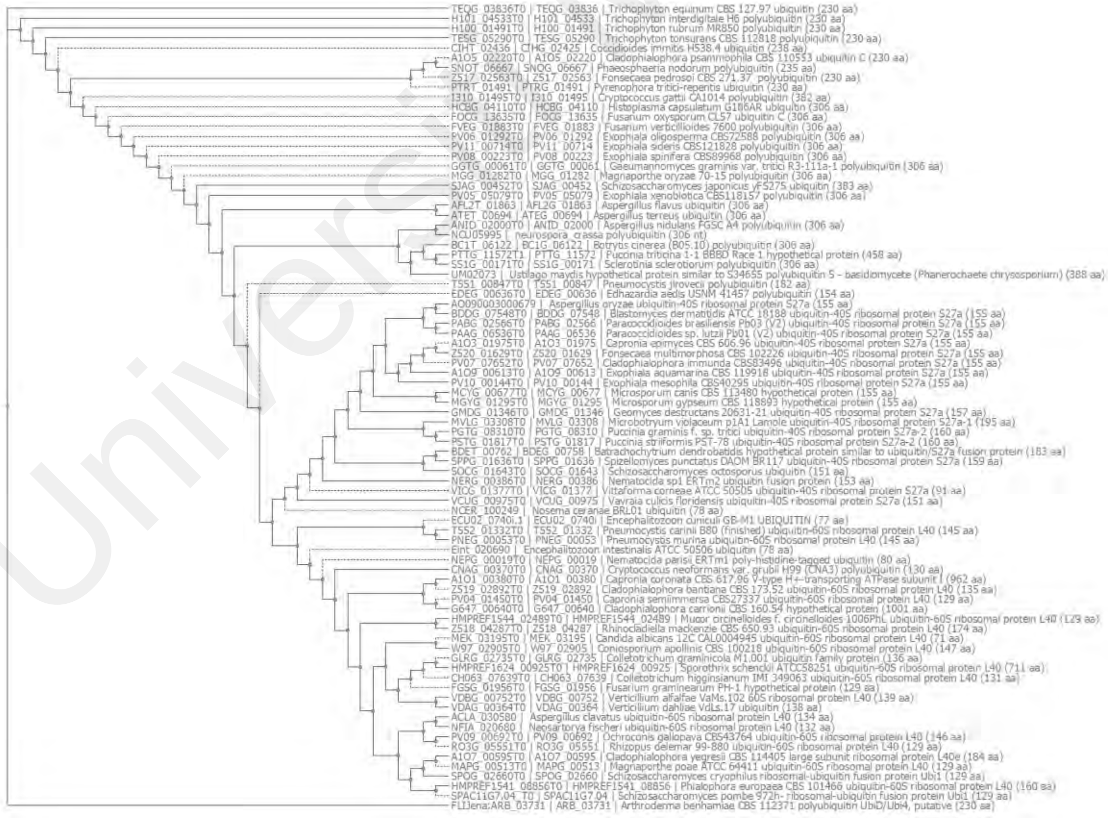

TEQG_03836T0 | TEQG_03836 | Trichophyton equinum CBS 127.97 ubiquitin (230 aa)
H101_04533T0 | H101_04533 | Trichophyton interdigitale H6 polyubiquitin (230 aa)
H100_01491T0 | H100_01491 | Trichophyton rubrum MR850 polyubiquitin (230 aa)
TESG_05290T0 | TESG_05290 | Trichophyton tonsurans CBS 112818 polyubiquitin (230 aa)
CIHT_02436 | CIHG_02425 | Coccidioides immitis H538.4 ubiquitin (238 aa)
A1O5_02220T0 | A1O5_02220 | Cladophialophora psammophila CBS 110553 ubiquitin C (230 aa)
SNOT_06667 | SNOG_06667 | Phaeosphaeria nodorum polyubiquitin (235 aa)
Z517_02563T0 | Z517_02563 | Fonsecaea pedrosoi CBS 271.37 polyubiquitin (230 aa)
PTRT_01491 | PTRG_01491 | Pyrenophora tritici-repentis ubiquitin (230 aa)
I310_01495T0 | I310_01495 | Cryptococcus gattii CA1014 polyubiquitin (382 aa)
HCBG_04110T0 | HCBG_04110 | Histoplasma capsulatum G186AR ubiquitin (306 aa)
FOCG_13635T0 | FOCG_13635 | Fusarium oxysporum CL57 ubiquitin C (306 aa)
FVEG_01883T0 | FVEG_01883 | Fusarium verticillioides 7600 polyubiquitin (306 aa)
PV06_01292T0 | PV06_01292 | Exophiala oligosperma CBS72588 polyubiquitin (306 aa)
PV11_00714T0 | PV11_00714 | Exophiala sideris CBS121828 polyubiquitin (306 aa)
PV08_00223T0 | PV08_00223 | Exophiala spinifera CBS89968 polyubiquitin (306 aa)
GGTG_00061T0 | GGTG_00061 | Gaeumannomyces graminis var. tritici R3-111a-1 polyubiquitin (306 aa)
MGG_01282T0 | MGG_01282 | Magnaporthe oryzae 70-15 polyubiquitin (306 aa)
SJAG_00452T0 | SJAG_00452 | Schizosaccharomyces japonicus yFS275 ubiquitin (383 aa)
PV05_05079T0 | PV05_05079 | Exophiala xenobiotica CBS118157 polyubiquitin (306 aa)
AFL2T_01863 | AFL2G_01863 | Aspergillus flavus ubiquitin (306 aa)
ATET_00694 | ATEG_00694 | Aspergillus terreus ubiquitin (306 aa)
ANID_02000T0 | ANID_02000 | Aspergillus nidulans FGSC A4 polyubiquitin (306 aa)
NCU05995 | neurospora crassa polyubiquitin (306 nt)
BC1T_06122 | BC1G_06122 | Botrytis cinerea (B05.10) polyubiquitin (306 aa)
PTTG_11572T1 | PTTG_11572 | Puccinia triticina 1-1 BBBD Race 1 hypothetical protein (458 aa)
SS1G_00171T0 | SS1G_00171 | Sclerotinia sclerotiorum polyubiquitin (306 aa)
UM02073 | Ustilago maydis hypothetical protein similar to S34655 polyubiquitin 5 - basidiomycete (Phanerochaete chrysosporium) (388 aa)
T551_00847T0 | T551_00847 | Pneumocystis jirovecii polyubiquitin (182 aa)
EDEG_00636T0 | EDEG_00636 | Edhazardia aedis USNM 41457 polyubiquitin (154 aa)
AO090003000679 | Aspergillus oryzae ubiquitin-40S ribosomal protein S27a (155 aa)
BDDG_07548T0 | BDDG_07548 | Blastomyces dermatitidis ATCC 18188 ubiquitin-40S ribosomal protein S27a (155 aa)
PABG_02566T0 | PABG_02566 | Paracoccidioides brasiliensis Pb03 (V2) ubiquitin-40S ribosomal protein S27a (155 aa)
PAAG_06536T0 | PAAG_06536 | Paracoccidioides sp. lutzii Pb01 (V2) ubiquitin-40S ribosomal protein S27a (155 aa)
A1O3_01975T0 | A1O3_01975 | Capronia epimyces CBS 606.96 ubiquitin-40S ribosomal protein S27a (155 aa)
Z520_01629T0 | Z520_01629 | Fonsecaea multimorphosa CBS 102226 ubiquitin-40S ribosomal protein S27a (155 aa)
PV07_07652T0 | PV07_07652 | Cladophialophora immunda CBS83496 ubiquitin-40S ribosomal protein S27a (155 aa)
A1O9_00613T0 | A1O9_00613 | Exophiala aquamarina CBS 119918 ubiquitin-40S ribosomal protein S27a (155 aa)
PV10_00144T0 | PV10_00144 | Exophiala mesophila CBS40295 ubiquitin-40S ribosomal protein S27a (155 aa)
MCYG_00677T0 | MCYG_00677 | Microsporum canis CBS 113480 hypothetical protein (155 aa)
MGYG_01295T0 | MGYG_01295 | Microsporum gypseum CBS 118893 hypothetical protein (155 aa)
GMDG_01346T0 | GMDG_01346 | Geomyces destructans 20631-21 ubiquitin-40S ribosomal protein S27a (157 aa)
MVLG_03308T0 | MVLG_03308 | Microbotryum violaceum p1A1 Lamole ubiquitin-40S ribosomal protein S27a-1 (195 aa)
PGTG_08310T0 | PGTG_08310 | Puccinia graminis f. sp. tritici ubiquitin-40S ribosomal protein S27a-2 (160 aa)
PSTG_01817T0 | PSTG_01817 | Puccinia striiformis PST-78 ubiquitin-40S ribosomal protein S27a-2 (160 aa)
BDET_00762 | BDEG_00758 | Batrachochytrium dendrobatidis hypothetical protein similar to ubiquitin/S27a fusion protein (183 aa)
SPPG_01636T0 | SPPG_01636 | Spizellomyces punctatus DAOM BR117 ubiquitin-40S ribosomal protein S27a (159 aa)
SOCG_01643T0 | SOCG_01643 | Schizosaccharomyces octosporus ubiquitin (151 aa)
NERG_00386T0 | NERG_00386 | Nematocida sp1 ERTm2 ubiquitin fusion protein (153 aa)
VICG_01377T0 | VICG_01377 | Vittaforma corneae ATCC 50505 ubiquitin-40S ribosomal protein S27a (91 aa)
VCUG_00975T0 | VCUG_00975 | Vavraia culicis floridensis ubiquitin-40S ribosomal protein S27a (151 aa)
NCER_100249 | Nosema ceranae BRL01 ubiquitin (78 aa)
ECU02_0740i.1 | ECU02_0740i | Encephalitozoon cuniculi GB-M1 UBIQUITIN (77 aa)
T552_01332T0 | T552_01332 | Pneumocystis carinii B80 (finished) ubiquitin-60S ribosomal protein L40 (145 aa)
PNEG_00053T0 | PNEG_00053 | Pneumocystis murina ubiquitin-60S ribosomal protein L40 (145 aa)
Eint_020690 | Encephalitozoon intestinalis ATCC 50506 ubiquitin (78 aa)
NEPG_00019T0 | NEPG_00019 | Nematocida parisii ERTm1 poly-histidine-tagged ubiquitin (80 aa)
CNAG_00370T0 | CNAG_00370 | Cryptococcus neoformans var. grubii H99 (CNA3) polyubiquitin (130 aa)
A1O1_00380T0 | A1O1_00380 | Capronia coronata CBS 617.96 V-type H+-transporting ATPase subunit I (962 aa)
Z519_02892T0 | Z519_02892 | Cladophialophora bantiana CBS 173.52 ubiquitin-60S ribosomal protein L40 (135 aa)
PV04_01450T0 | PV04_01450 | Capronia semiimmersa CBS27337 ubiquitin-60S ribosomal protein L40 (129 aa)
G647_00640T0 | G647_00640 | Cladophialophora carrionii CBS 160.54 hypothetical protein (1001 aa)
HMPREF1544_02489T0 | HMPREF1544_02489 | Mucor circinelloides f. circinelloides 1006PhL ubiquitin-60S ribosomal protein L40 (129 aa)
Z518_04287T0 | Z518_04287 | Rhinocladiella mackenzie CBS 650.93 ubiquitin-60S ribosomal protein L40 (174 aa)
MEK_03195T0 | MEK_03195 | Candida albicans 12C CAL0004945 ubiquitin-60S ribosomal protein L40 (71 aa)
W97_02905T0 | W97_02905 | Coniosporium apollinis CBS 100218 ubiquitin-60S ribosomal protein L40 (147 aa)
GLRG_02735T0 | GLRG_02735 | Colletotrichum graminicola M1.001 ubiquitin family protein (136 aa)
HMPREF1624_00925T0 | HMPREF1624_00925 | Sporothrix schenckii ATCC58251 ubiquitin-60S ribosomal protein L40 (711 aa)
CH063_07639T0 | CH063_07639 | Colletotrichum higginsianum IMI 349063 ubiquitin-60S ribosomal protein L40 (131 aa)
FGSG_01956T0 | FGSG_01956 | Fusarium graminearum PH-1 hypothetical protein (129 aa)
VDBG_00752T0 | VDBG_00752 | Verticillium alfalfae VaMs.102 60S ribosomal protein L40 (139 aa)
VDAG_00364T0 | VDAG_00364 | Verticillium dahliae VdLs.17 ubiquitin (138 aa)
ACLA_030580 | Aspergillus clavatus ubiquitin-60S ribosomal protein L40 (134 aa)
NFIA_020680 | Neosartorya fischeri ubiquitin-60S ribosomal protein L40 (132 aa)
PV09_00692T0 | PV09_00692 | Ochroconis gallopava CBS43764 ubiquitin-60S ribosomal protein L40 (146 aa)
RO3G_05551T0 | RO3G_05551 | Rhizopus delemar 99-880 ubiquitin-60S ribosomal protein L40 (129 aa)
A1O7_00595T0 | A1O7_00595 | Cladophialophora yegresii CBS 114405 large subunit ribosomal protein L40e (184 aa)
MAPG_00513T0 | MAPG_00513 | Magnaporthe poae ATCC 64411 ubiquitin-60S ribosomal protein L40 (129 aa)
SPOG_02660T0 | SPOG_02660 | Schizosaccharomyces cryophilus ribosomal-ubiquitin fusion protein Ubi1 (129 aa)
HMPREF1541_08856T0 | HMPREF1541_08856 | Phialophora europaea CBS 101466 ubiquitin-60S ribosomal protein L40 (160 aa)
SPAC11G7.04_T0 | SPAC11G7.04 | Schizosaccharomyces pombe 972h- ribosomal-ubiquitin fusion protein Ubi1 (129 aa)
FLIJena:ARB_03731 | ARB_03731 | Arthroderma benhamiae CBS 112371 polyubiquitin UbiD/Ubi4, putative (230 aa)

Figure 4.12: Phylogenetic Tree for CFPG00037

Figure 4.13 shows the multiple sequence alignment result for CFPG00008 with the gene name SlsnVgp028 which codes for the protein Ubiquitin GP37 fusion protein, which is also involved in modification of proteins for proteasomal degradation and non-proteolytic functions (Finley, et al. 2012). The multiple sequence alignment of the homologs from across 76 of the 86 species where the high confidence homologs were identified (with 80% Identity) showed that the sequences are also highly conserved only with varying length of protein sequences across different species of fungi. In Figure 4.14 the phylogenetics tree is displaying similar clustering with three different parent nodes branching out to three clades. However, distinctions were not obvious as distances between the parent nodes are also relatively low.

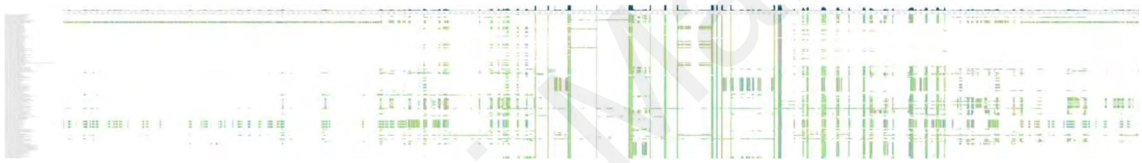


Figure 4.13: Multiple Sequence Alignment for CFPG00008

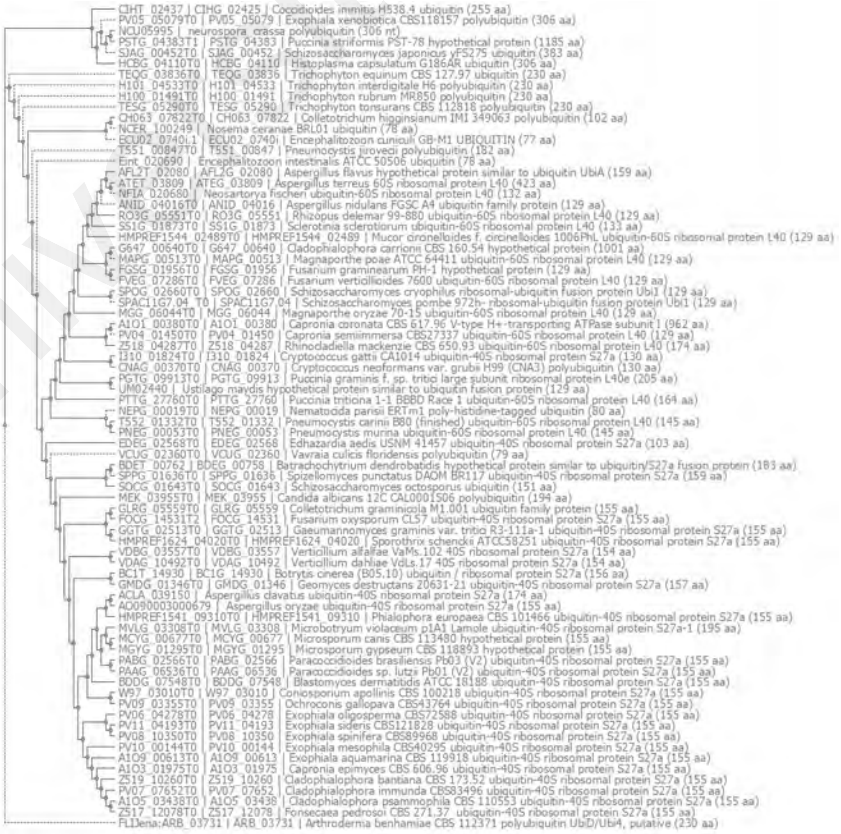

CIHT_02437 | CIHG_02425 | Coccidioides immitis H538.4 ubiquitin (255 aa)
PV05_05079T0 | PV05_05079 | Exophiala xenobiotica CBS118157 polyubiquitin (306 aa)
NCU05995 | neurospora_crassa polyubiquitin (306 nt)
PSTG_04383T1 | PSTG_04383 | Puccinia striiformis PST-78 hypothetical protein (1185 aa)
SJAG_00452T0 | SJAG_00452 | Schizosaccharomyces japonicus yFS275 ubiquitin (383 aa)
HCBG_04110T0 | HCBG_04110 | Histoplasma capsulatum G186AR ubiquitin (306 aa)
TEQG_03836T0 | TEQG_03836 | Trichophyton equinum CBS 127.97 ubiquitin (230 aa)
H101_04533T0 | H101_04533 | Trichophyton interdigitale H6 polyubiquitin (230 aa)
H100_01491T0 | H100_01491 | Trichophyton rubrum MR850 polyubiquitin (230 aa)
TESG_05290T0 | TESG_05290 | Trichophyton tonsurans CBS 112818 polyubiquitin (230 aa)
CH063_07822T0 | CH063_07822 | Colletotrichum higginsianum IMI 349063 polyubiquitin (102 aa)
NCER_100249 | Nosema ceranae BRL01 ubiquitin (78 aa)
ECU02_0740i.1 | ECU02_0740i | Encephalitozoon cuniculi GB-M1 UBIQUITIN (77 aa)
T551_00847T0 | T551_00847 | Pneumocystis jirovecii polyubiquitin (182 aa)
Eint_020690 | Encephalitozoon intestinalis ATCC 50506 ubiquitin (78 aa)
AFL2T_02080 | AFL2G_02080 | Aspergillus flavus hypothetical protein similar to ubiquitin UbiA (159 aa)
ATET_03809 | ATEG_03809 | Aspergillus terreus 60S ribosomal protein L40 (423 aa)
NFIA_020680 | Neosartorya fischeri ubiquitin-60S ribosomal protein L40 (132 aa)
ANID_04016T0 | ANID_04016 | Aspergillus nidulans FGSC A4 ubiquitin family protein (129 aa)
RO3G_05551T0 | RO3G_05551 | Rhizopus delemar 99-880 ubiquitin-60S ribosomal protein L40 (129 aa)
SS1G_01873T0 | SS1G_01873 | Sclerotinia sclerotiorum ubiquitin-60S ribosomal protein L40 (133 aa)
HMPREF1544_02489T0 | HMPREF1544_02489 | Mucor circinelloides f. circinelloides 1006PhL ubiquitin-60S ribosomal protein L40 (129 aa)
G647_00640T0 | G647_00640 | Cladophialophora carrionii CBS 160.54 hypothetical protein (1001 aa)
MAPG_00513T0 | MAPG_00513 | Magnaporthe poae ATCC 64411 ubiquitin-60S ribosomal protein L40 (129 aa)
FGSG_01956T0 | FGSG_01956 | Fusarium graminearum PH-1 hypothetical protein (129 aa)
FVEG_07286T0 | FVEG_07286 | Fusarium verticillioides 7600 ubiquitin-60S ribosomal protein L40 (129 aa)
SPOG_02660T0 | SPOG_02660 | Schizosaccharomyces cryophilus ribosomal-ubiquitin fusion protein Ubi1 (129 aa)
SPAC11G7.04_T0 | SPAC11G7.04 | Schizosaccharomyces pombe 972h- ribosomal-ubiquitin fusion protein Ubi1 (129 aa)
MGG_06044T0 | MGG_06044 | Magnaporthe oryzae 70-15 ubiquitin-60S ribosomal protein L40 (129 aa)
A1O1_00380T0 | A1O1_00380 | Capronia coronata CBS 617.96 V-type H+-transporting ATPase subunit I (962 aa)
PV04_01450T0 | PV04_01450 | Capronia semiimmersa CBS27337 ubiquitin-60S ribosomal protein L40 (129 aa)
Z518_04287T0 | Z518_04287 | Rhinocladiella mackenzie CBS 650.93 ubiquitin-60S ribosomal protein L40 (174 aa)
I310_01824T0 | I310_01824 | Cryptococcus gattii CA1014 ubiquitin-40S ribosomal protein S27a (130 aa)
CNAG_00370T0 | CNAG_00370 | Cryptococcus neoformans var. grubii H99 (CNA3) polyubiquitin (130 aa)
PGTG_09913T0 | PGTG_09913 | Puccinia graminis f. sp. tritici large subunit ribosomal protein L40e (205 aa)
UM02440 | Ustilago maydis hypothetical protein similar to ubiquitin fusion protein (129 aa)
PTTG_27760T0 | PTTG_27760 | Puccinia triticina 1-1 BBBD Race 1 ubiquitin-60S ribosomal protein L40 (164 aa)
NEPG_00019T0 | NEPG_00019 | Nematocida parisii ERTm1 poly-histidine-tagged ubiquitin (80 aa)
T552_01332T0 | T552_01332 | Pneumocystis carinii B80 (finished) ubiquitin-60S ribosomal protein L40 (145 aa)
PNEG_00053T0 | PNEG_00053 | Pneumocystis murina ubiquitin-60S ribosomal protein L40 (145 aa)
EDEG_02568T0 | EDEG_02568 | Edhazardia aedis USNM 41457 ubiquitin-40S ribosomal protein S27a (103 aa)
VCUG_02360T0 | VCUG_02360 | Vavraia culicis floridensis polyubiquitin (79 aa)
BDET_00762 | BDEG_00758 | Batrachochytrium dendrobatidis hypothetical protein similar to ubiquitin/S27a fusion protein (183 aa)
SPPG_01636T0 | SPPG_01636 | Spizellomyces punctatus DAOM BR117 ubiquitin-40S ribosomal protein S27a (159 aa)
SOCG_01643T0 | SOCG_01643 | Schizosaccharomyces octosporus ubiquitin (151 aa)
MEK_03955T0 | MEK_03955 | Candida albicans 12C CAL0001S06 polyubiquitin (194 aa)
GLRG_05559T0 | GLRG_05559 | Colletotrichum graminicola M1.001 ubiquitin family protein (155 aa)
FOCG_14531T2 | FOCG_14531 | Fusarium oxysporum CL57 ubiquitin-40S ribosomal protein S27a (155 aa)
GGTG_02513T0 | GGTG_02513 | Gaeumannomyces graminis var. tritici R3-111a-1 ubiquitin-40S ribosomal protein S27a (155 aa)
HMPREF1624_04020T0 | HMPREF1624_04020 | Sporothrix schenckii ATCC58251 ubiquitin-40S ribosomal protein S27a (155 aa)
VDBG_03557T0 | VDBG_03557 | Verticillium alfalfae VaMs.102 40S ribosomal protein S27a (154 aa)
VDAG_10492T0 | VDAG_10492 | Verticillium dahliae VdLs.17 40S ribosomal protein S27a (154 aa)
BC1T_14930 | BC1G_14930 | Botrytis cinerea (B05.10) ubiquitin / ribosomal protein S27a (156 aa)
GMDG_01346T0 | GMDG_01346 | Geomyces destructans 20631-21 ubiquitin-40S ribosomal protein S27a (157 aa)
ACLA_039150 | Aspergillus clavatus ubiquitin-40S ribosomal protein S27a (174 aa)
AO090003000679 | Aspergillus oryzae ubiquitin-40S ribosomal protein S27a (155 aa)
HMPREF1541_09310T0 | HMPREF1541_09310 | Phialophora europaea CBS 101466 ubiquitin-40S ribosomal protein S27a (155 aa)
MVLG_03308T0 | MVLG_03308 | Microbotryum violaceum p1A1 Lamole ubiquitin-40S ribosomal protein S27a-1 (195 aa)
MCYG_00677T0 | MCYG_00677 | Microsporum canis CBS 113480 hypothetical protein (155 aa)
MGYG_01295T0 | MGYG_01295 | Microsporum gypseum CBS 118893 hypothetical protein (155 aa)
PABG_02566T0 | PABG_02566 | Paracoccidioides brasiliensis Pb03 (V2) ubiquitin-40S ribosomal protein S27a (155 aa)
PAAG_06536T0 | PAAG_06536 | Paracoccidioides sp. lutzii Pb01 (V2) ubiquitin-40S ribosomal protein S27a (155 aa)
BDDG_07548T0 | BDDG_07548 | Blastomyces dermatitidis ATCC 18188 ubiquitin-40S ribosomal protein S27a (155 aa)
W97_03010T0 | W97_03010 | Coniosporium apollinis CBS 100218 ubiquitin-40S ribosomal protein S27a (155 aa)
PV09_03355T0 | PV09_03355 | Ochroconis gallopava CBS43764 ubiquitin-40S ribosomal protein S27a (155 aa)
PV06_04278T0 | PV06_04278 | Exophiala oligosperma CBS72588 ubiquitin-40S ribosomal protein S27a (155 aa)
PV11_04193T0 | PV11_04193 | Exophiala sideris CBS121828 ubiquitin-40S ribosomal protein S27a (155 aa)
PV08_10350T0 | PV08_10350 | Exophiala spinifera CBS89968 ubiquitin-40S ribosomal protein S27a (155 aa)
PV10_00144T0 | PV10_00144 | Exophiala mesophila CBS40295 ubiquitin-40S ribosomal protein S27a (155 aa)
A1O9_00613T0 | A1O9_00613 | Exophiala aquamarina CBS 119918 ubiquitin-40S ribosomal protein S27a (155 aa)
A1O3_01975T0 | A1O3_01975 | Capronia epimyces CBS 606.96 ubiquitin-40S ribosomal protein S27a (155 aa)
Z519_10260T0 | Z519_10260 | Cladophialophora bantiana CBS 173.52 ubiquitin-40S ribosomal protein S27a (155 aa)
PV07_07652T0 | PV07_07652 | Cladophialophora immunda CBS83496 ubiquitin-40S ribosomal protein S27a (155 aa)
A1O5_03438T0 | A1O5_03438 | Cladophialophora psammophila CBS 110553 ubiquitin-40S ribosomal protein S27a (155 aa)
Z517_12078T0 | Z517_12078 | Fonsecaea pedrosoi CBS 271.37 ubiquitin-40S ribosomal protein S27a (155 aa)
FLIJena:ARB_03731 | ARB_03731 | Arthroderma benhamiae CBS 112371 polyubiquitin UbiD/Ubi4, putative (230 aa)

Figure 4.14: Phylogenetic Tree for CFPG00008

Similarly, Figure 4.15 display the multiple sequence alignment for CFPG00004. The gene is a homolog for a yet to determined gene name with Q91BI7 UniProt ID which codes for the protein Ubiquitin GP37 fusion protein. This is like CFPG00008 as the gene product is also involved in modification of proteins for proteasomal degradation and non-proteolytic functions (Finley, et al. 2012). Multiple sequence alignment result and the phylogenetics tree construct were similar to CFPG00008 showing similarities and close relationships among the species of fungi found. See Appendix E for full list of mapping.

Figure 4.15: Multiple Sequence Alignment for CFPG00004

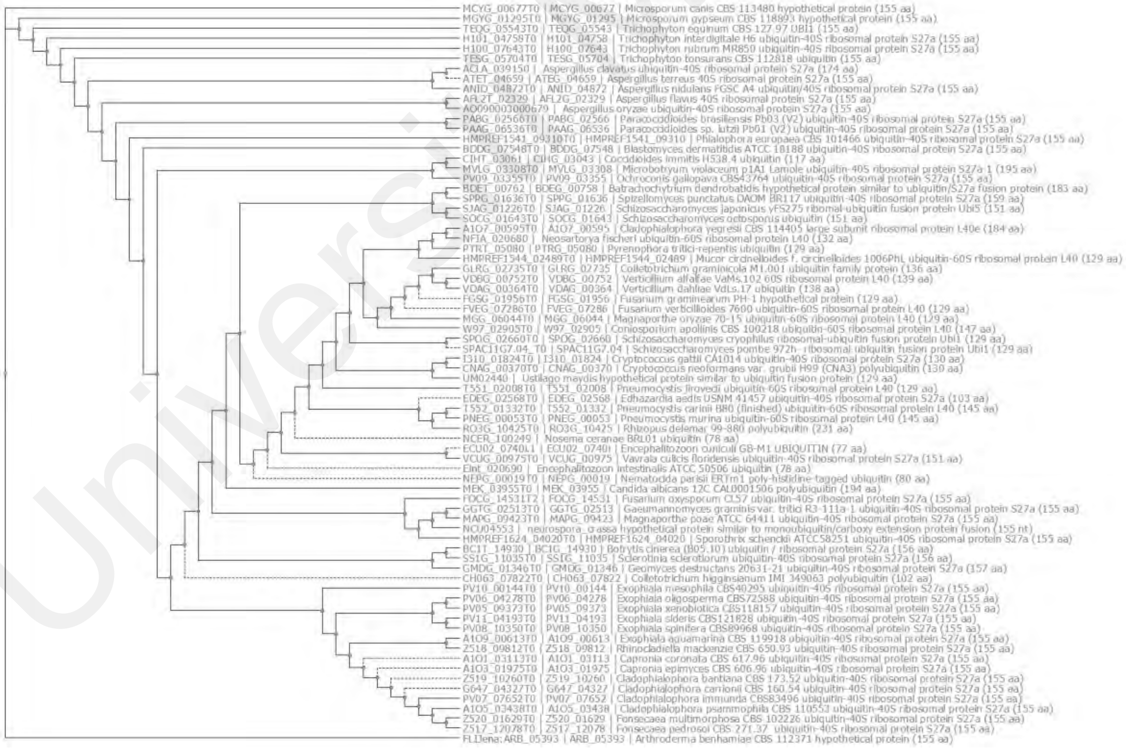

MCYG_00677T0 | MCYG_00677 | Microsporum canis CBS 113480 hypothetical protein (155 aa)
MGYG_01295T0 | MGYG_01295 | Microsporum gypseum CBS 118893 hypothetical protein (155 aa)
TEQG_05543T0 | TEQG_05543 | Trichophyton equinum CBS 127.97 UBI1 (155 aa)
H101_04758T0 | H101_04758 | Trichophyton interdigitale H6 ubiquitin-40S ribosomal protein S27a (155 aa)
H100_07643T0 | H100_07643 | Trichophyton rubrum MR850 ubiquitin-40S ribosomal protein S27a (155 aa)
TESG_05704T0 | TESG_05704 | Trichophyton tonsurans CBS 112818 ubiquitin (155 aa)
ACLA_039150 | Aspergillus clavatus ubiquitin-40S ribosomal protein S27a (174 aa)
ATET_04659 | ATEG_04659 | Aspergillus terreus 40S ribosomal protein S27a (155 aa)
ANID_04872T0 | ANID_04872 | Aspergillus nidulans FGSC A4 ubiquitin/40S ribosomal protein S27a (155 aa)
AFL2T_02329 | AFL2G_02329 | Aspergillus flavus 40S ribosomal protein S27a (155 aa)
AO090003000679 | Aspergillus oryzae ubiquitin-40S ribosomal protein S27a (155 aa)
PABG_02566T0 | PABG_02566 | Paracoccidioides brasiliensis Pb03 (V2) ubiquitin-40S ribosomal protein S27a (155 aa)
PAAG_06536T0 | PAAG_06536 | Paracoccidioides sp. lutzii Pb01 (V2) ubiquitin-40S ribosomal protein S27a (155 aa)
HMPREF1541_09310T0 | HMPREF1541_09310 | Phialophora europaea CBS 101466 ubiquitin-40S ribosomal protein S27a (155 aa)
BDDG_07548T0 | BDDG_07548 | Blastomyces dermatitidis ATCC 18188 ubiquitin-40S ribosomal protein S27a (155 aa)
CIHT_03061 | CIHG_03043 | Coccidioides immitis H538.4 ubiquitin (117 aa)
MVLG_03308T0 | MVLG_03308 | Microbotryum violaceum p1A1 Lamole ubiquitin-40S ribosomal protein S27a 1 (195 aa)
PV09_03355T0 | PV09_03355 | Ochroconis gallopava CBS43764 ubiquitin-40S ribosomal protein S27a (155 aa)
BDET_00762 | BDEG_00758 | Batrachochytrium dendrobatidis hypothetical protein similar to ubiquitin/S27a fusion protein (183 aa)
SPPG_01636T0 | SPPG_01636 | Spizellomyces punctatus DAOM BR117 ubiquitin-40S ribosomal protein S27a (159 aa)
SJAG_01226T0 | SJAG_01226 | Schizosaccharomyces japonicus yFS275 ribomal-ubiquitin fusion protein Ubi5 (151 aa)
SOCG_01643T0 | SOCG_01643 | Schizosaccharomyces octosporus ubiquitin (151 aa)
A1O7_00595T0 | A1O7_00595 | Cladophialophora yegresii CBS 114405 large subunit ribosomal protein L40e (184 aa)
NFIA_020680 | Neosartorya fischeri ubiquitin-60S ribosomal protein L40 (132 aa)
PTRT_05080 | PTRG_05080 | Pyrenophora tritici-repentis ubiquitin (129 aa)
HMPREF1544_02489T0 | HMPREF1544_02489 | Mucor circinelloides f. circinelloides 1006PhL ubiquitin-60S ribosomal protein L40 (129 aa)
GLRG_02735T0 | GLRG_02735 | Colletotrichum graminicola M1.001 ubiquitin family protein (136 aa)
VDBG_00752T0 | VDBG_00752 | Verticillium alfalfae VaMs.102 60S ribosomal protein L40 (139 aa)
VDAG_00364T0 | VDAG_00364 | Verticillium dahliae VdLs.17 ubiquitin (138 aa)
FGSG_01956T0 | FGSG_01956 | Fusarium graminearum PH-1 hypothetical protein (129 aa)
FVEG_07286T0 | FVEG_07286 | Fusarium verticillioides 7600 ubiquitin-60S ribosomal protein L40 (129 aa)
MGG_06044T0 | MGG_06044 | Magnaporthe oryzae 70-15 ubiquitin-60S ribosomal protein L40 (129 aa)
W97_02905T0 | W97_02905 | Coniosporium apollinis CBS 100218 ubiquitin-60S ribosomal protein L40 (147 aa)
SPOG_02660T0 | SPOG_02660 | Schizosaccharomyces cryophilus ribosomal-ubiquitin fusion protein Ubi1 (129 aa)
SPAC11G7.04_T0 | SPAC11G7.04 | Schizosaccharomyces pombe 972h- ribosomal ubiquitin fusion protein Ubi1 (129 aa)
I310_01824T0 | I310_01824 | Cryptococcus gattii CA1014 ubiquitin-40S ribosomal protein S27a (130 aa)
CNAG_00370T0 | CNAG_00370 | Cryptococcus neoformans var. grubii H99 (CNA3) polyubiquitin (130 aa)
UM02440 | Ustilago maydis hypothetical protein similar to ubiquitin fusion protein (129 aa)
T551_02008T0 | T551_02008 | Pneumocystis jirovecii ubiquitin-60S ribosomal protein L40 (129 aa)
EDEG_02568T0 | EDEG_02568 | Edhazardia aedis USNM 41457 ubiquitin-40S ribosomal protein S27a (103 aa)
T552_01332T0 | T552_01332 | Pneumocystis carinii B80 (finished) ubiquitin-60S ribosomal protein L40 (145 aa)
PNEG_00053T0 | PNEG_00053 | Pneumocystis murina ubiquitin-60S ribosomal protein L40 (145 aa)
RO3G_10425T0 | RO3G_10425 | Rhizopus delemar 99-880 polyubiquitin (231 aa)
NCER_100249 | Nosema ceranae BRL01 ubiquitin (78 aa)
ECU02_0740i1 | ECU02_0740i | Encephalitozoon cuniculi GB-M1 UBIQUITIN (77 aa)
VCUG_00975T0 | VCUG_00975 | Vavraia culicis floridensis ubiquitin-40S ribosomal protein S27a (151 aa)
EInt_020690 | Encephalitozoon intestinalis ATCC 50506 ubiquitin (78 aa)
NEPG_00019T0 | NEPG_00019 | Nematocida parisii ERTm1 poly-histidine-tagged ubiquitin (80 aa)
MEK_03955T0 | MEK_03955 | Candida albicans 12C CAL0001506 polyubiquitin (194 aa)
FOCG_14531T2 | FOCG_14531 | Fusarium oxysporum CL57 ubiquitin-40S ribosomal protein S27a (155 aa)
GGTG_02513T0 | GGTG_02513 | Gaeumannomyces graminis var. tritici R3-111a-1 ubiquitin-40S ribosomal protein S27a (155 aa)
MAPG_09423T0 | MAPG_09423 | Magnaporthe poae ATCC 64411 ubiquitin-40S ribosomal protein S27a (155 aa)
NCU04553 | neurospora_crassa hypothetical protein similar to monoubiquitin/carboxy extension protein fusion (155 nt)
HMPREF1624_04020T0 | HMPREF1624_04020 | Sporothrix schenckii ATCC58251 ubiquitin-40S ribosomal protein S27a (155 aa)
BC1T_14930 | BC1G_14930 | Botrytis cinerea (B05.10) ubiquitin / ribosomal protein S27a (156 aa)
SS1G_11035T0 | SS1G_11035 | Sclerotinia sclerotiorum ubiquitin-40S ribosomal protein S27a (156 aa)
GMDG_01346T0 | GMDG_01346 | Geomyces destructans 20631-21 ubiquitin-40S ribosomal protein S27a (157 aa)
CH063_07822T0 | CH063_07822 | Colletotrichum higginsianum IMI 349063 polyubiquitin (102 aa)
PV10_00144T0 | PV10_00144 | Exophiala mesophila CBS40295 ubiquitin-40S ribosomal protein S27a (155 aa)
PV06_04278T0 | PV06_04278 | Exophiala oligosperma CBS72588 ubiquitin-40S ribosomal protein S27a (155 aa)
PV05_09373T0 | PV05_09373 | Exophiala xenobiotica CBS118157 ubiquitin-40S ribosomal protein S27a (155 aa)
PV11_04193T0 | PV11_04193 | Exophiala sideris CBS121828 ubiquitin-40S ribosomal protein S27a (155 aa)
PV08_10350T0 | PV08_10350 | Exophiala spinifera CBS89968 ubiquitin-40S ribosomal protein S27a (155 aa)
A1O9_00613T0 | A1O9_00613 | Exophiala aquamarina CBS 119918 ubiquitin-40S ribosomal protein S27a (155 aa)
Z518_09812T0 | Z518_09812 | Rhinocladiella mackenzie CBS 650.93 ubiquitin-40S ribosomal protein S27a (155 aa)
A1O1_03113T0 | A1O1_03113 | Capronia coronata CBS 617.96 ubiquitin-40S ribosomal protein S27a (155 aa)
A1O3_01975T0 | A1O3_01975 | Capronia epimyces CBS 606.96 ubiquitin-40S ribosomal protein S27a (155 aa)
Z519_10260T0 | Z519_10260 | Cladophialophora bantiana CBS 173.52 ubiquitin-40S ribosomal protein S27a (155 aa)
G647_04327T0 | G647_04327 | Cladophialophora carrionii CBS 160.54 ubiquitin-40S ribosomal protein S27a (155 aa)
PV07_07652T0 | PV07_07652 | Cladophialophora immunda CBS83496 ubiquitin-40S ribosomal protein S27a (155 aa)
A1O5_03438T0 | A1O5_03438 | Cladophialophora psammophila CBS 110553 ubiquitin-40S ribosomal protein S27a (155 aa)
Z520_01629T0 | Z520_01629 | Fonsecaea multimorphosa CBS 102226 ubiquitin-40S ribosomal protein S27a (155 aa)
Z517_12078T0 | Z517_12078 | Fonsecaea pedrosoi CBS 271.37 ubiquitin-40S ribosomal protein S27a (155 aa)
FLIJena:ARB_05393 | ARB_05393 | Arthroderma benhamiae CBS 112371 hypothetical protein (155 aa)

Figure 4.16: Phylogenetic Tree for CFPG00004

Figure 4.17 shows the multiple sequence alignment result for CFPG00052 with the gene name AFUA_1G04950 which codes for the protein Serine/threonine-protein phosphatase. The protein has been found to play an important part in fungal pathogenicity in a study on pathogenicity of *Magnaporthe oryzae* (Du, et al. 2013) where it showed the deletion mutants of the gene failed to penetrate into host plant cells implying the pathogenicity impact of the protein in fungal pathogenicity. The multiple sequence alignment of the homologs from across 73 of the 86 species where the high confidence homologs were identified (with 80% Identity) showed that the sequences are highly conserved across majority of the species especially in the mid-region of the genes though variations are observed as well. In Figure 4.18 the phylogenetics tree show clustering of the sequences across species from different phyla in studies including the top two phylum where most members come from (i.e. Ascomycota/Basidiomycota).

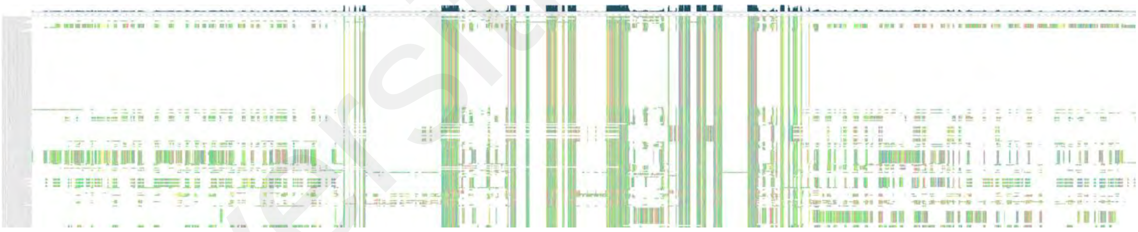


Figure 4.17: Multiple Sequence Alignment for CFPG00052

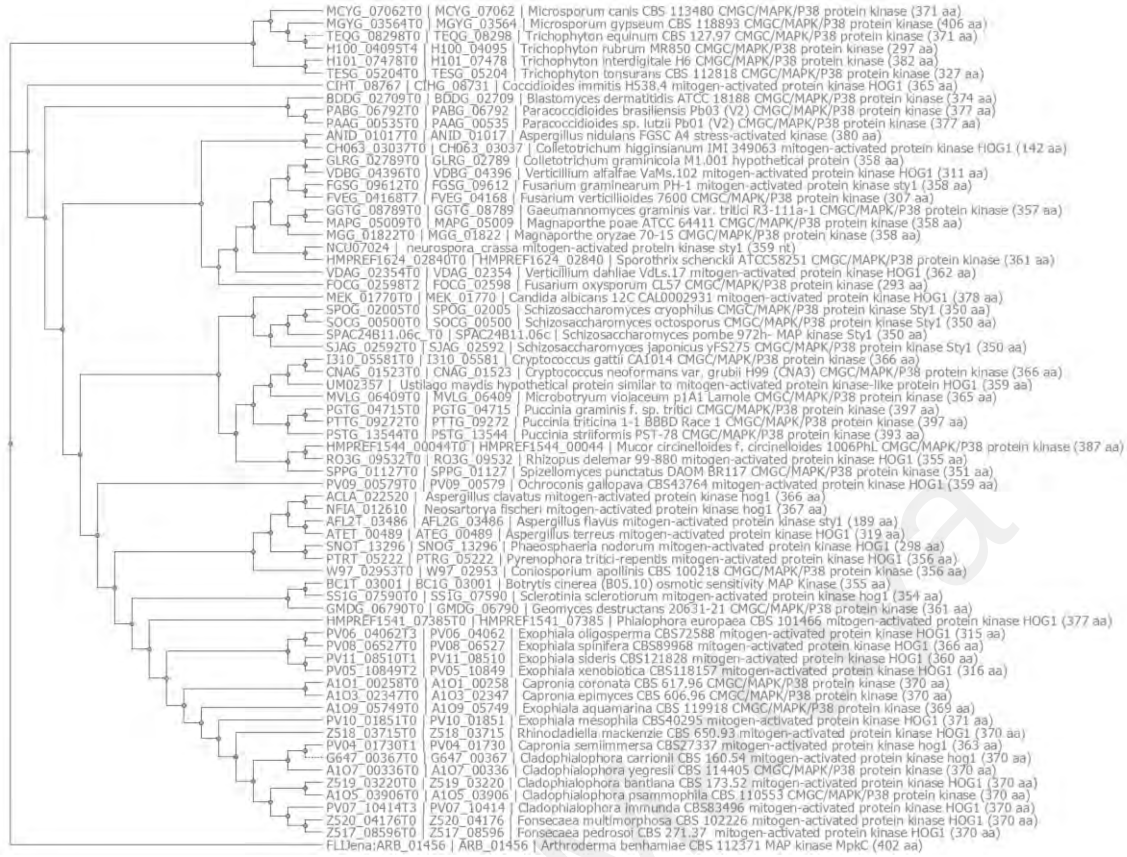

MCYG_07062T0 | MCYG_07062 | Microsporum canis CBS 113480 CMGC/MAPK/P38 protein kinase (371 aa)
MGYG_03564T0 | MGYG_03564 | Microsporum gypseum CBS 118893 CMGC/MAPK/P38 protein kinase (406 aa)
TEQG_08298T0 | TEQG_08298 | Trichophyton equinum CBS 127.97 CMGC/MAPK/P38 protein kinase (371 aa)
H100_04095T4 | H100_04095 | Trichophyton rubrum MR850 CMGC/MAPK/P38 protein kinase (297 aa)
H101_07478T0 | H101_07478 | Trichophyton interdigitale H6 CMGC/MAPK/P38 protein kinase (382 aa)
TESG_05204T0 | TESG_05204 | Trichophyton tonsurans CBS 112818 CMGC/MAPK/P38 protein kinase (327 aa)
CIHT_08767 | CIHG_08731 | Coccidioides immitis H538.4 mitogen-activated protein kinase HOG1 (365 aa)
BDDG_02709T0 | BDDG_02709 | Blastomyces dermatitidis ATCC 18188 CMGC/MAPK/P38 protein kinase (374 aa)
PABG_06792T0 | PABG_06792 | Paracoccidioides brasiliensis Pb03 (V2) CMGC/MAPK/P38 protein kinase (377 aa)
PAAG_00535T0 | PAAG_00535 | Paracoccidioides sp. lutzii Pb01 (V2) CMGC/MAPK/P38 protein kinase (377 aa)
ANID_01017T0 | ANID_01017 | Aspergillus nidulans FGSC A4 stress-activated kinase (380 aa)
CH063_03037T0 | CH063_03037 | Colletotrichum higginsianum IMI 349063 mitogen-activated protein kinase HOG1 (142 aa)
GLRG_02789T0 | GLRG_02789 | Colletotrichum graminicola M1.001 hypothetical protein (358 aa)
VDBG_04396T0 | VDBG_04396 | Verticillium alfalfae VaMs.102 mitogen-activated protein kinase HOG1 (311 aa)
FGSG_09612T0 | FGSG_09612 | Fusarium graminearum PH-1 mitogen-activated protein kinase sty1 (358 aa)
FVEG_04168T7 | FVEG_04168 | Fusarium verticillioides 7600 CMGC/MAPK/P38 protein kinase (307 aa)
GGTG_08789T0 | GGTG_08789 | Gaeumannomyces graminis var. tritici R3-111a-1 CMGC/MAPK/P38 protein kinase (357 aa)
MAPG_05009T0 | MAPG_05009 | Magnaporthe poae ATCC 64411 CMGC/MAPK/P38 protein kinase (358 aa)
MGG_01822T0 | MGG_01822 | Magnaporthe oryzae 70-15 CMGC/MAPK/P38 protein kinase (358 aa)
NCU07024 | neurospora_crassa mitogen-activated protein kinase sty1 (359 nt)
HMPREF1624_02840T0 | HMPREF1624_02840 | Sporothrix schenckii ATCC58251 CMGC/MAPK/P38 protein kinase (361 aa)
VDAG_02354T0 | VDAG_02354 | Verticillium dahliae VdLs.17 mitogen-activated protein kinase HOG1 (362 aa)
FOCG_02598T2 | FOCG_02598 | Fusarium oxysporum CL57 CMGC/MAPK/P38 protein kinase (293 aa)
MEK_01770T0 | MEK_01770 | Candida albicans 12C CAL0002931 mitogen-activated protein kinase HOG1 (378 aa)
SPOG_02005T0 | SPOG_02005 | Schizosaccharomyces cryophilus CMGC/MAPK/P38 protein kinase Sty1 (350 aa)
SOCG_00500T0 | SOCG_00500 | Schizosaccharomyces octosporus CMGC/MAPK/P38 protein kinase Sty1 (350 aa)
SPAC24B11.06c_T0 | SPAC24B11.06c | Schizosaccharomyces pombe 972h- MAP kinase Sty1 (350 aa)
SJAG_02592T0 | SJAG_02592 | Schizosaccharomyces japonicus yFS275 CMGC/MAPK/P38 protein kinase Sty1 (350 aa)
I310_05581T0 | I310_05581 | Cryptococcus gattii CA1014 CMGC/MAPK/P38 protein kinase (366 aa)
CNAG_01523T0 | CNAG_01523 | Cryptococcus neoformans var. grubii H99 (CNA3) CMGC/MAPK/P38 protein kinase (366 aa)
UM02357 | Ustilago maydis hypothetical protein similar to mitogen-activated protein kinase-like protein HOG1 (359 aa)
MVLG_06409T0 | MVLG_06409 | Microbotryum violaceum p1A1 Lamole CMGC/MAPK/P38 protein kinase (365 aa)
PGTG_04715T0 | PGTG_04715 | Puccinia graminis f. sp. tritici CMGC/MAPK/P38 protein kinase (397 aa)
PTTG_09272T0 | PTTG_09272 | Puccinia triticina 1-1 BBBD Race 1 CMGC/MAPK/P38 protein kinase (397 aa)
PSTG_13544T0 | PSTG_13544 | Puccinia striiformis PST-78 CMGC/MAPK/P38 protein kinase (393 aa)
HMPREF1544_00044T0 | HMPREF1544_00044 | Mucor circinelloides f. circinelloides 1006PhL CMGC/MAPK/P38 protein kinase (387 aa)
RO3G_09532T0 | RO3G_09532 | Rhizopus delemar 99-880 mitogen-activated protein kinase HOG1 (355 aa)
SPPG_01127T0 | SPPG_01127 | Spizellomyces punctatus DAOM BR117 CMGC/MAPK/P38 protein kinase (351 aa)
PV09_00579T0 | PV09_00579 | Ochroconis gallopava CBS43764 mitogen-activated protein kinase HOG1 (359 aa)
ACLA_022520 | Aspergillus clavatus mitogen-activated protein kinase hog1 (366 aa)
NFIA_012610 | Neosartorya fischeri mitogen-activated protein kinase hog1 (367 aa)
AFL2T_03486 | AFL2G_03486 | Aspergillus flavus mitogen-activated protein kinase sty1 (189 aa)
ATET_00489 | ATEG_00489 | Aspergillus terreus mitogen-activated protein kinase HOG1 (319 aa)
SNOT_13296 | SNOG_13296 | Phaeosphaeria nodorum mitogen-activated protein kinase HOG1 (298 aa)
PTRT_05222 | PTRG_05222 | Pyrenophora tritici-repentis mitogen-activated protein kinase HOG1 (356 aa)
W97_02953T0 | W97_02953 | Coniosporium apollinis CBS 100218 CMGC/MAPK/P38 protein kinase (356 aa)
BC1T_03001 | BC1G_03001 | Botrytis cinerea (B05.10) osmotic sensitivity MAP Kinase (355 aa)
SS1G_07590T0 | SS1G_07590 | Sclerotinia sclerotiorum mitogen-activated protein kinase hog1 (354 aa)
GMDG_06790T0 | GMDG_06790 | Geomyces destructans 20631-21 CMGC/MAPK/P38 protein kinase (361 aa)
HMPREF1541_07385T0 | HMPREF1541_07385 | Phialophora europaea CBS 101466 mitogen-activated protein kinase HOG1 (377 aa)
PV06_04062T3 | PV06_04062 | Exophiala oligosperma CBS72588 mitogen-activated protein kinase HOG1 (315 aa)
PV08_06527T0 | PV08_06527 | Exophiala spinifera CBS89968 mitogen-activated protein kinase HOG1 (366 aa)
PV11_08510T1 | PV11_08510 | Exophiala sideris CBS121828 mitogen-activated protein kinase HOG1 (360 aa)
PV05_10849T2 | PV05_10849 | Exophiala xenobiotica CBS118157 mitogen-activated protein kinase HOG1 (316 aa)
A1O1_00258T0 | A1O1_00258 | Capronia coronata CBS 617.96 CMGC/MAPK/P38 protein kinase (370 aa)
A1O3_02347T0 | A1O3_02347 | Capronia epimyces CBS 606.96 CMGC/MAPK/P38 protein kinase (370 aa)
A1O9_05749T0 | A1O9_05749 | Exophiala aquamarina CBS 119918 CMGC/MAPK/P38 protein kinase (369 aa)
PV10_01851T0 | PV10_01851 | Exophiala mesophila CBS40295 mitogen-activated protein kinase HOG1 (371 aa)
Z518_03715T0 | Z518_03715 | Rhinocladiella mackenzie CBS 650.93 mitogen-activated protein kinase HOG1 (370 aa)
PV04_01730T1 | PV04_01730 | Capronia semiimmersa CBS27337 mitogen-activated protein kinase hog1 (363 aa)
G647_00367T0 | G647_00367 | Cladophialophora carrionii CBS 160.54 mitogen-activated protein kinase hog1 (370 aa)
A1O7_00336T0 | A1O7_00336 | Cladophialophora yegresii CBS 114405 CMGC/MAPK/P38 protein kinase (370 aa)
Z519_03220T0 | Z519_03220 | Cladophialophora bantiana CBS 173.52 mitogen-activated protein kinase HOG1 (370 aa)
A1O5_03906T0 | A1O5_03906 | Cladophialophora psammophila CBS 110553 CMGC/MAPK/P38 protein kinase (370 aa)
PV07_10414T3 | PV07_10414 | Cladophialophora immunda CBS83496 mitogen-activated protein kinase HOG1 (370 aa)
Z520_04176T0 | Z520_04176 | Fonsecaea multimorphosa CBS 102226 mitogen-activated protein kinase HOG1 (370 aa)
Z517_08596T0 | Z517_08596 | Fonsecaea pedrosoi CBS 271.37 mitogen-activated protein kinase HOG1 (370 aa)
FLIJena:ARB_01456 | ARB_01456 | Arthroderma benhamiae CBS 112371 MAP kinase MpkC (402 aa)

Figure 4.18: Phylogenetic Tree for CFPG00052

4.9 Summary of Results

The Fungal Pathogenic-Related Genes Comparative Pipeline had successfully identified candidate common fungal pathogenicity genes across fungal species from different genus and phylum, supporting the hypothesis of the research that pathogenicity genes are highly conserved across fungal species across multiple genus and phylum. Through multiple sequence alignment, SNP mining and phylogenetics analysis of the common pathogenicity related genes further supports the observation that these pathogenicity-related genes are well-conserved through different fungal species. Although there were outliers in the result, these outliers do not impact the overall observation and results of the conservation of pathogenicity genes.

CHAPTER 5

DISCUSSION

5.1 Genomics Diversity and Relationship of Pathogenic Fungus through Comparative Genomics.

This study aimed to study fungal pathogenicity from a broader perspective, to expand and continue to fill the gap of knowledge in understanding fungal pathogenicity through comparative genomics, by leveraging on both raw and curated sequence data to build a comparative genomics pipeline specifically for fungal pathogenicity and creating a database portal that allow access to the identified candidate pathogenicity-related genes. Throughout the study and analysis there were many findings that worth discussing and henceforth as detailed in the following subsections.

5.1.1 Pathogenic Fungus Genomics Diversity through Homology Searches using Protein Sequences.

Homology searches compares either nucleotide sequences or protein sequences. Comparing protein sequences provides higher level of resolution for homology search. Nucleotide searches can produce different combination of triplet codons, which could potentially be translated to the same amino acid. Hence by using protein sequences, this can avoid any translational and or transcriptional impact for the eventual protein structure (Nature Education, 2014)). As mutation occurs naturally throughout the lifetime and through generations of a particular organism, the underlying changes in nucleotides does not contribute to vast phenotypic changes.

Using nucleotide sequences for homology searches could contribute to higher percentage of false negative results and therefore homologs could be missed (Pearson, 2013). Using protein sequences for homology searches yield higher sensitivity than nucleotide sequence comparison, thus picking up more homologs that would otherwise be missed by nucleotide sequence similarity search. Homology searches using protein sequences is more useful when searching for conserved protein-coding genes across organisms that has higher variability in genomic sequences as it is more targeted focusing only on protein coding regions and the variability in the coded proteins.

E-value and Percent Identity were the parameters used to determine the best candidate homologs from the BLAST sequence alignment results of protein sequences of the 86 fungal species against all three fungal pathogenicity-related database. These were common parameters to determine if two sequences have high degree of similarities in which inferring homology and suggests evolutionary relationship between the organisms in study (Pearson, 2013). The study utilizes both Percent Identity and E-value for a better inference of homologous relationships between sequences as usage of Precent Identity alone would produce false negatives, and the stringent criteria used in the study (i.e., Percent Identity of 80%, E-value of 10^{-5}), along with an additional criteria where the candidate Pathogenicity-Related Genes are found in at least 80% (68 of 86) of all fungal species studied here results in high confidence candidate Common Fungal Pathogenicity-related Genes.

The homology searches allow identification of common fungal pathogenicity-related genes by aligning them against various verified databases. This study shows that fungal pathogenicity genes are generally well conserved across the kingdom of fungus, regardless of which member of phylum or species the fungus belongs to. The conservation is apparent looking at the multiple sequence alignment results of the identified common fungal pathogenicity-related genes and based on the results, SNPs were identified and expected species that belongs to the same phylum are situated closer than the rest.

5.1.2 Inter-Phylum Fungus Comparative Genomics Pipeline

Most fungus genomics study focus on comparing genomics sequences from different isolates from a certain species of fungi or among closely related fungal species trying to understand various essential features such as insights into fungus lifestyle (Knapp, et al. 2018) and understanding of genomics properties of a certain fungal species. The datasets chosen for the study was 86 species of fungi that the Fungal Genome Initiative had collected and sequenced that portrayed the importance of their existence for applications development in medicine, agriculture, and industry (Broad Institute, 2014). There are other databases that contains more fungal genome sequences such as FungiDB (Basenko, et al. 2018) which contains sequence information for 186 fungal species across different phyla, regardless of the pathogenic significance of the each of the species hence instead of using all sequences in FungiDB, this study focused on studying fungal species that have pathogenicity significance to a range of hosts thus the selection was made to utilize data from the Fungal Genome Initiative.

The 86 fungal species in this study comprise of fungi from different phyla ranging from Ascomycota, Basidiomycota, Chytridiomycota, Microsporidia, and Mucormycotina. These different species of fungi all share a common trait where all 86 of the compared fungi have various level of pathogenicity properties by living on other organisms. These host organisms range from plants, humans, and animals (Refer to Table 3.1). From the entire list of fungi, it was identified that while comparing to available databases namely PHI-base (Urban, 2017), CAZy (Lombard, 2014) and DFVF (Lu, et al. 2012) species of fungi belonging to the phylum of Ascomycota and Basidiomycota have more homologous hits to the databases than the rest of the phyla, and among the 86 species of fungi, 61 are known to live on and or infect animal or human hosts, with the remaining 25 fungal species are known to live on and or infect plants. The advantage of identifying conserved pathogenic genes across different phyla of fungal species is that it allows for discovery of broad-spectrum antifungal agents or broad-spectrum diagnostic tools as effective PCR primers can be designed for detection and identification of pathogenic fungi or other specific isolates of fungus as done by other studies (Lee, et al. 2008).

Across the different species of fungi although the fungus host ranges from animal, human and plant, unsurprisingly there is a large amount of overlapping pathogenicity-related genes among the phylum of Ascomycota and Basidiomycota, both Dikarya. This also aligns with the general observation reported by Dean et al. (2012) that had summarized the top 10 fungal pathogens in molecular plant pathology whereby all 10 of the shortlisted fungal species came from either the phylum of Ascomycota or Basidiomycota, as listed in Table 2.1. Ascomycetes are also most represented in this study, which does not come as a surprise as the phylum is the largest in the Kingdom of Fungi (Watkinson, et al. 2015) and the study also includes members of other phyla in the Kingdom of Fungi such as Chytridiomycetes, Microsporidia, and Mucoromycotina, showed low homologous count to all three fungal pathogenicity-related gene databases. This may be due to the databases that were utilized contains little to no pathogenicity data from species of these three phyla as they are generally less represented in the databases.

This observation seems to be consistent with the continuous effort to understand the diversity of the Kingdom of Fungi. A study by Choi & Kim (2017) attempted to construct phylogenetics relationship by comparing whole-genome data. The results from the study showed that there are only three major groups namely Monokarya, Basidiomycota, and Ascomycota (Petersen, 2013). Monokaryotic fungus which includes Chytridiomycetes does not produce dikayons during the life cycles thus has high level of variability in the mechanisms of infections. Microsporidia on the other hand are a group of spore-forming unicellular organisms and infect range of hosts including human and is identified as a basal branch of the fungi or as a sister group (Han, et al. 2020) hence it is also not a surprise that the member of the species from Microsporidia showed very low level of homology to the pathogenicity-related databases. As fungal pathogenicity is highly associated with the life cycle of the species of fungi it is vital to understand pathogenicity from the angle of the life cycle.

5.1.3 Comparison between Animal Fungal Pathogenicity and Plant Fungal Pathogenicity

The results from the comparative genomics study showed that although fungal pathogen generally shares high similarities in genes composition across different fungi, it was observed that the type of fungal pathogen hosts has different mechanism of pathogenicity, thus the presence of pathogenic genes alone does not mean that these genes are the causative could cause disease onset in its host.

A pathogenic fungi that infects animal hosts, including human has higher degree of variability in its method of infection when compared to plant pathogenic fungi. This is an observation that supports the understanding that hosts determine the mode of infections, rather than the nature of the fungus itself. This may be attributed by millions of years of adaptation and evolution that had led to specialized pathogenicity among fungi and make eradicating fungal diseases extremely difficult. Many species of fungi that display pathogenicity towards plant hosts requires development of specific structure to invade the plant host. In *Ustilago maydis* for instance requires the development of dikaryotic filament to penetrate the plant cell wall and this process is often controlled by the regulation of transcription factors (Pérez-Martín & de Sena-Tomás, 2011). Similarly for fungal species that affect animal or human hosts infection it is related to the fungal life cycle through either one of the three ways: Replication of Fungus, Immune

Response caused by Fungus Infection, and Competition for Resources (FutureLearn, 2021).

The number of Common Fungal Pathogenicity-related Gene identified in this study are genes that are crucial to the life cycle of pathogenic fungi such as Ubiquitin which is a general protein that is required for breaking down of proteins to amino acids and Histones, which as described by Gargolionis et al. (2012) where the modification by either acetylation or methylation would cause the onset of pathogenicity. This again is an example where pathogenicity is related to proteins that are crucial for the maintenance of life among both plant and animal fungal pathogen. This study has found the similarities in the genes or proteins that are participating in fungal pathogenicity, despite the differences in the range of host organisms.

5.2 Common Fungal Pathogenicity-related Genes across Kingdom of Fungus.

One of the main objectives of the study is to identify Common Fungal Pathogenicity-related Genes across the kingdom of fungus and build resources that can be leveraged on in the future post-study to continue uncovering and refining the pool of common pathogenic genes which can be utilized by the research community to create molecular diagnostic method that are targeted for broad spectrum usage. This study has identified 59 high confidence common fungal pathogenicity-related genes that can serve as a foundation for further research.

5.2.1 Comparative genomics pipeline for Fungal pathogenicity study and publicly available data

In the Chapter 2 of this thesis, we discussed various comparative genomics tools and pipeline that are available currently and a knowledge gap was identified for a tool dedicated for the comparative genomics effort of fungal pathogenicity-related genes. Leveraging on publicly available data for fungal pathogenicity a Fungal Pathogenic Gene Comparative Pipeline was created using a combination of commonly used bioinformatics tools and shell scripting where all scripts developed are attached in Appendix A. Once the pipeline is established the next step for the pipeline was to be automated to build a model of fungal pathogenicity genes that can be used to identify candidate genes with higher efficiency.

Development of a Database portal for all data generated was pivotal and is one of the objectives for the study. To create a database portal that is efficient and user friendly a XAMPP architecture was deployed and using Content Management Software tool with Joomla! a database portal was created and configured. The database portal is hosted on cfgp.leapomics.com at the moment and will be able to serve as a platform for public access and collaboration for the scientific community.

5.2.2 Challenges of Identification and Development of Universal Genomic Markers for Pathogenic Fungi

Discovery of genomics markers for Pathogenic Fungi has always been the goal of many comparative genomics researchers, as the saying “Prevention is better than cure” if scientists can detect pathogenic markers in the early stage of development, then recovery plans can be put in place to prevent disease infestations across a wide area of plantations and farmlands. As an example, the Internal Transcribed Spacer (ITS) region, namely ITS1 and ITS2 of the rRNA gene are both target region for species identification of *Candida albicans, Candida glabrata, Candida parapsilosis, Candida tropicalis,* and *Aspergillus fumigatus*. This region is important to identify the differences in genome sequence and is often used as the universal DNA barcode marker for difference species of fungi (Schoch, et al. 2012) and it is what this study is trying to achieve and further differentiate between pathogenic and non-pathogenic species of fungi.

This study had identified homologous pathogenicity genes that display high degree of homology across multiple fungal species and phyla. However this is not done, as more genome sequences for other species of fungi is published it can be included to develop the model. Thi can increase the confidence to use common fungal pathogenicity-related genes as tool to develop methodology for fungal pathogenicity diagnosis where primers can be designed to target conserved pathogenicity genes, and using the methodology to fill the gap in pathogenicity identification in hard-to-detect diseases.

The challenge with developing a universal marker to detect pathogenicity lies with the identification of a unique genomic attribute among pathogenic fungi. From the polymorphic markers identified from this study we discovered multiple conserved SNPs across all the sequences of the 59 CFPG entries. Given the highly conserved nature of the genomic marker further validation work can be made against non-pathogenic isolates to check if these SNPs can indeed be used as a polymorphic marker to differentiate between pathogenic and the non-pathogenic fungi.

Additionally, although pathogenicity genes are well-conserved, but the onset of disease are triggered by highly complex biological pathways and triggers. The presence of the pathogenicity-related genes is only a single dimensional observation on pathogenicity conservation in terms of the presence and absence of genes and to add into that observation it requires genes expression profiling of the pathogenicity-related genes. Transcriptome analysis will provide more insights into the onset of pathogenicity of these genes such as those done with *Magnaporthe oryzae* (Jeon, et al. 2020) and *Sclerotinia sclerotiorum* (Chittem, et al. 2020) allowing further understanding of disease onset and by developing intervention during these pathways will create an avenue for disease prevention and treatment.

5.3 Cloud-based and Public Domain Data-Driven Research

For years genomics studies have been dominated by on-premise computing resources, which are extremely expensive and hence massive DNA sequencing studies high barrier to entry for researchers with modest resources to procure and maintain super computing resources. The landscape of genomics studies, however, has transformed and evolved with maturing cloud computing technologies offered by companies such as Amazon (Amazon Web Services), Google (Google Cloud), Alibaba (Ali Cloud) which is available as long as there is an internet connection. This has allowed lower barrier to entry for bioinformatics research and analysis and hence further pushing innovation in the space of study. This study utilizes a combination of Cloud-based and local computing resources where the initial processes that requires uninterrupted running uses the Cloud compute resource and the downstream analysis utilizing local computing resources. This strategy allows efficient usage of resources both financially and time as Cloud-based systems are not the most cost-effective to maintain.

The availability of public genomics data allows the scientific community afforded the opportunity and resources to accelerate scientific research. This study utilizes all publicly available data from various sources to perform secondary analysis of fungal pathogenicity data to advance the understanding of fungal pathogenicity, and the conservation of fungal pathogenicity and infer new methodology for identification of pathogenic fungi across different phyla.

The key in using publicly available genomic data requires data collection and proper data clean up. For sequences that were utilized in this study all fungal sequences that were downloaded were not utilized for analysis and the main reason for that was due to the varied level of completeness in data and duplication. This requires extensive data clean-up process to identify the final dataset for the study, which reduced the number of fungal sequences from 247 to 86. This is the challenges with dealing with huge dataset and it is a process that must be adopted in any data analysis.

The study had highlighted the importance of understanding the differences between nucleotide and protein sequences, and what would be the best approach to use each data types as well as discussed various challenges faced by the scientific community in the understanding of fungal pathogenicity. By leveraging on publicly available data this study adds to the understanding of fungal pathogenicity at the genomic level and contributes to the betterment in advancing knowledge of the subject. However, as fungus generally have plastic genomes and fast evolution time the understanding and knowledge will need to be developed continuously, and hence more work and effort is still required to study the subject of fungal pathogenicity.

CHAPTER 6

CONCLUSION

This study has revealed conservation of fungal pathogenicity across species of fungi from different phyla, regardless of the host is plant or human. The study also identified that most publicly available fungal pathogenicity-related databases lack representation across different phyla of fungus. The initial objective of understanding the genomics diversity and relationship among pathogenic fungi was achieved as this study unveiled clear pattern of genomics conservation with the identification of 59 Common Fungal Pathogenicity-related Genes, which was studied in detailed using phylogenetics trees that showed the relation distances between each member of study.

The Fungal Pathogenic Gene Comparative Pipeline was constructed and can be used for re-processing additional species of fungi or with newly identified pathogenic genes in public domain. Based on the identification of the Common Fungal Pathogenicity-related Genes, a Web Database Application has been developed and can be accessed here at cfpg.leapomics.com and these data are available for download. This database can serve as a foundation for further research and development to increase the level of confidence of the identified CFPG entries and validation through transcriptomics studies to confirm at the phenotypic level will further enhances our understanding of conservation of fungal pathogenicity. Hence all objectives of the study have been achieved.

Recent development in the realm of information technology towards the direction of artificial intelligence provide an avenue for automation of data discovery hence allow continuous development of tools and datasets that can benefit the community in shorter timeframe than executing end-to-end workflow manually. This study aims to develop a fungal pathogenicity comparative genomics pipeline that can be utilized as a platform for automation of discovery of pathogenicity-related genes across a diverse group of fungus phylum and species.

The automated tool has the potential to be developed further to continuous identify different signals within the datasets of genomics data, with the possibility of uncovering more pathogenicity-related information through AI. This is well demonstrated in the recent development of AlphaFold (Jumper, et al. 2021) which uses deep learning algorithm in predicting protein structures.

This study is not a conclusion of the effort in understanding fungal pathogenicity, due to the plastic nature of fungal genome. Hence the pipeline and database portal need continuous improvement to cater for needs in the future.

REFERENCES

Ahmadi, P., Muharam, F. M., Ahmad, K., Mansor, S., & Seman, I. A. (2017). Early Detection of Ganoderma Basal Stem Rot of Oil Palms using Artificial Neural Network Spectral Analysis. *Plant Disease*, *101*, 1009-1016. doi: http://dx.doi.org/10.1094/PDIS-12-16-1699-RE

Alastruey-Izquierdo, A., Alcazar-Fuoli, L., & Cuenca-Estrella, M. (2014). Antifungal susceptibility profile of cryptic species of Aspergillus. Mycopathologia, 178(5-6), 427–433. https://doi.org/10.1007/s11046-014-9775-z

Amselem, J., Cuomo, C. A., van Kan, J. A. L., Viaud, M., Benito, E. P., Couloux, A., Coutinho, P. M., de Vries, R. P., Dyer, P. S., Filinger, S., Fournier, E., Gout, L., Hahn, M., Kohn, L., Lapalu, N., Plummer, K. M., Pradier, J. M., Quevillon, E., Sharon, A., Simon, A., ten Have, A., Tudzynski, B., Beffa, R., Benoit, I., Bouzid, O., Brault, B., Chen, Z., Choquer, M., Collemare, J., Cotton, P., Danchin, E. G., Da Silva, C., Gautier, A., Giraud, C., Giraud, T., Gonzalez, C., Grossetete, S., Guldener, U., Henrissat, B., Howlett, B. J., Kodira, C., Krestchmer, M., Lappartient, A., Leroch, M., Levis, C., Mauceli, E., Neuveglise, C., Oeser, B., Pearson, M., Poulain, J., Poussereau, N., Quesneville, H., Rascle, C., Schumacher, J., Segurens, B., Sexton, A., Silva, E., Sirven, C., Soanes, D. M., Talbot, N. J., Templeton, M., Yandava, C., Yarden, O., Zeng, Q., Rollins, J. A, Lebrun, M. & Dickman, M. (2011). Genomic Analysis of the Necrotrophic 43 Fungal Pathogens Sclerotinia sclerotiorum and Botrytis cinerea. *PLoS Genetics*, *7*(8). 1-27.

Andrew, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Retrieved from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Apache Friends (2023). XAMPP. Retrieved from: https://www.apachefriends.org/index.html

Bailey, L. (1963). Infectious Diseases of the Honey-bee. Land Books.

Basenko, E. Y., Pulman, J. A., Shanmugasundram, A., Harb, O. S., Crouch, K., Starns, D., Warrenfeltz, S., Aurrecoechea, C., Stoeckert, C. J., Jr, Kissinger, J. C., Roos, D. S., & Hertz-Fowler, C. (2018). FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. Journal of fungi (Basel, Switzerland), 4(1), 39. https://doi.org/10.3390/jof4010039

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., & Kauserud, H. (2010). ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. BMC microbiology, 10, 189. https://doi.org/10.1186/1471-2180-10-189

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., & Sayers, E. W. (2018). GenBank. Nucleic acids research, 46(D1), D41–D47. https://doi.org/10.1093/nar/gkx1094

Bernardi, G., Wiley, E. O., Mansour, H., Miller, M. R., Orti, G., Haussler, D., O'Brien, S. J., Ryder, O. A., & Venkatesh, B. (2012). The fishes of Genome 10K. Marine genomics, 7, 3–6. https://doi.org/10.1016/j.margen.2012.02.002

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., & Trees, E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases, 24(4), 335–341. https://doi.org/10.1016/j.cmi.2017.10.013

BGI (2011). 10,000 Microbial Genome Project. Retrieved from: http://ldl.genomics.org.cn/page/M-research.jsp

Blackwell, M. (2011). The fungi: 1, 2, 3 … 5.1 million species? *Am J Bot*, *98*(3), 426-438. doi: 10.3732/ajb.1000298

Brauer, V. S., Rezende, C. P., Pessoni, A. M., De Paula, R. G., Rangappa, K. S., Nayaka, S. C., Gupta, V. K., & Almeida, F. (2019). Antifungal Agents in Agriculture: Friends and Foes of Public Health. Biomolecules, 9(10), 521. https://doi.org/10.3390/biom9100521

Broad Institute (2014). Fungal Genomics. Retrieved from: http://www.broadinstitute.org/scientific-community/science/projects/fungal-genomeinitiative/fungal-genomics

Brown, G. D., Denning, D. W., Gow, N. A., Levitz, S. M., Netea, M. G., & White, T. C. (2012). Hidden killers: human fungal infections. Science translational medicine, 4(165), 165rv13. https://doi.org/10.1126/scitranslmed.3004404

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. BMC bioinformatics, 10, 421. https://doi.org/10.1186/1471-2105-10-421

Carris, L. M., C. R. Little & C. M. Stiles (2012). Introduction to Fungi. The Plant Health Instructor. DOI:10. 1094/PHI-I-2012-0426-01

Carver, T., Harris, S. R., Berriman, M., Parkhill, J., & McQuillan, J. A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics (Oxford, England), 28(4), 464–469. https://doi.org/10.1093/bioinformatics/btr703

Chen, X. L., Wang, Z., & Liu, C. (2016). Roles of Peroxisomes in the Rice Blast Fungus. *BioMed research international*, 2016, 9343417. doi:10.1155/2016/9343417

Chittem, K., Yajima, W. R., Goswami, R. S., & Del Río Mendoza, L. E. (2020). Transcriptome analysis of the plant pathogen Sclerotinia sclerotiorum interaction with resistant and susceptible canola (Brassica napus) lines. PloS one, 15(3), e0229844. https://doi.org/10.1371/journal.pone.0229844

Choi, J., & Kim, S. H. (2017). A genome Tree of Life for the Fungi kingdom. Proceedings of the National Academy of Sciences of the United States of America, 114(35), 9391–9396. https://doi.org/10.1073/pnas.1711939114

Chybowska, A. D., Childers, D. S., & Farrer, R. A. (2020). Nine Things Genomics Can Tell Us About Candida auris. Frontiers in genetics, 11, 351. https://doi.org/10.3389/fgene.2020.00351

Corley, R.H.V. and Tinker, P.B. (2003) The Oil Palm. 4th Edition, Wiley, Hoboken, 562 p. https://doi.org/10.1002/9780470750971

Davey, J., Hohenlohe, P., Etter, P., Boone, P. D., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet*, *12*, 499–510. doi:10.1038/nrg3012

Dean, R. A., Talbot, N. J., Ebbole, D. J., Farman, M. L., Mitchell, T. K., Orbach, M. J., Thon, M., Kulkarni, R., Xu, J. R., Pan, H., Read, N. D., Lee, Y. H., Carbone, I., Brown, D., Oh, Y. Y., Donofrio, N., Jeong, J. S., Soanes, D. M., Djonovic, S., Kolomiets, E., Rehmeyer, C., Li, W., Harding, M., Kim, S., Lebrun, M. H., Bohnert, H., Coughian, S., Butler, J., Calvo, S., Ma, L. J., Nicol, R., Purcel, S., Nusbaum, C., Galagan, J. E., & Birren C.W. (2005). The Genome Sequence of the Rice Blast Fungus Magnaporthe grisea. *Nature, 434*, 980-986. doi: 10.1038/nature03449

Dean, R., Van Kan, J. A., Pretorius, Z. A., Hammond-Kosack, K. E., Di Pietro, A., Spanu, P. D., Rudd, J. J., Dickman, M., Kahmann, R., Ellis, J., & Foster, G. D. (2012). The Top 10 fungal pathogens in molecular plant pathology. *Molecular plant pathology*, *13*(4), 414–430. doi:10.1111/j.1364-3703.2011.00783.x

Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., Nisanci, G., Cetin-Atalay, R., & Ozturk, M. (2002). PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. Bioinformatics (Oxford, England), 18(7), 996–1003. https://doi.org/10.1093/bioinformatics/18.7.996

Dieckmann, M. A., Beyvers, S., Nkouamedjo-Fankep, R. C., Hanel, P. H. G., Jelonek, L., Blom, J., & Goesmann, A. (2021). EDGAR3.0: comparative genomics and phylogenomics on a scalable infrastructure. Nucleic acids research, 49(W1), W185–W192. https://doi.org/10.1093/nar/gkab341

Donovan, P. D., Gonzalez, G., Higgins, D. G., Butler, G., & Ito, K. (2018). Identification of fungi in shotgun metagenomics datasets. PloS one, 13(2), e0192898. https://doi.org/10.1371/journal.pone.0192898

Dos Santos, R. A. C., Steenwyk, J. L., Rivero-Menendez, O., Mead, M. E., Silva, L. P., Bastos, R. W., Alastruey-Izquierdo, A., Goldman, G. H., & Rokas, A. (2020). Genomic and Phenotypic Heterogeneity of Clinical Isolates of the Human Pathogens Aspergillus fumigatus, Aspergillus lentulus, and Aspergillus fumigatiaffinis. Frontiers in genetics, 11, 459. https://doi.org/10.3389/fgene.2020.00459

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., Joly, D. L., Hacquard, S., Amselem, J., Cantarel, B. L., Chiu, R., Coutinho, P. M., Feau, N., Field, M., Frey, P., Gelhaye, E., Goldberg, J., Grabherr, M. G., Kodira, C. D., Kohler, A., Kües, U., Lindquist, E. A., Lucas, S. M., Mago, R., Mauceli, E., Morin, E., Murat, C., Pangilinan, J. L., Park, R., Pearson, M., Quesneville, H., Rouhier, N., Sakthikumar, S., Salamov, A. A., Schmutz, J., Selles, B., Shapiro, H., Tanguay, P., Tuskan, G. A., Henrissat, B., Van de Peer, Y., Rouzé, P., Ellis, J. G., Dodds, P. N., Schein, J. E., Zhong, S., Hamelin, R. C., Grigoriev, I. V., Szabo, L. J., & Martin, F. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(22), 9166–9171. doi:10.1073/pnas.1019315108

Du, Y., Shi, Y., Yang, J., Chen, X., Xue, M., Zhou, W., & Peng, Y. L. (2013). A serine/threonine-protein phosphatase PP2A catalytic subunit is essential for asexual development and plant infection in Magnaporthe oryzae. Current genetics, 59(1-2), 33–41. https://doi.org/10.1007/s00294-012-0385-3

Edgar R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Felsenstein, J. (1989). PHYLIP – Phylogeny Inference Package. *Cladistics*, *5*, 164-166.

Ferrer, C., Colom, F., Frasés, S., Mulet, E., Abad, J. L., & Alió, J. L. (2001). Detection and identification of fungal pathogens by PCR and by ITS2 and 5.8S ribosomal DNA typing in ocular infections. *Journal of clinical microbiology*, *39*(8), 2873–2879. doi:10.1128/JCM.39.8.2873-2879.2001

Finley, D., Ulrich, H. D., Sommer, T., Kaiser, P. (2012). The Ubiquitin-Proteasome System of Saccharomyces Cerevisiae. Genetics 192, 319–360. doi: 10.1534/genetics.112.140467

Fisher, M. C., Hawkins, N. J., Sanglard, D., & Gurr, S. J. (2018). Worldwide emergence of resistance to antifungal drugs challenges human health and food security. Science (New York, N.Y.), 360(6390), 739–742. https://doi.org/10.1126/science.aap7999

Fisher, M. C., Henk, D. A., Briggs, C. J., Brownstein, J. S., Madoff, L. C., McCraw, S. L., & Gurr, S. J. (2012). Emerging fungal threats to animal, plant and ecosystem health. Nature, 484(7393), 186–194. doi:10.1038/nature10947

Fukuda, A., Kodama, Y., Mashima, J., Fujisawa, T., & Ogasawara, O. (2021). DDBJ update: streamlining submission and access of human data. Nucleic acids research, 49(D1), D71–D75. https://doi.org/10.1093/nar/gkaa982

FutureLearn (2021). How do Fungi cause disease? – Part 1. Retrieved from: https://www.futurelearn.com/info/courses/antifungal-stewardship/0/steps/55623

Garcia-Solache, M. A., & Casadevall, A. (2010). Global warming will bring new fungal diseases for mammals. *mBio*, *1*(1), e00061-10. doi:10.1128/mBio.00061-10

Gargalionis, A. N., Piperi, C., Adamopoulos, C., & Papavassiliou, A. G. (2012). Histone modifications as a pathogenic mechanism of colorectal tumorigenesis. The international journal of biochemistry & cell biology, 44(8), 1276–1289. https://doi.org/10.1016/j.biocel.2012.05.002

GenomeWeb (2021). Retrieved from: https://www.genomeweb.com/

Guirao-Abad, J. P., Weichert, M., Luengo-Gil, G., Sze Wah Wong, S., Aimanianda, V., Grisham, C., Malev, N., Reddy, S., Woollett, L., & Askew, D. S. (2021). Pleiotropic Effects of the P5-Type ATPase SpfA on Stress Response Networks Contribute to Virulence in the Pathogenic Mold Aspergillus fumigatus. mBio, 12(5), e0273521. https://doi.org/10.1128/mBio.02735-21

Gupta, A. K., Jain, H. C., Lynde, C. W., Macdonald, P., Cooper, E. A., & Summerbell, R. C. (2000). Prevalence and epidemiology of onychomycosis in patients visiting physicians' offices: a multicenter canadian survey of 15,000 patients. Journal of the American Academy of Dermatology, 43(2 Pt 1), 244–248. https://doi.org/10.1067/mjd.2000.104794

Han, B., Takvorian, P. M., & Weiss, L. M. (2020). Invasion of Host Cells by Microsporidia. Frontiers in microbiology, 11, 172. https://doi.org/10.3389/fmicb.2020.00172

Hariharan, G., & Prasannath, K. (2021). Recent Advances in Molecular Diagnostics of Fungal Plant Pathogens: A Mini Review. Frontiers in cellular and infection microbiology, 10, 600234. https://doi.org/10.3389/fcimb.2020.600234

Hawksworth, D. L. (2001). The magnitude of fungal diversity: the 1.5 million spcies estimate revisited. *Mycological Research, 105*(12), 1422-1432. doi: https://doi.org/10.1017/S0953756201004725

Hingamp, P., van den Broek, A. E., Stoesser, G., & Baker, W. (1999). The EMBL Nucleotide Sequence Database. Contributing and accessing data. Molecular biotechnology, 12(3), 255–267. https://doi.org/10.1385/MB:12:3:255

Hogan, D. A., & Sundstrom, P. (2009). The Ras/cAMP/PKA signaling pathway and virulence in Candida albicans. Future microbiology, 4(10), 1263–1270. https://doi.org/10.2217/fmb.09.106

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., Gall, A., … Flicek, P. (2021). Ensembl 2021. Nucleic acids research, 49(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Hu Z. (2014). Using VisANT to Analyze Networks. Current protocols in bioinformatics, 45(88), 8.8.1–8.8.39. https://doi.org/10.1002/0471250953.bi0808s45

Hushiarian, R., Yusof, N. A., & Dutse, S. W. (2013). Detection and control of Ganoderma boninense: strategies and perspectives. SpringerPlus, 2, 555. https://doi.org/10.1186/2193-1801-2-555

JaeJin, C., & Sung-Hou, K. (2017). Fungal Tree of Life: A "genome tree". *Proceedings of the National Academy of Sciences*, *114*(35), 9391-9396. doi: 10.1073/pnas.1711939114

Jeon, J., Lee, G. W., Kim, K. T., Park, S. Y., Kim, S., Kwon, S., Huh, A., Chung, H., Lee, D. Y., Kim, C. Y., & Lee, Y. H. (2020). Transcriptome Profiling of the Rice Blast Fungus Magnaporthe oryzae and Its Host Oryza sativa During Infection. Molecular plant-microbe interactions : MPMI, 33(2), 141–144. https://doi.org/10.1094/MPMI-07-19-0207-A

Rochen (2017). Joomla!. Retrieved from: https://www.joomla.org

Joseleau, J.P. & Pérez, S. (2016). The Plant Cell Walls. Accessed: www.glycopedia.eu

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaczmarek, A. M., King, K. M., West, J. S., Stevens, M., Sparkes, D., & Dickinson, M. J. (2019). A Loop-Mediated Isothermal Amplification (LAMP) Assay for Rapid and Specific Detection of Airborne Inoculum of Uromyces betae (Sugar Beet Rust). Plant disease, 103(3), 417–421. https://doi.org/10.1094/PDIS-02-18-0337-RE

Kamper, J., Kahmann, R., Bolker, M., Ma, L., Brefort, T., Saville, B. J., Banuett, F., Kronstad, J. W., Gold, S. E., Muller, O., Perlin, M. H., Wosten, H. A. B., de Vries, R., Ruiz-Herrera, J., Reynaga-Pena, C. G., Snetselaar, K., McCann, M., Perez-Martin, J., Feldbrugge, M., Basse, C. W., Steinberg, G., Ibeas, J. I., Holloman, W., Guzman, P., Farman, M., Stajich, J. E., Sentandreu, R., Gonzalez-Preito, J. M., Kennell, J. C., Molina, L., Schirawski, J., Mendoza-Mendoza, A., Greilinger, D., Munch, K., Rossel, N., Scherer, M., Vranes, M., Ladendorf, O., Vincon, V., Fuchs, U., Sandrock, B., Meng S., Ho, E. C. H., Cahill, M. J., Boyce, K. J., Klose, J., Klosterman, S. J., Deelstra, H. J., OrtizCastellanos, L., Li, W., Sanchez-Alonso, P., Schreier, P. H., Hauser-Hahn, I., Vaupel, M., Koopmann, E., Friedrich, G., Voss, H., Schluter, T., Margolis, J., Platt, D., Swimmer, C., Gnirke, A., Chen, F., Vysotskaia, V., Mewes, H., Mauceli, E. W., DeCaprio, D., Wade, C. M., Butler, J., Young, S., Jaffe, D. B., Calvo, S., Nusbaum, C., Galagan, J. & Birren, B. W. (2006). Insights from the Genome of the Biotrophic Fungal Plant Pathogen *Ustilago maydis*. *Nature*, *444*, 97-101. doi:10.1038/nature05248

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772–780. doi:10.1093/molbev/mst010

Kidd, S. E., Chen, S. C., Meyer, W., & Halliday, C. L. (2020). A New Age in Molecular Diagnostics for Invasive Fungal Disease: Are We Ready?. Frontiers in microbiology, 10, 2903. https://doi.org/10.3389/fmicb.2019.02903

King, R., Brown, N. A., Urban, M., & Hammond-Kosack, K. E. (2018). Inter-genome comparison of the Quorn fungus *Fusarium venenatum* and the closely related plant infecting pathogen *Fusarium graminearum*. *BMC Genomics*, *19*, 269. doi: 10.1186/s12864-018-4612-2

Knapp, D. G., Németh, J. B., Barry, K, Hainaut, M., Henrissat, B., Johnson, J., Kuo, A., Lim, J. H. P, Lipzen, A., Nolan, M., Ohm, R. A., Tamás, L., Grigoriev, I. V., Spatafora, J. W., Nagy, L. G. & Kovács, G. M. (2018). Comparative genomics provides insights into the lifestyle and reveals functional heterogeneity of dark septate endophytic fungi. *Scientific Reports*, *8*, 6321. doi:10.1038/s41598-018-24686-4

Knapp, D. G., Németh, J. B., Barry, K., Hainaut, M., Henrissat, B., Johnson, J., Kuo, A., Lim, J., Lipzen, A., Nolan, M., Ohm, R. A., Tamás, L., Grigoriev, I. V., Spatafora, J. W., Nagy, L. G., & Kovács, G. M. (2018). Comparative genomics provides insights into the lifestyle and reveals functional heterogeneity of dark septate endophytic fungi. Scientific reports, 8(1), 6321. https://doi.org/10.1038/s41598-018-24686-4

Koepfli, K. P., Paten, B., Genome 10K Community of Scientists, & O'Brien, S. J. (2015). The Genome 10K Project: a way forward. *Annual review of animal biosciences*, *3*, 57–111. doi:10.1146/annurev-animal-090414-014900

Kozel, T. R., & Wickes, B. (2014). Fungal diagnostics. Cold Spring Harbor perspectives in medicine, 4(4), a019299. https://doi.org/10.1101/cshperspect.a019299

Kronstadt, J. W. (1997). Virulence and cAMP in Smuts, Blasts and Blights. *Trends Plant Sci*, *2*, 193-199.

Lee, J., Lee, S., & Young, J. P. (2008). Improved PCR primers for the detection and identification of arbuscular mycorrhizal fungi. FEMS microbiology ecology, 65(2), 339–349. https://doi.org/10.1111/j.1574-6941.2008.00531.x

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup

(2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (Oxford, England), *25*(16), 2078–2079. doi:10.1093/bioinformatics/btp352

Lin, C. J., & Chen, Y. L. (2018). Conserved and Divergent Functions of the cAMP/PKA Signaling Pathway in Candida albicans and Candida tropicalis. Journal of fungi (Basel, Switzerland), 4(2), 68. https://doi.org/10.3390/jof4020068

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic acids research, 42(Database issue), D490–D495. https://doi.org/10.1093/nar/gkt1178

Löytynoja A. (2014). Phylogeny-aware alignment with PRANK. Methods in molecular biology (Clifton, N.J.), 1079, 155–170. https://doi.org/10.1007/978-1-62703-646-7_10

Lu, T., Yao, B., & Zhang, C. (2012). DFVF: database of fungal virulence factors. Database : the journal of biological databases and curation, 2012, bas032. https://doi.org/10.1093/database/bas032

Martin, K. J., & Rygiewicz, P. T. (2005). Fungal-specific PCR primers developed for analysis of the ITS region of environmental DNA extracts. *BMC microbiology*, *5*, 28. doi:10.1186/1471-2180-5-28

Meng, S., Brown, D. E., Ebbole, D. J., Torto-Alalibo, T., Oh, Y. Y., Deng, J., Mitchell, T. K., & Dean, R. A. (2009). Gene Ontology annotation of the rice blast fungus, Magnaporthe oryzae. BMC microbiology, 9 Suppl 1(Suppl 1), S8. https://doi.org/10.1186/1471-2180-9-S1-S8

Moran, G. P., Coleman, D. C., & Sullivan, D. J. (2011). Comparative genomics and the evolution of pathogenicity in human pathogenic fungi. Eukaryotic cell, 10(1), 34–42. https://doi.org/10.1128/EC.00242-10

National Human Genome Research Institute (2019). DNA Sequencing Costs. Retrieved from: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Nature Education (2014). Retrieved from: https://www.nature.com/scitable/definition/silent-mutation-10/

Nikitin, A., Egorov, S., Daraselia, N., & Mazo, I. (2003). Pathway studio--the analysis and navigation of molecular networks. Bioinformatics (Oxford, England), 19(16), 2155–2157. https://doi.org/10.1093/bioinformatics/btg290

Okonechikov, K., Golosova, O., Fursov, M. & the UGENE team (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, *28*, 1166-1167. doi: 10.1093/bioinformatics/bts091

Ou, S. H. (1980). Pathogen Variability and host Resistance in Rice Blast Disease. *Ann. Rev. Phytopathol*, *18*, 167-187. doi: https://doi.org/10.1146/annurev.py.18.090180.001123

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., & Harris, S. R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial genomics*, *2*(4), e000056. doi:10.1099/mgen.0.000056

Pearson WR. An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinformatics. (2013); Chapter 3:Unit3.1. doi:10.1002/0471250953.bi0301s42

Pérez-Martín, J., & de Sena-Tomás, C. (2011). Dikaryotic cell cycle in the phytopathogenic fungi Ustilago maydis is controlled by the DNA damage response cascade. Plant signaling & behavior, 6(10), 1574–1577. https://doi.org/10.4161/psb.6.10.17055

Petersen, J.H. (2013). The Kingdom of Fungi. (Princeton Univ Press, Princeton, NJ).

Plesken, C., Weber, R. W. S., Rupp, S., Leroch, M., & Hahn, M. (2015). *Applied and Environmental Microbiology*. *81*(20), 7048-7056. doi: 10.1128/AEM.01719-15

Prakash, H., & Chakrabarti, A. (2021). Epidemiology of Mucormycosis in India. Microorganisms, 9(3), 523. https://doi.org/10.3390/microorganisms9030523

Ramzi, A. B., Che Me, M. L., Ruslan, U. S., Baharum, S. N., & Nor Muhammad, N. A. (2019). Insight into plant cell wall degradation and pathogenesis of Ganoderma boninense via comparative genome analysis. PeerJ, 7, e8065. https://doi.org/10.7717/peerj.8065

Reddy, M. K., & Ananthanarayanan, T. V. (1984). Detection of *Ganoderma lucidum* in Betelnut by the Fluorescent Antibody Technique. *Transactions of the British Mycological Society*, *82*(3), 559-561. doi: https://doi.org/10.1016/S0007-1536(84)80026-1

Rodriguez-Moreno, L., Ebert, M. K., Bolton, M. D., & Thomma, B. P. H. J. (2018). Tools of the crook- infection strategies of fungal plant pathogens. The Plant journal : for cell and molecular biology, 93(4), 664–674. https://doi.org/10.1111/tpj.13810

Sanusi, M. S., Mustafa, N., Mad Zin, A., & Mansor, P. (2016). The Death Grip of Ants Infected by Brain-manipulating Fungus Ophiocordyceps unilateralis. *Conservation Malaysia*, *24*, 3-4.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database : the journal of biological databases and curation, 2020, baaa062. https://doi.org/10.1093/database/baaa062

Schoch, C.L., Seifert, K.A., Huhndork, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W. and Fungal Barcoding Consortium (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences Apr 2012*, 109 (16) 6241-6246; DOI: 10.1073/pnas.1117018109

Schuelke TA, Wu G, Westbrook A, Woeste K, Plachetzki DC, Broders K, MacManes MD. Comparative Genomics of Pathogenic and Nonpathogenic Beetle-Vectored Fungi in the Genus Geosmithia. Genome Biol Evol. 2017 Dec 1;9(12):3312-3327. doi: 10.1093/gbe/evx242. PMID: 29186370; PMCID: PMC5737690.

Scientific Beekeeping (2023). Available at: https://www.scientificbeekeeping.co.uk/stonebrood.html. Accessed 12 April 2023.

Sexton, A. C., & Howlett, B. J. (2006). Parallels in fungal pathogenesis on plant and animal hosts. Eukaryotic cell, 5(12), 1941–1949. https://doi.org/10.1128/EC.00277-06

Seyedmousavi, S., Guillot, J., Arné, P., de Hoog, G. S., Mouton, J. W., Melchers, W. J., & Verweij, P. E. (2015). Aspergillus and aspergilloses in wild and domestic animals: a global health concern with parallels to human disease. Medical mycology, 53(8), 765–797. https://doi.org/10.1093/mmy/myv067

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Sharon, A., & Shlezinger, N. (2013). Fungi infecting plants and animals: killers, non-killers, and cell death. PLoS pathogens, 9(8), e1003517. https://doi.org/10.1371/journal.ppat.1003517

Shi-Kunne, X., van Kooten, M., Depotter, J. R. L., Thomma, B. P. H. J., & Seidl, M. F. (2019) The Genome of the Fungal Pathogen Verticillium dahliae Reveals Extensive Bacterial to Fungal Gene Transfer. *Genome Biology and Evolution*, *11*(3), 855–868. doi: https://doi.org/10.1093/gbe/evz040

Lu, T., Yao, B., & Zhang, C. (2012). DFVF: database of fungal virulence factors. Database : the journal of biological databases and curation, 2012, bas032. https://doi.org/10.1093/database/bas032

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research, 22(22), 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Tudzynski, B., Hedden, P., Carrera, E., & Gaskin, P. (2001). The P450-4 gene of Gibberella fujikuroi encodes ent-kaurene oxidase in the gibberellin biosynthesis pathway. *Applied and environmental microbiology*, *67*(8), 3514–3522. doi:10.1128/AEM.67.8.3514-3522.2001

Tudzynski, B., Mihlan, M., Rojas, M. C., Linnemannstöns, P., Gaskin, P., & Hedden, P. (2003). Characterization of the final two genes of the gibberellin biosynthesis gene cluster of Gibberella fujikuroi: des and P450-3 encode GA4 desaturase and the 13-hydroxylase, respectively. *J Biol Chem*, *278*, 28635–28643.

Turrà, D., Segorbe, D., & Di Pietro, A. (2014). Protein kinases in plant-pathogenic fungi: conserved regulators of infection. Annual review of phytopathology, 52, 267–288. https://doi.org/10.1146/annurev-phyto-102313-050143

UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic acids research, 49(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Urban, M., Cuzick, A., Rutherford, K., Irvine, A., Pedro, H., Pant, R., Sadanadan, V., Khamari, L., Billal, S., Mohanty, S., & Hammond-Kosack, K. E. (2017). PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database. Nucleic acids research, 45(D1), D604–D610. https://doi.org/10.1093/nar/gkw1089

Utomo, C., & Niepold, F. (2001). Development of Diagnostic Methods for Detecting *Ganoderma*-infected Oil Palms. *Journal of Phytopathology*, *148*(9-10). doi: https://doi.org/10.1046/j.1439-0434.2000.00478.x

Utomo, C., Werner, S., Niepold, F., & Deising, H. B. (2005). Identification of *Ganoderma*, the Causal Agent of Basal Stem Rot Disease in Oil Palm using a Molecular Method. *Mycopathologia*, *159*(1), 159-170.

van Dam, P., de Sain, M., Ter Horst, A., van der Gragt, M., & Rep, M. (2017). Use of Comparative Genomics-Based Markers for Discrimination of Host Specificity in Fusarium oxysporum. Applied and environmental microbiology, 84(1), e01868-17. doi:10.1128/AEM.01868-17

van der Does, H. C., & Rep, M. (2007). Virulence genes and the evolution of host specificity in plant-pathogenic fungi. Molecular plant-microbe interactions : MPMI, 20(10), 1175–1182. https://doi.org/10.1094/MPMI-20-10-1175

Walcott, R. R. (2003). Detection of Seadborne Pathogens. *HortTechnology*, *13*(1), 40-47. doi: 10.21273/HORTTECH.13.1.0040

Walker, T. S., Bais, H. P., Déziel, E., Schweizer, H. P., Rahme, L. G., Fall, R., & Vivanco, J. M. (2004). Pseudomonas aeruginosa-plant root interactions. Pathogenicity, biofilm formation, and root exudation. Plant physiology, 134(1), 320–331. https://doi.org/10.1104/pp.103.027888

Watkinson, S.C., Boddy, L. and Money, N.P. (2015). The Fungi (3rd Edition). Elsevier Ltd.

Wei, L., Liu, Y., Dubchak, I., Shon, J., & Park, J. (2002). Comparative genomics approaches to study organism similarities and differences. *Journal of Biomedical Informatics*, *35*(2), 142-150. doi: https://doi.org/10.1016/S1532-0464(02)00506-3

Wetterstrand, K.A. (2020). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 19 October 2021.

Winnenburg, R., Baldwin, T. K., Urban, M., Rawlings, C., Köhler, J., & Hammond-Kosack, K. E. (2006). PHI-base: a new database for pathogen host interactions. Nucleic acids research, 34(Database issue), D459–D464. https://doi.org/10.1093/nar/gkj047

Xu, J. R. (2002). MAP Kinases in Fungal Pathogens. *Fungal Genetics and Biology*, *31*(3), 137-152. doi: https://doi.org/10.1006/fgbi.2000.1237

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet*, *13*, 329–342. doi:10.1038/nrg3174

Ye, J., McGinnis, S., & Madden, T. L. (2006). BLAST: improvements for better sequence analysis. Nucleic acids research, 34(Web Server issue), W6–W9. https://doi.org/10.1093/nar/gkl164

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., & Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic acids research, 40(Web Server issue), W445–W451. https://doi.org/10.1093/nar/gks479

Yu, C. H., Sephton-Clark, P., Tenor, J. L., Toffaletti, D. L., Giamberardino, C., Haverkamp, M., Cuomo, C. A., & Perfect, J. R. (2021). Gene Expression of Diverse Cryptococcus Isolates during Infection of the Human Central Nervous System. mBio, 12(6), e0231321. https://doi.org/10.1128/mBio.02313-21

Zerillo, M. M., Adhikari, B. N., Hamilton, J. P., Buell, C. R., Lévesque, C. A., & Tisserat, N. (2013). Carbohydrate-Active Enzymes in *Pythium* and Their Role in

Plant Cell Wall and Storage Polysaccharide Degradation. *PLOS ONE*, *8*(9). doi: https://doi.org/10.1371/journal.pone.0072572