An Accurate Fungal Classification Tool and Evolutionary Dynamics of Aflatoxin Genes

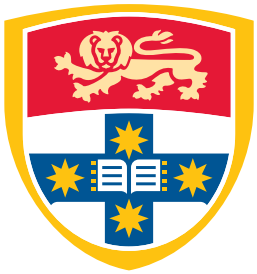
VINITA DESHPANDE
SID: 309249104

Supervisor: A/Prof Michael Charleston
Associate Supervisor: Paul Greenfield

This thesis is submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Information Technology (Honours)

School of Information Technologies
The University of Sydney
Australia

5 November 2013



THE UNIVERSITY OF
SYDNEY

Student Plagiarism: Compliance Statement

I certify that:

I have read and understood the University of Sydney Student Plagiarism: Coursework Policy and Procedure;

I understand that failure to comply with the Student Plagiarism: Academic Board Policy: Academic Dishonesty and Plagiarism can lead to the University commencing proceedings against me for potential student misconduct under the 2012 Academic Dishonesty and Plagiarism in Coursework Policy.

This Work is substantially my own, and to the extent that any part of this Work is not my own I have indicated that it is not my own by Acknowledging the Source of that part or those parts of the Work.

Name: Vinita Deshpande

Signature: **Date**:

Abstract

Fungal organisms play a critical role in agriculture, medicine and the food industry. Their impacts, however, are both beneficial and harmful to humans and the environment. For example, species such as *Aspergillus oryzae* and *Aspergillus sojae* are used in food fermentation processes, while their close relatives *Aspergillus flavus* and *Aspergillus parasiticus* naturally produce carcinogenic 'aflatoxins' that contaminate crops and foods consumed by humans. Thus the ability to accurately differentiate between fungal organisms at the species level is paramount.

Existing genome sequence-based tools for fungal classification suffer from computational inefficiency or inaccurate or unreliable classifications. Although the recently developed LSU classifier overcomes these drawbacks, the use of '28S LSU' gene sequences in the training set of this classifier restricts its classifications to genus rank only. The first part of this thesis describes the approach taken to adapt the LSU classifier to use the highly variable 'ITS' sequences instead for making classifications. The new ITS classifier developed is not only more accurate, but classifies below the genus rank to species rank.

A highly debated topic is the relationship between *A oryzae* (non-toxigenic) and *A flavus* (toxigenic). Part 2 focuses on their evolutionary relationships as a means to resolve this issue. A comparison of the evolutionary dynamics of seven aflatoxin genes from novel fungal genomes was performed using phylogenetic analysis. The findings suggest that the aflatoxin genes are experiencing similar evolutionary forces, and that *A oryzae* and *A flavus* are members of the same species, rather than separate species.

It is envisaged that the ITS classifier developed in Part 1 will prove valuable and be adopted in fungal genome studies as an efficient classification step. The findings obtained from the evolutionary dynamics of the aflatoxin genes in Part 2 has revealed insights into the underlying evolutionary mechanisms acting on the aflatoxin genes, which can help to better understand the relationships between toxigenic and non-toxigenic species.

Acknowledgements

I would firstly like to express my most sincere thanks and gratitude to my supervisor A/Prof Michael Charleston, for his invaluable guidance, support and unwavering enthusiasm throughout this project and what has become one of the most rewarding experiences of my undergraduate degree.

I would also like to express my deepest gratitude to my co-supervisor Paul Greenfield from CSIRO Computational Informatics, for once again allowing me to work under his guidance and supervision. Thank you for your patience and for sharing practical knowledge and advice, both technical and non-technical.

I am deeply thankful to Dr Nai Tran-Dinh and Dr David Midgley from CSIRO Animal, Food and Health Sciences for sharing their mycological insights and expertise, and providing continual feedback and suggestions for improving the training set and the classifier, and for devoting the time to construct the validation set used to evaluate the classifier. Thank you also for organising lab experiments for me to conduct, without which I would not have gained a better understanding and appreciation of the difficulties of generating good quality data.

Special thanks to the School of Information Technologies and CSIRO Computational Informatics for providing the resources required to complete this project. I would also like to thank the Office of the Chief Executive, CSIRO Transformational Biology and CSIRO Computational Informatics for supporting this research project.

And finally, I am deeply grateful to my parents, friends and family, for their constant support and encouragement throughout this journey.

CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background and Motivation

All living organisms are identified and classified using a biological taxonomy that assigns a name, or 'taxon', at each level in the hierarchy from 'domain' and down to 'species' (Figure 1.1). At the domain level, living organisms are divided into Bacteria, Archaea or Eukaryote, which form the three branches of life under the three-domain system proposed by Woese et al. in 1990 (Figure 1.2). Fungi are one of the oldest and largest group of organisms belonging to the Eukaryote domain, and share a common ancestor with animals and plants that is believed to have existed around one billion years ago (Moore, 2013).

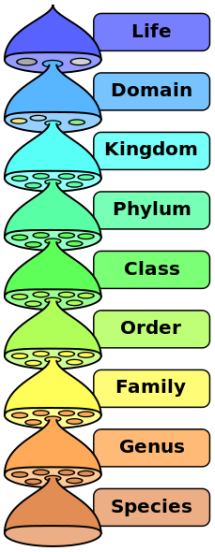


FIGURE 1.1. Biological Classification. Image obtained from `http://en.wikipedia.org/wiki/Biological_classification`.

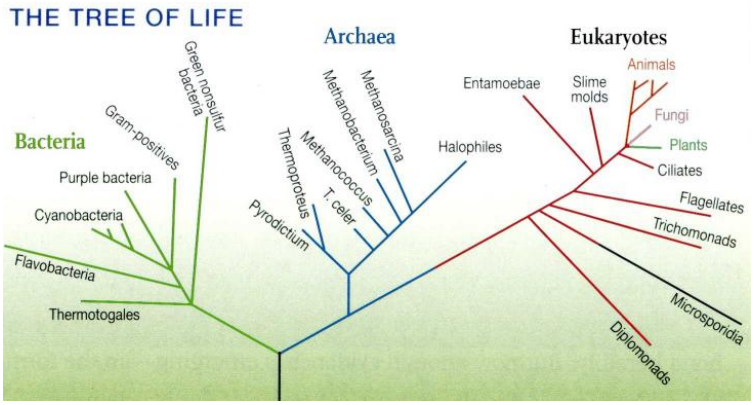


FIGURE 1.2. Phylogenetic Tree of Life showing the three domains: Bacteria, Archaea and Eukaryotes. Image obtained from `http://njsas.org/life/images/tree_of_life_sm.jpg`.

The fungal kingdom is of critical biological importance to humans and the environment as these organisms play a vital role in functions and processes required to sustain life on earth. An estimated 90% of all land plant species are in symbiosis ('living together') with mycorrhizal fungi, a mutually beneficial relationship in which fungi reside on or inside the plant's roots (Behie et al., 2013). This interaction mediates the transfer of nutrients between the plant and fungi, where the fungus receives sugars in exchange for enhancing the plant's ability to absorb water and nutrients, and possibly provide protection from pesticides and pathogens. Thus each is dependent on the other for growth and survival. Fungi are primary biodegraders, together with bacteria, that facilitate the decomposition of organic materials in the environment, utilising these as sources of nutrients and recycling essential elements such as carbon, oxygen, nitrogen and phosphorous by releasing them into the soil and atmosphere (Ahmadjian, 2013; Tariq, 2013). Industrial bioremediation processes have also benefitted from fungi for decomposing toxic compounds from industrial waste products before they are released into the environment.

One of the most important roles of fungi is as a direct source of food for humans, for example mushrooms, and *Saccharomyces cerevisiae* (yeast) in food fermentation processes for the production of bread, cheese, beer and wine, and some soya bean products (Ahmadjian, 2013). The application of fungi in medicine has revolutionised health care through the exploitation of fungal metabolic compounds to develop anti-fungal drugs (e.g. griseofulvin), cholesterol-lowering drugs (e.g. lovastatin), and anti-tumour drugs (e.g. terrequinone A) (Yu, 2012), with the most noteworthy being the discovery of the antibiotic penicillin derived from the species *Penicillium notatum* (Moore, 2013). Species such as *Saccharomyces*

cerevisiae and *Neurospora* have served as model organisms for studying cellular metabolism and biochemistry to shed light on how these processes work in more complex eukaryotes such as humans, thus greatly contributing to biological knowledge. Fungi and fungal products have also recently been identified as the most efficient and safe way of improving utilisation of natural resources, for example as renewable replacements for fossil fuels, for a more sustainable future (Lange et al., 2012).

Not all fungi, however, have beneficial applications. Species of *Aspergillus*, including *Aspergillus flavus* and its close relative *Aspergillus minisclerotigenes*, as well as *Aspergillus parasiticus* and *Aspergillus nomius*, produce toxic compounds called aflatoxin as a secondary metabolite. These toxins are of the most potent, naturally occurring carcinogens known that contaminate important food and feed crops consumed by humans and animals, such as maize and peanuts (Georgianna et al., 2010). The economic costs incurred from yield losses caused by aflatoxin contamination is in the order of several millions of dollars in the United States alone (Amaike and Keller, 2011). Aflatoxin contamination is a much greater problem in developing countries due to inadequate regulations and measures for testing aflatoxin concentrations. As a consequence, approximately 5 billion people worldwide are exposed to uncontrolled levels of aflatoxin in their daily diet (D Midgley, personal communication). Aflatoxins are identified as a potential immunosuppressive agent (Yu, 2012) and consumption of aflatoxin-contaminated food has been implicated in diseases such as aspergillosis, aflatoxicosis and heptacellular carcinoma (HCC), a type of liver cancer (Amaike and Keller, 2011). This problem is worsened with the ability of aflatoxins to act synergistically with hepatitis B and C viruses, to strengthen and promote HCC development. It is reported that aflatoxin exposure was responsible for approximately 25,000 to 155,000 cases (4.6% – 28.2%) of all HCC cases worldwide in 2010 (Amaike and Keller, 2011). Therefore, the detrimental effects of aflatoxins on both agricultural crops and human and animal health necessitates further research to enhance understanding of aflatoxin production, in order to minimise aflatoxin contamination and thereby improve food safety and sustainability.

It was originally estimated in 1991 that the fungal kingdom comprises of 1.5 million species, however more recent estimates have increased this number to 5.1 million species (Blackwell, 2011). Traditional methods to identify and classify fungi are based on morphological and cultural characteristics (Chang and Ehrlich, 2010) of fungal specimens in collections or cultures. As these are techniques are prone to misclassification, several molecular methods were developed, including DNA restriction fragment

length polymorphism (RFLP), amplified fragment length polymorphism (AFLP) and amplification of evolutionary markers such as the ribosomal RNA (rRNA) internal transcribed spacer (ITS) (Chang et al., 2007). Classifications based solely on such molecular methods, although relatively more accurate, are time consuming.

Of the total number of fungal species, only approximately 100,000 species are thought to have been characterised thus far (Hibbett and Taylor, 2013). It is believed that the vast majority, about 80 – 90%, of fungal diversity is harboured in environmental habitats in which the constituent fungal species cannot be cultured (Magnuson and Lasure, 2002). The advent of DNA sequencing technologies has revolutionised the way in which characterisation of fungal communities, analysis of fungal diversity, and classification of species is performed, by eliminating the need for culturing and cultivation (Bates et al., 2013) and instead using sequence-based classification. This has been facilitated by the new research field of 'metagenomics', in which an environmental sample is characterised to identify all the micro-organisms present as a whole, rather than identifying each species in isolation. In more recent years, the rapid uptake of high-throughput next generation DNA sequencing technologies has resulted in unprecedented volumes of fungal genomic data; the data analysis of which has been unable to keep up. Current bioinformatics classification tools are either computationally inefficient or lack discriminatory power to accurately classify fungal organisms down to the species rank in the biological taxonomy (Figure 1.1).

The beneficial and detrimental impact of fungi on humans and environment demands accurate and reliable ways of classifying novel and unknown fungal organisms. The need for a tool that achieves both rapid and accurate classification down to the species rank, together with a greater understanding of the genes responsible for aflatoxin production, form the motivations and objectives of the current study. This study forms part of a larger research project being conducted at CSIRO Animal, Food and Health Sciences (North Ryde, NSW), that involved the sequencing of 39 medically and economically relevant, novel fungal genomes which are summarised in Table 1.1.

TABLE 1.1. Summary of the 39 novel fungal genomes used in this study. Data courtesy of Dr Nai Tran-Dinh and Dr David Midgley.

Sample Number	Strain ID	Genus	Species	Source	Country	Year
FGS1	342	*Aspergillus*	*flavus*	Rice	Australia	1970
FGS2	2757	*Aspergillus*	*flavus*	Peanuts	Australia	1984
FGS3	2807	*Aspergillus*	*flavus*	Cellophane	S. Pacific	1944
FGS4	503	*Aspergillus*	*parasiticus*	Mealy bug	USA	1912
FGS5	2744	*Aspergillus*	*parasiticus*	Peanuts	Australia	1984
FGS6	3282	*Aspergillus*	*parasiticus*	Cottonseed	USA	1980
FGS7	369	*Aspergillus*	*carbonarius*	Paper	USA	1916
FGS8	5171	*Aspergillus*	*carbonarius*	Wash water	Australia	1998
FGS9	2940	*Penicillium*	*verrucosum*	Barley	Denmark	1982
FGS10	1690	*Penicillium*	*verrucosum*	Dried sausage	Germany	1973
FGS11	3543	*Aspergillus*	*nomius*	Bee	USA	1986
FGS12	3546	*Aspergillus*	*nomius*	Wheat	USA	1986
FGS13	3547	*Aspergillus*	*nomius*	Peanuts	USA	1986
FGS14	4086	*Aspergillus*	*minisclerotigenes*	Peanuts	Australia	1990
FGS15	4937	*Aspergillus*	*minisclerotigenes*	Peanuts	Argentina	1993
FGS16	543	*Aspergillus*	*ochraceus*	Hay fodder	Australia	1970
FGS17	2249	*Aspergillus*	*malvicolor*	Soil	Australia	1979
FGS18	3382	*Aspergillus*	*minisclerotigenes*	Sunflower seed	S. Africa	1987
FGS19	3425	*Aspergillus*	*hancockii*	Soil	Australia	1989
FGS20	3643	*Aspergillus*	*near flavus*	Flour	Australia	1988
FGS21	3815	*Aspergillus*	*ochraceus*	Sand	Australia	1990
FGS22	4430	*Penicillium*	*nordicum*	Salami	Australia	1993
FGS23	5399	*Aspergillus*	*ochraceus*	Soil	Australia	2000
FGS24	5400	*Aspergillus*	*near hancockii*	Soil	Australia	2000
FGS25	5421	*Aspergillus*	*Q102*	Soil	Australia	2001
FGS26	5424	*Aspergillus*	*Q105*	Soil	Australia	2001
FGS28	5641	*Aspergillus*	*N184*	Soil	Australia	2001
FGS29	3913	*Aspergillus*	*carbonarius*	Sultanas	Australia	1978
FGS30	N69	*Aspergillus*	*N69*	Soil	Australia	2001
FGS31	965	*Penicillium*	*verrucosum*	Neotype, unrecorded	Belgium	1920
FGS32	5944	*Aspergillus*	*westerdijkiae*	Coffee beans	Vietnam	2007
FGS33	5945	*Aspergillus*	*westerdijkiae*	Coffee beans	Vietnam	2007
FGS34	5946	*Aspergillus*	*westerdijkiae*	Coffee beans	Vietnam	2007
FGS35	5948	*Aspergillus*	*steynii*	Coffee beans	Vietnam	2007
FGS37	5191	*Aspergillus*	*parasiticus*			
AF4351	4351	*Aspergillus*	*flavus*	Peanut shell	Australia	1991
AF3698	3698	*Aspergillus*	*flavus-oryzae*	Soy sauce fermentation	Singapore	
AF2336	2336	*Aspergillus*	*flavus-oryzae*	Miso koji	Korea	
AF3809	3809	*Aspergillus*	*flavus-oryzae*		Thailand	

1.2 Scientific Contributions and Discoveries

The results and findings of this study have made the following important contributions to scientific knowledge:

- The construction of a high quality training set as part of the development of a rapid and accurate Naïve Bayes classification tool. When an unknown query fungal DNA sequence is passed to the classifier, it formulates a full taxonomic assignment from ranks of domain down to species, to identify the species that the query sequence originates from. The ability to classify at the lowest taxonomic level of species represents an improvement over the currently state-of-the-art classifier which only assigns taxonomies down to the genus level. The training set and classifier together are expected to be a valuable asset to fungal biologists that can be integrated into fungal analysis and metagenomic studies as an efficient and accurate classification step.

- An enhanced understanding of the evolutionary dynamics of the aflatoxin genes through a focus on novel fungal genomes. From the results obtained, it is apparent that the genes studied are undergoing similar evolutionary processes. In particular, the findings suggest that *Aspergillus oryzae* is not a distinct species from *Aspergillus flavus*, which represents a discovery in an unresolved area of fungal taxonomy that is subject to much debate. It is envisaged that the findings presented will prove useful in achieving the ultimate goal of devising new, more effective strategies to minimise the risk of aflatoxin contamination in crops and improve the quality of foods derived from them.

1.3 Thesis Structure

Due to the double-faceted nature of this study, this thesis has been divided into 2 parts to address the objectives of both studies. Part 1 describes the Naïve Bayes fungal classification tool, which begins in Chapter 2 with a review of existing classification tools and a justification of the tool that was selected for adaptation to achieve the goals of this study. The chapter concludes with a detailed explanation of the underlying algorithm of the classification tool to be developed. The methods utilised in developing and evaluating the training sets, validation sets and the classifier itself are presented in Chapter 3. The results obtained from the various tests and experiments to evaluate the performance of the classifier are discussed in detail in Chapter 4, and is concluded with a discussion of the limitations and suggestions for future work to improve the performance of the classifier.

Having devised an effective tool for fungal species classification, Part 2 of the thesis forms a natural extension of the work from Part 1 by focusing on the group of agriculturally and medically important species from the *Aspergillus* genus that are known to produce aflatoxins (Section 1.1). Chapter 5 summarises current knowledge about the aflatoxin genes and their evolutionary behaviour. This is followed, in Chapter 6, by an explanation of the *in silico* analysis pipeline employed to study the evolutionary dynamics of the aflatoxin genes in the novel fungal genomes. Finally, the findings, limitations and conclusions of the study are presented in Chapter 7.

The thesis concludes in Chapter 8 by synthesising the findings of both studies and how we plan to further extend the work in the present study.

Part 1

Fungal Classification

CHAPTER 2

Introduction to Fungal Classification

2.1 Overview

This chapter provides a deeper background to fungal classification in terms of the problems associated with fungal taxonomy and effective target genes or regions of DNA to use as the basis of classification. The literature review presented in Section 2.4 summarises and provides a critical evaluation of existing classification tools. The chapter concludes with a detailed explanation of the algorithm of classification tool deemed most appropriate for the purposes of the present study in Section 2.5.

2.2 Fungal Taxonomy

Up until the middle of the 20th century, fungi were considered as belonging to the plant kingdom. It was not until after this time that fungi were separated and classified as a distinct kingdom (Moore, 2013) within the Eukaryote domain, alongside the animal and plant kingdoms (Figure 1.2). Although fungi are recognised as distinct from plants, fungal taxonomy still has its origins in plant taxonomy, and has thus retained the nuances of plant taxonomy. The primary criticism of this taxonomy is the dual nomenclature, which permits different names for the anamorphs (asexual stage) and teleomorphs (sexual stage) of the same species. For example, *Aspergillus nidulans* is the anamorph of *Emericella nidulans*, while *Nectria haematococca* is the teleomorph of *Fusarium solani*. This problem is amplified as some organisms, such as *Fusarium*, have up to seven teleomorphs (Hibbett and Taylor, 2013), and some fungal species having different names in different parts of the world.

As explained in the Introduction, about 100,000 fungal species have been taxonomically identified under the current nomenclature, however there are more than 400,000 fungal species names that have been

recorded in literature (Hibbett and Taylor, 2013). This discrepancy between the numbers reflects the extent of redundancy and synonymous names that have arisen under the dual nomenclature system. This is highly problematic especially when evaluating the accuracy of fungal classification tools, as it is difficult to judge whether an incorrectly classified sequence is truly a misclassification or an artefact of the teleomorph-anamorph relationship, in which case it should not be treated as incorrect. This was a major problem encountered during the evaluation of the training sets and classifier developed in this study as described in Section 4.3.

In 2011, the International Botanical Congress abolished the dual nomenclature system and instead adopted the 'one fungus, one name' principle (Hibbett and Taylor, 2013). This single nomenclature system has numerous benefits, including more effective and ease of communication. Hibbett and Taylor (2013) report that fungal biologists will need to work together to resolve fungal species names for the new nomenclature, especially medically, agriculturally and industrially important organisms and pathogens. In the meantime, this feature of the taxonomy must be kept in mind and dealt with appropriately.

2.3 Ribosomal RNA Genes for Classification

Traditional methods of classification and identification of fungal organisms was achieved using their cultural and morphological characteristics such as shape and size. These techniques, however, are unreliable and highly prone to misclassification (Chang and Ehrlich, 2010).

The advent of DNA sequencing technologies caused a shift towards sequence-based strategies for species comparison and identification, which have become the norm today (Balajee et al., 2009). Briefly, this method involves selecting a target gene or region of DNA, extracting and amplifying this region using a process called Polymerase Chain Reaction (PCR) and sequencing the resulting amplicons. The organism from which the sequences originate are then identified by querying against a database and using a similarity metric to assign the species, or using other techniques such as phylogenetic analyses.

The success of sequence-based identification techniques lies primarily in the selection of an appropriate gene target, which highly impacts the ability to accurately classify unknown or novel genome sequences

FIGURE 2.1. Structure of the rRNA region in fungi. The highly variable ITS1 and ITS2 (light blue) flanking the highly conserved 5.8S region (dark blue) together comprise the ITS region. The entire ITS region itself is flanked by the more conserved 18S SSU and 28S LSU genes (green).

from a particular group of organisms. Balajee et al. (2009) summarise the criteria that such an ideal target gene should satisfy:

- Evolution by common descent i.e. is orthologous
- Occur universally in all organisms of interest
- Have high levels of interspecies variation and low levels of intraspecies variation in order to distinguish between the organisms with high sensitivity
- Not undergo recombination, which is the exchange of genetic material to produce new combinations of genetic content
- Easy to extract, amplify and sequence
- Have a length that falls within the range of sequence reads produced by the common sequencing technologies, i.e. 600 – 800 base pairs (bp)
- Be easily aligned with similar sequences in a database for comparison

In bacterial and fungal genomes, a commonly used marker is the region containing the ribosomal RNA (rRNA) genes, which has been successfully applied in taxonomic assignments of microbial genome sequences and other phylogenetic analyses (Porter and Golding, 2012; Schoch et al., 2012). In bacteria, the 16S rRNA gene is commonly used as the gene target for species identification. The corresponding rRNA structure in fungi (Figure 2.1) consists of the variable Internal Transcribed Spacer (ITS), defined as the ITS1-5.8S-ITS2 region, flanked by the more conserved 18S small ribosomal subunit (SSU) and 28S large ribosomal subunit (LSU) genes.

The 18S SSU and 28S LSU are important genes encoding the small and large ribosomal subunits respectively that make up the cellular machinery required for the production of proteins. In contrast, the ITS1 and ITS2 regions presently have no known function, and are highly variable in their lengths. ITS1 is about 180 bp in length while ITS2 is about 170 bp (Nilsson et al., 2010). The length of the

entire ITS region typically ranges between 400 bp and 1500 bp (D Midgley and N Tran-Dinh, personal communication).

To address the lack of a standard, official barcode for fungal species identification, a recent international consortium assessed six DNA regions as potential barcodes or fungal sequence signatures (Schoch et al., 2012). Their findings were formalised by Schoch et al. (2012) who further investigated the "barcoding performance" of three rRNA regions (SSU, LSU, and ITS) along with the RNA polymerase II large subunit (RPB1) gene, using Probability of Correct Identification (PCI) and barcode gap analysis as metrics for their evaluation. The fungal SSU gene, despite reported applications in phylogenetic analysis, lacks the extent of hypervariable regions that are present in the equivalent bacterial 16S gene. The LSU and ITS regions, either on their own or combined with each other, are popular choices for species identification, however the LSU has been more successfully applied at taxonomic ranks of genus or above (Figure 1.2). The ubiquitous nature and low selection pressure (slow divergence rate) of the RPB1 gene justifies its candidature as a potential fungal barcode marker. The study found that RPB1 and ITS consistently had higher PCI scores for all species except higher ancestral species (early diverging lineages) where SSU and LSU had greater discriminatory power. The results of the barcode gap analysis provided additional confirmation, with only RPB1 and ITS having statistically significant barcode gaps that captured both inter- and intra-species variation. This is in agreement with the fact that the ITS region evolves rapidly, giving rise to variation between species or populations (White et al., 1990).

Despite the RBP1 gene having the most success at species identification, Schoch *et al* report that unlike the SSU, LSU and ITS regions, amplification and sequencing of RBP1 had the most problems and lowest sequencing success rate (Schoch et al., 2012). This inability to efficiently and successfully extract the RBP1 gene sequence from fungal organisms thereby invalidates its use as a barcode. The ITS region, which performed second best to RBP1, is also not without its limitations. Some taxonomic groups have low ITS interspecific variation, including the *Aspergillus* and *Penicillium* genera in which ITS sequences are nearly identical across several species. In *Aspergillus flavus* and *Aspergillus minisclerotigenes* for example, the ITS sequences differ by 1 base only (D Midgley, personal communication). The ITS region also lacks sufficient variability to distinguish between the species *Rhizopus azygosporus* from *Rhizopus microsporus* (Balajee et al., 2009). Furthermore, the high genetic diversity among fungi may render a single barcode system as inadequate, demanding a secondary marker or combination of markers, such

as ITS and LSU, to accurately distinguish closely related species or those in narrow taxonomic groups. Taking all these factors and challenges into consideration, the authors proposed ITS to be the standard barcode as the conclusion of their study, which also satisfies most of the above criteria proposed by Balajee et al..

Given the relative ease of sequencing the ITS region compared to its neighbouring SSU and LSU genes (Schoch et al., 2012), it comes as no surprise that ITS has now become the most widely sequenced DNA region in fungi, and is the preferred marker for molecular identification of most fungal organisms (Balajee et al., 2009; Bates et al., 2013). As a natural consequence, there are more ITS sequences available in public databases, and hence a classifier built using sequences from this region would be more robust and accurate as there is more data available for training.

2.4 Classification Tools for Species Identification

The previous section established ITS as the region against which sequence based identification strategies should be targeted. Following on, classification tools that have been developed for the purposes of species characterisation to serve microbial and metagenomic studies are now discussed. Tools using the ITS region as the basis for fungal species identification form the focal point of examination. This is complemented by a wider survey of tools based on other fungal genes, as well as for different types of organisms, which are investigated as a means to evaluate their potential to be adapted for fungal classification.

Classification tools are highly diverse in their underlying algorithm with respect to how the classifications are formulated. Most tools are based on either sequence similarity, phylogeny or sequence composition; these categories form the subsections of this review.

2.4.1 Similarity-based Classifiers

Similarity-based tools are commonly used for the classification metagenomic reads from environmental samples (Porter and Golding, 2012). They often perform searches against a reference database using the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) and use a devised measure of similarity to determine the best match, which then forms the taxonomic assignment. The results

of a BLAST search include the percentage of query coverage, percentage of identity in the alignment (the number of sites that are identical between the query and reference sequence), a bit score and an Expectation value (E-value). The E-value is essentially a significance threshold that reflects background noise in the database and measures the probability of a match being returned purely by random chance (`http://www.ncbi.nlm.nih.gov/BLAST`). Hence matches with lower E-values, corresponding to higher bit scores, are more likely to be true and not the result of random chance.

PlutoF (Abarenkov et al., 2010) was designed to provide a workbench to aid ecological and taxonomical analysis pipelines through cloud based databases and resources. The sequence analysis module has been optimised for identification of environmental fungal ITS sequences, and can be modified for different gene targets and groups of organisms. Taxonomy assignments are executed through a BLAST search, against the International Nucleotide Sequence Database (INSD) and UNITE database (Kõljalg et al., 2005), and the top hit as judged by the bit score is typically selected as the classification (Dannemiller et al., 2013). Often, however, a BLAST search will return different matches with the same top E-value, which presents a major problem to deciphering and choosing which of the top hits is the correct match upon which to formulate the assignment. There is no evidence to suggest that PlutoF attempts to resolve matches with identical top E-values; the topmost match is simply returned as the taxonomic assignment. It has been noted that the top BLAST match is not always the closest phylogenetic neighbour to the input query sequence that was searched (Porter and Golding, 2012). This observation demands more sophisticated methods to handle the output of BLAST searches.

Huson et al. (2007) developed Metagenome Analyser (MEGAN) to deal with tying top BLAST search hits. The authors claim their approach is more practical, especially for processing large data sets, as there is no requirement for assembly or specific phylogenetic markers to be targeted for sequencing. The program takes as input the BLAST output and processes it to collect all the sequences in the database that were returned as significant matches to the query sequences and had a higher score than the user-specified threshold. The sequences are retrieved along with an annotation or taxonomy as defined by the National Center for Biotechnology Information (NCBI). MEGAN then applies a Lowest Common Ancestor (LCA) algorithm to classify a query sequence by assigning it to the lowest rank that contains all the top hit sequences under it in the taxonomy. Huson *et al* implement an improvement to the LCA algorithm in the newer version of their program, called MEGAN4 (Huson et al., 2011). A sequence is

assigned to the lowest common rank only if the number of query sequences assigned to that taxon are greater than a minimum support threshold. If the number of query sequences at the current lowest rank is less than the threshold, all query sequences are re-assigned and propagated up the taxonomy to the parent rank which satisfies the support threshold. The rank at which the assignment is made is intended to reflect the level of conservation in the query sequence compared to the matched sequences. Sharma et al. (2012) view that MEGAN4's consideration of BLAST hits, which is based solely on bit scores, may lead to more non-specific taxonomic assignments due to the incorporation of both expected and unexpected hits that arise from re-assigning sequences to higher ranks.

Monzoorul Haque et al. (2009) also report the high false positive rate and low specificity assignments observed with MEGAN, especially when the query sequences were from novel organisms. This prompted the development of Sequence ORTholog based approach for binning and Improved Taxonomic Estimation of Metagenomic Sequences (SOrt-ITEMS) by Monzoorul Haque et al. to address these limitations. Rather than using the bit score of the BLAST search solely, SOrt-ITEMS assesses the quality of the alignment between query sequence and the matched sequences from BLAST, and uses additional alignment parameters, such as percentage identity, to determine an appropriate rank in the taxonomy where the query can be assigned. This allows for sequences with higher quality alignments with BLAST hits to be classified to more specific ranks such as genus and species. The algorithm then identifies hits with significant orthology, or reciprocal similarity, with the query sequence, and these orthologous hits are then passed through the LCA algorithm to make the final assignment of the query sequence.

The authors of SOrt-ITEMS improved the algorithm in their new program Distance Score Ratio for Improved Binning And Taxonomic Estimation (DiScRIBinATE) (Ghosh et al., 2010). The first stage of SOrt-ITEMS, which reduces the number of false positives and identifies the appropriate rank for classification, is retained in DiScRIBinATE. In the second step which ensures specificity of the assignment, rather than using the orthology-based approach which requires the time consuming reciprocal BLAST searches, DiScRIBinATE uses the quicker computation of the ratio of bit scores and a distance measure between each query sequence and its BLAST hits. Given these optimisations, Ghosh et al. report a reduction in misclassification error rate of 1.5 – 3 times compared to SOrt-ITEMS and 3 – 30 times compared to MEGAN.

Dannemiller et al. (2013) recently developed the tool Fungal Hi-throughput Taxonomic Identification tool for use with Next-Generation Sequencing (FHiTINGS), which adopts a similar procedure to MEGAN (Huson et al., 2007). FHiTINGS also requires the output of a BLAST search of the query sequences and like MEGAN, applies the Last Common Ancestor algorithm to resolve the identically scored but differently annotated best BLAST matches. Because the algorithm reassigns sequences to the lowest common rank, fewer assignments are made to low ranks such as genus and species, however the higher rank assignments have less error. The differentiating factors between FHiTINGS and MEGAN are the databases against which the BLAST searches are performed, and the taxonomy used to make the classification. MEGAN requires BLAST results from the NCBI database and uses the NCBI taxonomy for taxonomic assignments. On the other hand, FHiTINGS requires BLAST results from an ITS-only fungal database in which all sequences are annotated down to the species rank, and uses the Index Fungorum database (`http://www.indexfungorum.org/`) as the basis of its classifications. While MEGAN is more general and can be applied to any gene, Dannemiller et al. claim that having reference and taxonomic databases specific to fungal species benefits their classifier, especially as these databases are subject to extensive curation by fungal experts.

A different strategy to the LCA algorithm is employed by White et al. (2013) in their program CloVR-ITS, an automated pipeline for the comparative analysis of ITS sequences from metagenomic samples. This program constructs a custom reference database of ITS sequences by first obtaining ITS2 sequences and using the accession numbers to retrieve the NCBI entry containing the whole ITS region, including ITS1, and extending to the 18S SSU and 28S LSU genes. While the addition of the 28S LSU sequences may help enhance taxonomic resolution, the incorporation of the 18S SSU sequences is questionable due to its less variability and therefore less discriminatory power (Section 2.3). The final database contained 154,050 sequences spanning 8,440 genera and over 60,000 species. Query sequences are classified using a BLAST search against this custom ITS database using an E-value threshold of 10^5. The assignment is made to the BLAST match where the alignment of the query sequence and BLAST hit has at least 90% query coverage and minimum identities of 90% (species), 85% (genus), 75% (family), 70% (order) and 60% (class). These minimum identity thresholds for each rank were determined by conducting pairwise alignments of ITS2 sequences at that rank and taking the average identity found between all alignments. Whether this set of identity thresholds can be applied universally to taxonomic assignments remains to be elucidated.

2.4.2 Phylogeny-based Classifiers

It is believed that classifications using similarity-based BLAST searches against a reference database, although commonly used, may not be reliable and robust. First and foremost, resolving matches with top BLAST hits (those with identical E-values) has been identified as a significant issue, and an additional problem that BLAST-based classifiers must deal with. Even if only one top match is returned, this is not guaranteed to be the closest match in the phylogenetic sense to the query sequence (Porter and Golding, 2012). Furthermore, Porter and Golding report that BLAST does not have the ability to automatically classify the query sequence to higher taxonomic ranks where the accuracy would be higher than at the current classified rank.

Munch et al. (2008) argue the statistical problems of BLAST matches. The score of the match is calculated using local and not global alignments which causes a loss of information. The search against the database does not take into consideration the population genetic and phylogenetic issues pertaining to species identification. Lastly, the confidence values supplied with BLAST searches measure significance of local sequence similarity rather than significance of taxonomic assignment.

The concept of phylogeny-based classifiers is illustrated through the Statistical Assignment Package (SAP) (Munch et al., 2008), the development of which was motivated by the statistical problems of BLAST-based classification outlined above. The most pertinent feature of SAP, not offered by BLAST (Altschul et al., 1990) or MEGAN (Huson et al., 2007), is a measure of statistical confidence for each classification, a feature important for empowering the user to judge the reliability of the taxonomic assignments for themselves (Munch et al., 2008). Unlike BLAST, SAP detects situations where an assignment is ambiguous and has low confidence at the species level, to assign the sequence to higher ranks such as genus and family.

The procedure devised by Munch et al. (2008) begins by using a BLAST search and annotation information from the NCBI Taxonomy database to create a set of homologs, or near identical sequences to the query sequence. The query sequence with its homologs are aligned using the ClustalW aligner, and the resulting multiple alignment informs the generation of phylogenetic trees using either the Markov Chain Monte Carlo (MCMC) or Neighbour-Joining (NJ) methods. The assignment is made as the homolog

which forms the adjoining branch to the query sequence. Bootstrapped trials sample the trees and are used to calculate the probabilities that estimate the confidence values of the assignment.

In terms of the performance of SAP, only one set of results is presented in the paper, which displays confidence values around 30% for genus and species ranks, values too low to be reliable. While more data is required for a valid assessment, the classifier seems to be more accurate at family ranks and above. Hence this approach was not deemed viable for our purposes of building an accurate species level classifier.

2.4.3 Composition-based Classifiers

In contrast to both similarity-based and phylogeny-based classifiers, machine learning and statistical techniques are at the core of composition-based classifiers that are described in this section.

The Ribosomal Database Project (RDP) Classifier (Wang et al., 2007) available at `http://rdp.cme.msu.edu/classifier`, was developed as a fast and accurate classification tool to characterise bacterial organisms from environmental samples using 16S rRNA gene sequences. Its popularity and widespread use as a classification tool among biologists is evidenced from its 1648 citations. The classifier is available as an online service as well as a command line tool which offers the user the option between using the built in bacterial 16S training set or to train the classifier using their own training set.

One of the goals of the classifier is to be compliant with the short sequence reads produced by newly emerged Next-Generation Sequencing (NGS) technologies (refer to Section 6.2.1 for a full explanation about sequencing). Wang et al. therefore developed the classifier with the ability to handle sequences as short as 50 bp.

There is no mention in the paper about a data or sequence processing step prior to training and building the classifier, hence it is assumed that no such processing was performed. It is likely that this step was deemed unnecessary as the two training sets used – Bergey's corpus and the NCBI corpus – were obtained from published and readily available corpora, and contained 5014 and 23,095 16S sequences respectively. The corpora of 16S sequence permit taxonomic classification at the genus rank. The classifier is currently in the ninth iteration of the training set since it was made available in the public domain in 2007.

In terms of the underlying algorithm, the classifier is a Naïve Bayes Classifer, which uses information about the query sequence composition itself, hence categorised as a 'composition-based' classifier. This eliminates the need to perform computationally expensive alignments with sequences in a reference database as is done in similarity-based methods (Section 2.4.1), and the exclusion of this step permits faster classification times. Briefly, the classifier uses knowledge about prior conditional and joint probabilities of k-length subsequences (k-mers) at the genus rank to assign each query a rank from genus up to domain through the application of Bayes' Theorem. The performance of the training set was assessed by simulating sequences ranging from 50 bp to full 16S sequence and using the validation method of Leave-One-Out-Cross-Validation (LOOCV). Not surprisingly, the best results were obtained from the full 16S sequences which had an accuracy of 91.4% at the genus level.

In their phylogeny-based classifier SAP, (Munch et al., 2008) highlight the importance of providing a statistical measure of confidence along with the classifications, to provide more meaning to the classifications and enabling the user to judge the reliability of the taxonomic assignments using their biological expertise (Section 2.4.2). Like SAP, this attractive feature is also offered in the RDP Classifier, which performs bootstrapping to estimate confidence values that are supplied with the classification for each rank. A detailed and thorough explanation of the RDP Classifier algorithm is presented in following Section 2.5.

The capabilities of the RDP Naïve Bayes Classifier were extended by Liu et al. in (2012) to support fungal classification. This classifier is based on the fungal 28S LSU gene and as a result, is only able to classify down to genus level. The authors reason that the 28S LSU gene was chosen over the ITS region as it has sufficient discriminatory power and results in more robust alignments. A training set of 8,506 high quality 28S LSU gene sequences were manually curated and processed to produce a final training set of 7,737 sequences. Unlike the RDP 16S classifier (Wang et al., 2007) in which no extensive processing and curation was required, the LSU gene sequences were processed to remove sequences with none or conflicting taxonomies, and sequences that occurred uniquely in the training set. Since the majority of sequences were not more than 1400 bp in length, the starting position of the LSU was inferred and first 1400 bp of the gene extracted, which was made possible from the alignment of all sequences against a reference sequence.

Extensive testing was conducted to evaluate the performance of the training using LOOCV which was done for sequence lengths of 100 bp and 400 bp to reflect the nature of sequencing reads produced by commonly used NGS technologies. Furthermore, the alignment of the sequence enabled Liu et al. to test the classifier using a sliding window of 100 bp and 400 bp sequences to elucidate which specific region within the LSU gives the most accurate results. It was found that the best results were obtained using 400 bp sequences across the hypervariable D2 region, which gave an accuracy of 92% at the lowest rank of genus. While the training set evaluation is highly elaborate, the paper would benefit from a similar analysis of a test set consisting of new sequences not present in the training set as a more comprehensive evaluation of the classifier.

As mentioned previously, the Naïve Bayes classifier is computationally faster than its similarity-based equivalent tools (Section 2.4.1) as it does not need to perform sequence alignments of each query sequence against a database. This is evidenced by Liu et al.'s observation of a 460-fold increase in speed compared to the BLAST similarity-based approach, with similar or better accuracies achieved.

A similar k-mer probability based approach forms the basis of the QUadratic, K-mer – based, Iterative, Reconstruction (Quikr) program recently developed by Koslicki et al. (2013). Currently designed for bacterial 16S sequences, the program aims to characterise and reconstruct a metagenomic sample by identifying the species contained within it. Quikr assumes that the species in the given environmental sample are present in the training set, however it still performs well for novel species in the sample. The output of the classifier is a single probability for each sequence in training set, that is, the probability of each sequence in the training set being present in the test set (environmental sample). Although Koslicki et al. report that Quikr is much faster than the RDP Classifier (Wang et al., 2007), Quikr lacks features offered by the RDP Classifier including a full taxonomic prediction with confidence values at each rank, and the ability to evaluate the training set using an in-built validation method such as LOOCV.

2.4.4 Summary

Porter and Golding (2012) surveyed and compared the performance of BLAST + MEGAN (similarity-based), SAP (phylogeny-based) and the RDP Classifier (composition-based) using a compiled set of

TABLE 2.1. Summary of the classification tools discussed for each type of category. Note that the lowest rank can change depending on the lowest rank of sequences in the training set.

Program	Organisms	Genes Used	Lowest Rank	Type	Reference
PlutoF	Fungi	ITS	Species	Similarity	Abarenkov et al., 2010
MEGAN	All	Any	Species	Similarity	Huson et al., 2007
SOrt-ITEMS	All	Any	Species	Similarity	Monzoorul Haque et al., 2009
DiScRIBinATE	All	Any	Species	Similarity	Ghosh et al., 2010
FHiTINGS	Fungi	ITS	Species	Similarity	Dannemiller et al., 2013
CloVR-ITS	Fungi	ITS	Species	Similarity	White et al., 2013
SAP	All	Any	Species	Phylogeny	Munch et al., 2008
RDP Bacterial 16S	Bacteria	16S	Genus	Composition	Wang et al., 2007
RDP Fungal LSU	Fungi	28S LSU	Genus	Composition	Liu et al., 2012
Quikr	Bacteria	16S	Species	Composition	Koslicki et al., 2013

28S LSU gene sequences and measured the LOOCV accuracies at the genus rank. Their results found that BLAST + MEGAN had the lowest error rate and highest robustness in the presence of sequence errors. When longer LSU sequences were used, SAP achieved the highest accuracy. The RDP Classifier was the fastest by far, taking minutes for classifications that took hours to days for BLAST + MEGAN and SAP. It is important to note that with the RDP Classifier, if an incorrect taxonomic assignment is consistent among bootstrap replicates, the incorrect assignment will be applied and supported by a high confidence value. Regardless, Porter and Golding endorse the use of the RDP Classifier for rapid taxonomic assignment as the conclusion of their study.

This section critically reviewed available types of classification tools across three categories, highlighting the strengths and weaknesses of each tool. This led to the conclusion that the speed and high accuracy demonstrated by composition-based classifiers make them highly applicable for the objectives of this study as outlined in the Introduction. There is clear evidence that a composition-based classifier built using ITS sequences and that can resolve down to species level, has not been developed, and forms the motivation and objective of the current project.

Of the composition-based classifiers presented in Section 2.4.3, the RDP Classifier (Wang et al., 2007) has been highly successful, as evidenced by its wide use and trust among biologists. Furthermore, the classifier has undergone several stages of extensive testing. Although the RDP Classifier at present only contains bacterial 16S sequences and fungal 28S LSU sequences, its proven success makes it an attractive choice as the foundation for a new ITS classifier. Therefore, rather than building a completely

new classifier from scratch, the decision was made to adapt the RDP Classifier for fungal ITS sequences. Before proceeding to explain the methodology used to achieve this, a detailed explanation of the RDP Classifier algorithm is presented in the following section.

2.5 RDP Naïve Bayes Classifier

The RDP classifier is implemented as a Naïve Bayes classifier, named after the 'naïve' and unrealistic assumption made regarding the independence of the data attributes. Despite the violation of this assumption, the classifier has been shown to work reasonably well in practice, extending to applications such as spam email filters and text categorisation. The Naïve Bayes classifier is a supervised classification method, which means that it has prior knowledge of the class or label, and in this case taxonomic ranks, of all the sequences that comprise the training set. These labels are assumed to be the 'true' or 'gold standard' label, necessitating that each sequence has an accurate taxonomic assignment.

2.5.1 k-mer Feature Space

The algorithm is implemented using a k-mer based approach to create a feature space of all possible k-length subsequences or 'words', which forms the basis of training and classification. A k-mer is defined as a DNA sequence of length k bases. The choice of an appropriate value of k is a critical decision as it affects accuracy of taxonomic assignment. Small values of k may be too non-specific to achieve the desired sensitivity, while longer k-mers may suffer from over-specificity and may not result in sufficient matches. For their bacterial 16S classifier, Wang et al. conducted empirical experiments for values of k between 6 and 9, and it was found that $k = 6$ and $k = 7$ gave less accurate results, while $k = 8$ and $k = 9$ had results of comparable accuracy. Since the feature space created by 8-base words is much smaller than the corresponding feature space produced 9-base words, the authors chose 8 as the optimal value of k. Similarly, Liu et al. performed entropy measurements to determine the optimal value of k for their LSU classifier, and also found $k = 8$ as the optimal word size. Each position in an 8-base word can be any of the 4 nucleotides in DNA (A, C, G and T), giving a total of $4^8 = 65,536$ possible words in the feature space, which is considerably less than a feature space of 262,144 words that would result with $k = 9$.

2.5.2 Statistical-based Classification

The Naïve Bayes classifier uses a statistical classification approach based on Bayes' Theorem, so training of the classifier entails the calculation of two sets of probability values. The first is the prior probability P_i for each k-base word in each training sequence, which measures the likelihood of observing word $w_i \in W$ in a sequence in the training set. The prior probability is calculated using Equation (2.1), where $n(w_i)$ is the number of sequences containing word w_i and N is the total number of sequences in the training corpus. The values 0.5 and 1 are added to the numerator and denominator respectively to ensure the final probability values lie in the range of zero to one.

$$P_i = \frac{n(w_i) + 0.5}{N + 1} \tag{2.1}$$

The second set of probabilities calculates the conditional probability that a sequence Q belonging to a specified rank R contains word w_i (Equation (2.2)), and the joint probability that a sequence Q of rank R contains a set of words $v_i \in V, V \subseteq W$ (Equation (2.3)). Note that rank R can be any taxonomic rank specified; the RDP 16S (Wang et al., 2007) and RDP LSU (Liu et al., 2012) classify at the genus rank, while the new ITS classifier developed in this study classifies at the species rank. For a given class C at rank R represented by M sequences in the training set, $m(w_i)$ denotes the number of sequences from class C that contain word w_i.

$$P(w_i \mid R) = \frac{m(w_i) + P_i}{M + 1} \tag{2.2}$$

$$P(Q \mid R) = \prod P(v_i \mid R) \tag{2.3}$$

The probability that a query sequence Q belongs to rank R is calculated by applying Bayes' Theorem:

$$P(R \mid Q) = \frac{P(Q \mid R) \times P(R)}{P(Q)} \tag{2.4}$$

Under the assumption that all members of rank R are equally likely, the terms $P(R)$ and $P(Q)$ become constant and can be dropped from the equation, leaving only the term $P(Q \mid R)$ on the right hand side. Equation (2.4) then simplifies down to Equation (2.3):

$$P(R \mid Q) = P(Q \mid R) = \prod P(v_i \mid R) \tag{2.5}$$

The sequence Q is assigned to the class (at the specified rank) with the highest probability, and this classification is extended to make an assignment at every rank from R up to domain, to gain a full taxonomic prediction.

2.5.3 Bootstrapping

As mentioned previously, a feature of the RDP Classifier are confidence values accompanying the prediction for each rank, and are estimated using bootstrapping (Wang et al., 2007). Under the assumption of independent attributes, the size of the bootstrap sample is set to the number of features in the original sample. The overlapping set of k-mers from each sequence violates this assumption, with the number of independent features equalling the number of non-overlapping k-mers, which is calculated as the total number of k-mers N divided by k.

For each out of 100 bootstrap trials, a subset of words of size N/k is randomly selected with replacement, and this set of words is used in the calculation of the joint probability in Equation (2.3). The frequency with which a class C at rank R is assigned to the query sequence approximates the confidence of the assignment to class C. The confidence of rank above R up to domain are calculated by summing the confidence values of all members C that belong under that taxon.

CHAPTER 3

Methods for Fungal Classification

3.1 Overview

The literature review conducted in Chapter 2 established the effectiveness of compositional based species classification methods such as a Naïve Bayes classifier, which justifies its use in the current research project. The performance of our fungal ITS Naïve Bayes classifier is inherently dependent on the quality of the ITS sequences that form the training set. Otherwise stated, the accuracy of the classifications is bounded by the quality of the reference data set upon which the classifier is built. The Naïve Bayes classifier is a supervised classifier that relies on pre-classified, known training examples; so creating a set of fungal ITS sequences that were high quality, well-curated, had broad species representation and taxonomically accurate classifications, was therefore a challenging yet imperative task. This chapter describes the steps taken in the attainment and processing of ITS sequences, the creation and evaluation of training sets, and the evaluation of the classifier on a specifically designed validation set. The steps in each stage were automated with the aid of in-house, custom written Python scripts that I wrote to accomplish this task.

3.2 Fungal ITS Sequence Pre-processing and Curation

For bacterial 16S sequences, there are publicly available corpora containing high-quality, validated sequences along with their taxonomies, which were directly used by Wang et al. as their training sets (Section 2.4.3). In comparison, much less effort has been directed towards building high-quality reference sets for fungal sequences, including the 28S LSU gene and ITS region, which has been recognised

as a limitation in furthering fungal research (Bates et al., 2013). Thus curation of fungal sequences becomes a necessary step in the development of a training set; this was undertaken by Liu et al. for their fungal LSU classifier and we report the same here for our ITS classifier.

Currently, the largest and most popular databases for genome sequences available in the public domain are hosted by the National Centre for Biotechnology Information (NCBI), European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ). These databases collectively form the International Nucleotide Sequence Database (INSD). A major problem encountered more often than not is that genome sequences deposited in these public databases are of poor quality, contain sequencing errors, have incorrect or missing gene annotations and erroneous taxonomic assignments (Kõljalg et al., 2005; Bates et al., 2013). This is often accompanied by non-validated information, which only the person who deposited the sequence has the authority to correct and update (Balajee et al., 2009).

These factors motivated the construction of the UNITE database (`http://unite.ut.ee`), which contains high-quality ITS sequences manually checked by experts (Kõljalg et al., 2005), as well as all the ITS sequences present in INSD. The UNITE database was recently endorsed as a sound foundation upon which to build high quality ITS reference datasets (Bates et al., 2013). Following this ratification, a dataset was downloaded from UNITE which contained 343,809 fungal ITS sequences (as of 10 May 2013), and formed the starting point for the development of our training set. The sequences were subjected to extensive processing and curation discussed in detail in this section, and resulted in 24,447 sequences that formed the final training set of the classifier.

The file of sequences downloaded is in FASTA format, where each sequence is associated with a descriptive 'header' of the form:

```
> Unique accession number | INSD organism name | INSD taxonomy | UNITE organism name | UNITE taxonomy
```

Each sequence was assigned either an organism name and taxonomy from either INSD or UNITE, or both. Sequences which did not have an UNITE or INSD lineage were firstly discarded as these lacked any taxonomic information to base the classifications on, leaving 245,590 sequences in the dataset. The next step involved understanding the dataset in terms of the taxonomic composition and sequence

TABLE 3.1. Taxonomic composition of the UNITE ITS dataset. For simplicity, the taxonomy composition is shown at the phylum rank.

Phylum	# Sequences	% of Total
Arthropoda	171	0.070
Ascomycota	146,734	59.750
Basidiomycota	85,032	34.620
Blastocladiomycota	40	0.020
Chytridiomycota	776	0.320
Glomeromycota	7,776	3.170
Incertae sedis	47	0.020
Microspora	723	0.290
Microsporidia	7	0.003
Neocallimastigomycota	442	0.180
Zygomycota	3,842	1.560
Total	245,590	100.000

lengths, which formed further criteria that warranted whether or not a sequence is considered a candidate in the training set.

Table 3.1 clearly displays anomalies in the data set. Arthropoda is a phylum of the insect kingdom rather than the fungal kingdom, while Microspora is a type of algae that is part of the plant kingdom. It is unclear why sequences from these organisms are present in the dataset; one possibility is simply error in constructing the ITS dataset. Regardless, these sequences do not represent fungal ITS sequences and were subsequently removed. The remaining sequences in the data set were primarily from the Ascomycota and Basidiomycota phyla, which together accounted for 94.73% of the sequences. This is expected as the majority of fungal biologists study species that from these two phyla, hence there are more of these sequences available in databases.

An examination of the ITS sequences revealed their lengths to occur between 60 and 2995 bp, with the most frequent lengths being 540 bp and 546 bp. This histogram in Figure 3.1 shows a long right tail that reflects the average length of 1340 bp, which is towards the higher end of the expected range. As ITS sequences generally range from 400 to 1500 bp (D Midgley and N Tran-Dinh, personal communication), it was difficult to discern which exact regions, in terms of the ITS region and greater rRNA operon, the sequences spanned and where exactly the ITS region occurred. Figure 3.2 illustrates the possible

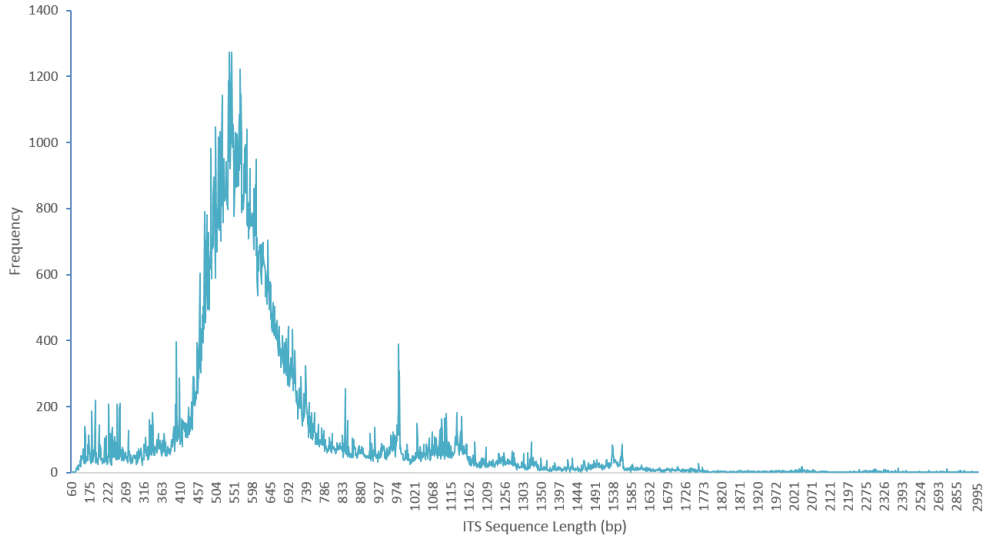


FIGURE 3.1. Length Distribution Histogram for the ITS Sequences in the UNITE dataset.

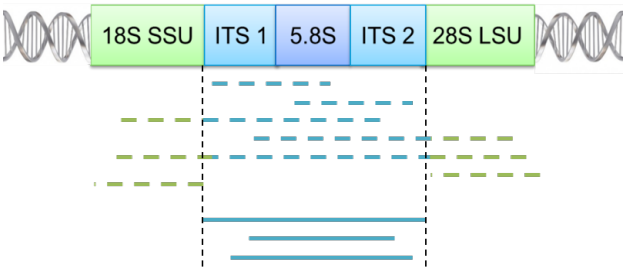


FIGURE 3.2. Possible regions of the rRNA locus the UNITE sequences may represent Dashed lines represent invalid sequences which either do not contain the full ITS region, or contain parts of the adjoining 18S SSU and/or 28S LSU genes. The solid blue lines between the vertical black dashed lines represent the target sequences we wish to obtain that span the ITS region only.

scenarios of what the sequences could represent; the objective is to have sequences that represent the ITS region only (ITS1 – 5.8S – ITS2), as the presence of the more conserved 18S SSU or 28S LSU genes may counteract the high variability offered by the ITS region.

One method of discerning what regions are covered by the sequences would be to align all the sequences against a reference ITS sequence. Given the extent of the variability of the ITS region in both sequence and length, performing an alignment would not be feasible (Bates et al., 2013), reiterating Liu et al.'s decision to base their fungal classifier on the LSU rather than ITS region. An alternative strategy was

to detect the conserved 5.8S subregion that occurs between the ITS1 and ITS2 in all ITS sequences. A set of short probes were designed based on Cullings and Vogler's 5.8S sequences. These probes were passed to a custom tool developed by Paul Greenfield, which filtered out sequences that had matches to any probe. The probes were at least 20 bp in length to ensure high specificity in the matching. Sequences with no matches to the 5.8S region were discarded, as the absence of the 5.8S region indicates that either the ITS1 or ITS2 is missing and automatically renders the sequence invalid. Sequences that had a total length less than 250 bp were also removed in this step as they are too short to be informative, reducing the number of sequences to 227,789.

A quick BLAST search (`http://blast.ncbi.nlm.nih.gov/`) of some of these sequences indeed revealed the presence of 18S SSU and 28S LSU sequences, which accounts for the long tail of sequences with lengths greater than 1500 bp in Figure 3.1. These genes were trimmed, or rather the ITS region was extracted, using a tool called FungalITSExtractor (Nilsson et al., 2010), which took approximately 6 hours to run on the 227,789 sequences. The program works by using short (18 – 25 bp) and long (30 – 50 bp) Hidden Markov Models (HMMs) to detect the boundaries of the SSU, 5.8S and LSU genes, and accordingly extract the ITS1 and ITS2 between them. The structure of DNA consists of two strands that are reverse complements of each other, and the program handles this to ensure sequences in the form of ITS2 – 5.8S – ITS1 are corrected in orientation before being returned. Partial sequences of ITS1 or ITS2 were considered 'valid'; as long as both ITS and ITS2 were present in the sequence along with the 5.8S region. This process retained 221,266 sequences. The output of this program was verified manually by taking conserved region upstream of the start of the ITS1 i.e. the end of the SSU gene, and another conserved region downstream of the end of the ITS2 i.e. the start of the LSU gene, and trimming the sequences based on the existence and positions of the conserved region. Both results in general were consistent with each other, with the sequence boundaries detected by the manual checking mostly occurring 11 bp to the left and 38 bp to the right of the boundaries detected by FungalITSExtractor. These minor discrepancies are likely attributed to the conserved regions used in the manual extraction not being at the very end of the SSU or the very start of the LSU, indicating the program used is more stringent with the ITS boundary detection.

Now that there was consistency in the composition of the sequences, a single taxonomy (either INSD or UNITE) was assigned to each sequence. When only a single taxonomy was present, that taxonomy was

chosen. When both the INSD and UNITE taxonomies were given in the header of the sequence, the more informative one was chosen, that is, the taxonomy that was down to a more specific rank. Sequences that had no species name or had taxonomic inconsistencies, such as other missing ranks, were also removed. This was guided by the known fact that orders have the suffix of '-ales' and family ranks have the suffix '-aceae'. Sequences that were identical to each other, but had conflicting or incompatible taxonomies were resolved using majority to choose the most frequent taxonomy. Taxonomies which were ambiguous were discarded as there was no reliable way of determining which was the correct taxonomy out of the conflicted ones. Since we are resolving to species level, this task was even more arduous due to the taxonomic groupings allowed in fungal nomenclature. When the exact identity of the species is unknown, the species name can include 'sp' (species), 'aff' ('affinis' = related to) and 'cf' ('confer' = compare with) to indicate the current uncharacterised species is similar to another known species, and 'var' (variety) and 'subsp' (subspecies) which refer to taxonomic ranks below species. This needs to be taken into account when resolving conflicting species names, since *Fusarium solani sp pisi* and *Fusarium solani*, and *Aspergillus flavus subsp oryzae* and *Aspergillus flavus*, are essentially the same species and should not be treated as conflicts.

Of the remaining 195,665 sequences, sequences which only occurred once, termed 'singletons', were discarded as there is a lack of reliability in the sequence and taxonomy. Sequences which only occurred twice and came from the same study as indicated by sequential accession numbers ('doubletons') were also excluded, again due to lack of reliability from sequences deposited from the same study. More confidence is gained if a sequence has the same taxonomic assignments at least thrice, or from at least two different studies conducted by different research groups. Following this step, there were 121,063 such valid sequences in total.

These sequences in this set were then candidates for the training set if the sequence only comprised of A's, C's, G's, and T's which are the four nucleotides of DNA. This step was conducted to remove sequences with N's, which can arise from ambiguous base calls or errors during the sequencing process. In terms of how to design the taxonomic composition of the training set, it was deliberated whether to have equal proportions of each phyla or maintain the proportions from the original dataset. To cater for the majority of fungal biologists who work with Ascomycota or Basidiomycota fungi, the latter option was chosen to maximise the taxonomic representation of species in the training set. Since having

all of the approximate 200,000 sequences in the training set would be computationally expensive and impractical, a maximum limit of 3 sequences per species was set. If 3 separate sequences were not available for a given species, then 2 separate sequences with different accession numbers were chosen, since singletons and doubletons were removed. As a further check of the validity of the doublet or triplet of sequence, alignment of these 2 or 3 sequences for each species was performed using ClustalW (Larkin et al., 2007). At the species level, the ITS region should have almost identical sequences, so a minimum alignment identity threshold of 98% was enforced, where 'identity' refers to all sequences in a given position in the alignment having the same nucleotide. The alignments of 6,297 species passed this minimum identity threshold, and the 17,504 sequences represented by these species formed Training Set version 1.

3.3 Training Set Creation and Evaluation

The RDP Classifier developed by Wang et al. requires the training sequence file as a FASTA file which is required to be of a specific format, where <ident> contains the full taxonomy from domain to species separated by ';':

```
'>' <ident>
<sequence>
```

The sequence file needs to be accompanied by a taxonomy file that contains each taxon name, its rank and its parent taxon name, which is used to build the hierarchical tree for making classifications. Unlike bacterial 16S sequences for which there is a standard taxonomy, there is currently none available for fungal ITS sequences for reasons outlined in Section 2.2. This meant that we had to create our own taxonomy from the taxonomy given by the labels of the sequences in the training set.

Like Wang et al. and Liu et al., our training set was evaluated by conducting Leave-One-Out-Cross-Validation. This validation method randomly removes one sequence from the training set, and attempts to classify it based on a classifier built from the remaining sequences. The process is then repeated until each sequence in the training set has been classified in this manner. The LOOCV is another reason why singletons were removed in the sequence curation process described above; when the singleton

is removed for classification, it no longer occurs in the training set and the probability of correctly classifying that sequence becomes zero.

The 2,842 sequences that were incorrectly classified from the LOOCV of Training Set version 1 were inspected and resolved. Sequences that were misclassified at family rank and above indicated unreliability in the sequences and were thus discarded. BLAST searches of a sample of sequences misclassified at the genus or species levels showed in the majority of cases that the classifier was in fact correct. Rather than discarding these sequences, the taxon names between the actual and predicted taxa were merged.

The species that were omitted from Training Set version 1 due to their alignment identities being lower than 98% were realigned using a different combination of 2 or 3 sequences and included in the training set if the new doublet or triplet of sequences achieved an alignment identity $\geq 98\%$. This process was repeated until an alignment with identity greater than the threshold was obtained, or a maximum of 50 iterations was reached. This was performed to recover as many species as possible that were missed in Training Set version 1. As a result, an additional 2,776 species were recovered, leading to a total of 9,073 species spanning 24,447 sequences which formed Training Set version 2.

Training Set version 2 was also evaluated using LOOCV, and produced results deemed sufficiently accurate in the scope of the present study. The classifier was then subsequently built, resulting in the generation of 4 files that correspond to the hierarchical taxonomy (in XML format) and the prior and conditional probabilities for each species as described in Section 2.5.2.

3.4 Validation Set Creation and Classifier Evaluation

To most effectively test the performance of the classifier, a validation set was constructed by Dr Nai Tran-Dinh and Dr David Midgley which targeted families that were most frequently misclassified by the LOOCV of Training Set version 2. The validation set was created using 1,757 GenBank sequences (`www.ncbi.nlm.nih.gov/genbank/`) where the taxonomic assignment was reliable and trusted as judged by the research group that performed the sequencing and deposited the sequence.

For consistency, the 1,757 sequences were passed through FungalITSExtractor (Nilsson et al., 2010), similar to the processing of the original downloaded set of sequences (Section 3.2). The 1400 sequences

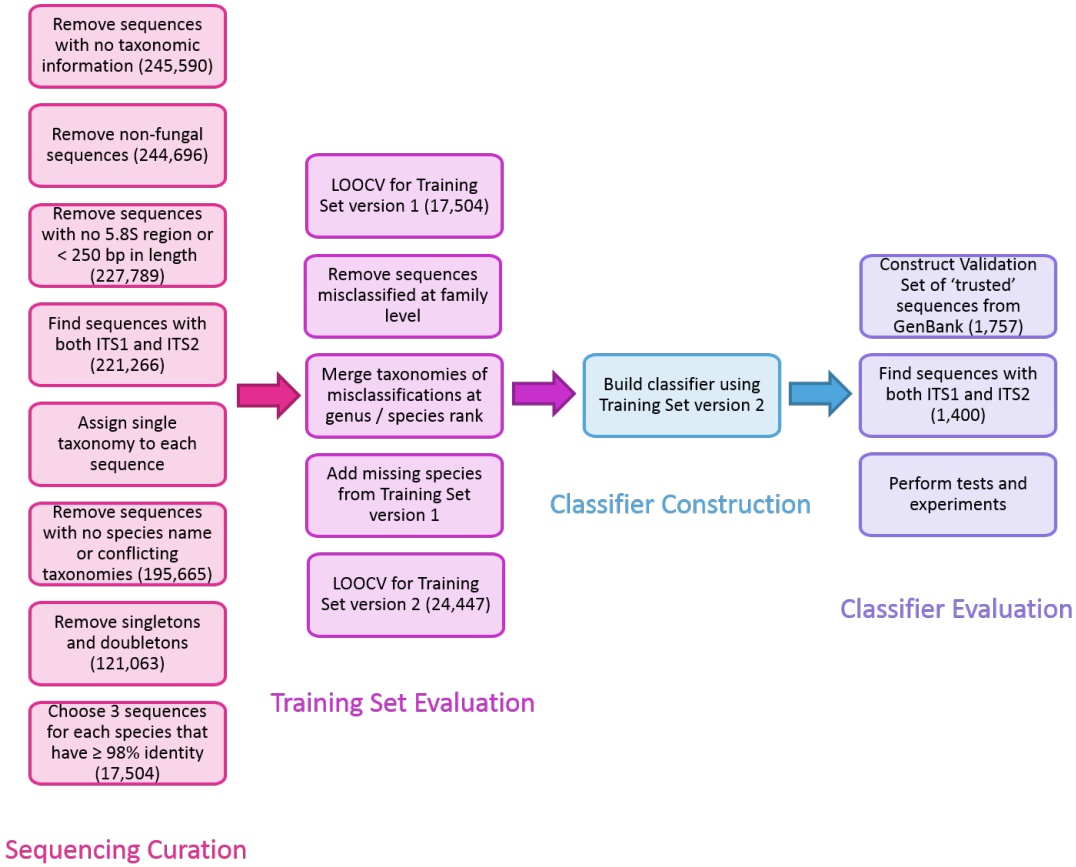


FIGURE 3.3. Workflow for the creation and evaluation of the ITS classifier.

that contained the entire ITS region, that is, ITS1–5.8S–ITS2, were retained to form the final validation set. Various experiments were conducted using this validation set, including accuracy measurements, simulation of amplicon sequencing reads and investigations of confidence values which are detailed in the following chapter.

The entire workflow from the ITS sequence curation to the building of the classifier through to its evaluation is summarised in Figure 3.3.

CHAPTER 4

Results and Discussion

4.1 Overview

The performance and usefulness of any classifier depends on the properties of the sequences that comprise the training set, including quality, length and the amount of taxonomic representation and coverage, as well as the accuracy of classification on new, query sequences that are submitted (Porter and Golding, 2012). The quality of the classifier itself can only be as good as the underlying training set used to build the classifier and formulate taxonomic assignments.

Chapter 3 described the extensive processing and curation procedures utilised to achieve the goal of a high-quality training set. The results of various tests to evaluate the performance of the training set, and the validation set used to evaluate the performance of the classifier are presented here. This chapter concludes with a discussion of the limitations of the classifier and suggestions for improvements.

4.2 Characteristics of the Training Set

The characterisation and assessment of the quality of the ITS sequences in the training sets constructed in the current study was guided by the metrics outlined by Porter and Golding (2012). These metrics also facilitated the comparison of the training sets. Note that the training set resulting from the first iteration is referred to as 'Training Set v1', while the second version of the training set is referred to as 'Training Set v2'.

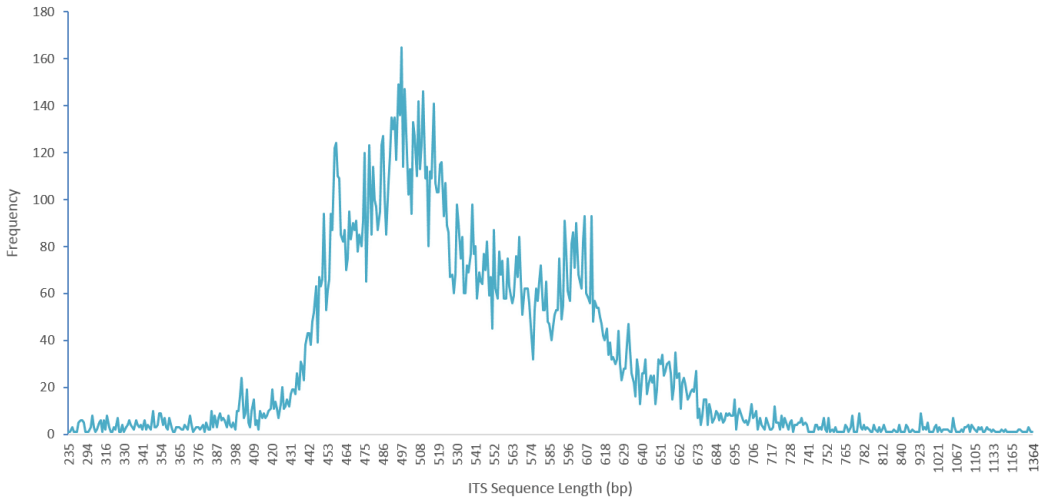


FIGURE 4.1. Length Distribution Histogram for the ITS Sequences in Training Set v1.

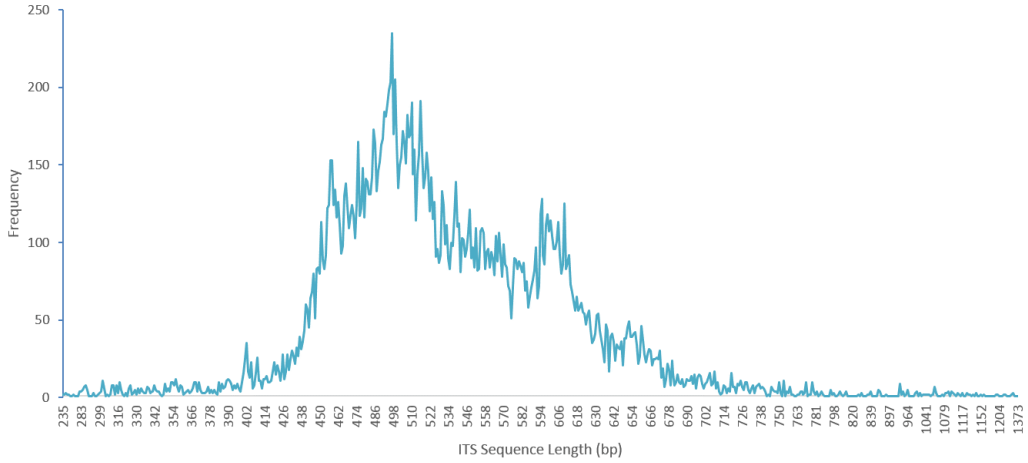


FIGURE 4.2. Length Distribution Histogram for the ITS Sequences in Training Set v2.

4.2.1 ITS Sequence Lengths

As mentioned in Section 3.2, the length of the full ITS region, consisting of ITS1-5.8S-ITS2, is highly variable and typically ranges from 400 to 1500 bp (D Midgley and N Tran-Dinh, personal communication). Figures 4.1 and 4.2 display the frequency distribution histograms of the lengths of ITS sequences in Training Set v1 and Training Set v2 respectively.

It is evident from Figures 4.1 and 4.2 that in both training sets, the bulk of the ITS sequences are in the range of 400 to 700 bp. Sequences in this range comprised 95.13% of all sequences in Training Set v1, and 94.93% of all sequences in Training Set v2. The sequence length most frequently observed was 497 bp in both training sets, forming the peak frequency of 165 in Training Set v1 and 235 in Training Set v2. The minimum sequence length of 235 bp is identical for both training sets, while the maximum sequence lengths differed only slightly, being 1364 bp in Training Set v1 and 1373 bp in Training Set v2. Furthermore, the average sequence lengths are 618 bp and 627 bp respectively in Training Set v1 and Training Set v2.

While the maximum lengths fall within the expected maximum length of 1500 bp, there are approximately 2% of sequences in both training sets whose lengths are shorter than the expected minimum length of 400 bp, and form the left tail of the histogram. This can be attributed to partial ITS1 and/or ITS2 regions, however as these short sequences contains both regions, they are considered 'valid' as defined in Section 3.2 and hence acceptable as members of the training set.

The quantitative metrics discussed here indicate that the length frequency distributions of Training Set v1 and Training Set v2 are similar, and which are consistent with the range of sequence lengths expected. In particular, the maximum lengths of the sequences in Training Set v1 (1364 bp) and Training Set v2 (1373 bp) are less than half of the maximum sequence length of 2995 bp observed in the original set of sequences downloaded from the UNITE database. This provides confidence that the sequences indeed represent the ITS region only and not its adjoining SSU or LSU regions (Figure 2.1), thereby validating the method used to extract the ITS region from the original downloaded set of sequences (Section 3.2).

4.2.2 Taxonomic Composition

Other critical factors that impact the quality of a classifier are the taxonomic breadth and depth encompassed by sequences in the training set upon which the classifier is built. This means that sequences in the training set need to be highly taxonomically representative such that sufficient interspecies and intraspecies variation is captured, in order to provide the classifier with power for resolving sequences down to the species level.

TABLE 4.1. Comparison of the taxonomic composition of Training Set v1 and Training Set v2 at the phylum rank.

Phylum	# Sequences (v1)	% of Total (v1)	# Sequences (v2)	% of Total (v2)
Ascomycota	10,715	61.21	14,615	59.78
Basidiomycota	6,499	37.13	9,195	37.61
Blastocladiomycota	2	0.01	2	0.01
Chytridiomycota	34	0.19	47	0.19
Fungi *Incertae sedis*	0	0.00	8	0.03
Glomeromycota	62	0.35	258	1.06
Neocallimastigomycota	0	0.00	5	0.02
Zygomycota	192	1.10	317	1.30
Total	17,504	100.00	24,447	100.00

Training Set v1 contains 17,504 sequences spanning 1,289 genera and 6,297 species, while its refined version Training Set v2 contains 24,447 sequences representing 1,601 genera with 9,073 species. The taxonomic composition of both sets of training sequences are characterised at the phylum level as summarised in 4.1.

The Ascomycota phylum was the most represented phylum in both training sets, accounting for about 60% of all sequences, followed by Basidiomycota which comprised about 37%. Together, these two phyla constituted 98.34% and 97.39% of Training Set v1 and Training Set v2 respectively. It was considered whether to design the training set to have equal proportions of all phyla. Most fungal research is directed towards studying organisms in the Ascomycota and Basidiomycota phyla, hence the majority of fungal biologists, or the end users of this tool, will be interested in identifying species originating from these two phyla. The current taxonomic composition was thus deemed more desirable and appropriate.

There is no representation of the Fungi *Incertae sedis* or Neocallimastigomycota in Training Set v1; this was rectified in Training set v2 which has 8 and 5 sequences respectively from these missing phyla. Although the *Incertae sedis* phyla is essentially a group of sequences 'of unknown (taxonomic) position', a small number of these sequences were included for testing purposes. Thus Training Set v2 has constituent sequences from all of the major fungal phyla, with the exception of Microsporidia, which was omitted during the ITS extraction phase as it did not appear to contain the full ITS region. The question of whether Microsporidia is truly a member of the fungal kingdom still remains debated, and sequences

TABLE 4.2. Comparison of the taxonomic composition of the ITS Training Set and LSU Training Set at the phylum rank.

Phylum	# Sequences (ITS)	% of Total (ITS)	# Sequences (LSU)	% of Total (LSU)
Ascomycota	14,615	59.78	2,395	28.16
Basidiomycota	9,195	37.61	6,024	70.82
Blastocladiomycota	2	0.01	8	0.09
Chytridiomycota	47	0.19	51	0.60
Fungi *Incertae sedis*	8	0.03	5	0.06
Glomeromycota	258	1.06	16	0.19
Neocallimastigomycota	5	0.02	2	0.02
Zygomycota	317	1.30	1	0.01
Eukaryota *Incertae sedis*	0	0.00	4	0.05
Total	24,447	100.00	8,506	100.00

from this phyla were also excluded by Schoch et al. (2012) in their studies that led to the proposal of ITS to be an official fungal barcode, as described in Section 2.3. This unresolved debate, in conjunction with the scepticism of ribosomal RNA genes, including ITS and LSU, as effective markers for Microsporidia species identification, warrants their exclusion from the training set.

To further assess the validity of the phylum distribution of Training Set v2, which was the final training set used to build the classifier, comparisons were made against the phylum distribution of the training set developed by Liu et al. for their RDP LSU classifier (Table 4.2).

The LSU training set contains 776 genera represented by 8,506 sequences. While this is almost one-third the number of sequences present in our Training Set v2, this lower taxonomic coverage is likely adequate for classifying down to genus ranks only.

In contrast to Training Set v1 and Training Set v2, the most represented phylum in the LSU training set is Basidiomycota, which is almost double the proportion of 70.82% compared to the ITS training sets. Sequences from Ascomycota, which is the most represented phylum in the ITS training sets, has half the proportion in the LSU training set at 28.16%. These two phyla however, still account for 98.98% of all sequences in the LSU training set, similar to the pattern of composition observed in both the ITS training sets (Table 1). The LSU training set also has 4 sequences from Eukaryota *Incertae sedis*, which

are of unknown taxonomic position in the wider Eukaryota domain and therefore not necessarily fungal LSU sequences.

Thus it is evident that Training Set v2 is widely representative with sequences originating from each of the major fungal phyla, and also has depth with the majority of sequences originating from the most relevant phyla. The training sequences used to build our new ITS classifier therefore fulfils the goal of wide taxonomic representation which is recognised as a feature paramount for the high accuracy of a classifier.

4.3 Training Set Performance

The performance of the training set needs to be thoroughly tested and evaluated prior to building the classifier. Several different validation techniques for evaluating classification tools have been developed, including 10-fold Cross Validation, Held Out Validation etc. The validation technique used to evaluate our ITS classifier was informed by Wang et al. (2007) and Liu et al. (2012), who utilised the Leave-One-Out-Cross-Validation (LOOCV) method to test their classifiers. Incorporating the same method in our testing pipeline also facilitated valid comparisons between the fungal classification tools.

LOOCV was firstly used to evaluate both our ITS training sets as a means of quantifying the differences in performance by comparing the accuracies achieved. The LOOCV process took 29 hours to complete on Training Set v1 (consisting of 17,504 sequences), and 50 hours to complete on Training Set v2 (consisting of 24,447 sequences) on a 2x Intel Xeon E5-2680 (2.7GHz) machine. Note that the LOOCV algorithm is implemented to run on a single core only; converting this to a multi-threaded program, whilst out of the scope for this study, may be a worthwhile endeavour in future.

As expected, the accuracy of classifications decrease as the specificity increases from domain down to species rank for both training sets. From domain to family ranks, both training sets have similar accuracies between 98% and 100%. Discernible differences only start to emerge at the genus and species levels, where Training Set v2 performs better than Training Set v1. There is a 1.6% gain in accuracy from 97.2% in Training Set v1 to 98.8% at the genus level, and a 5.8% increase in accuracy from 84.4% in Training Set v1 to 90.2% at the species level.

FIGURE 4.3. Comparison of LOOCV Accuracies between ITS Training Set v1 and ITS Training Set v2.

During the accuracy calculations, misclassifications caused by teleomorph-anamorph differences were resolved at the genus level, and species with different names that are synonyms of each other were also resolved to a certain extent by querying the predicted species and the labelled species against fungal nomenclature databases such as MycoBank (`http://www.mycobank.org`) and Index Fungorum (`http://www.indexfungorum.org`). Due to time constraints, it was not feasible to check all of the 2000-3000 species misclassifications manually. It is expected however that the majority of these misclassifications, or discrepancies between the assignment given by the ITS classifier and the UNITE/INSD annotations, are due to errors in these database annotations themselves, and should not be counted as incorrect. This claim is supported by observations that approximately 20% of fungal sequences deposited in GenBank, a member of INSD, are incorrectly assigned to the species rank (Balajee et al., 2009; Dannemiller et al., 2013). In light of this, the accuracies presented in Figure 4.3, and hereafter, should be regarded as minimum accuracies.

Note that this problem also presents a major challenge for the evaluation of the classifier. The reference taxonomies or annotations are treated to be the 'correct' or 'gold standard' label against which classifications are determined to be correct or incorrect. If there are errors in the reference taxonomies

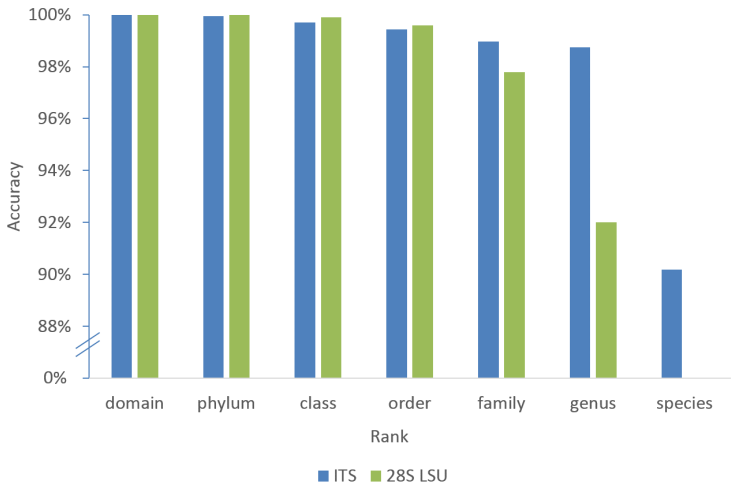


FIGURE 4.4. Comparison of LOOCV Accuracies between ITS Classifier and LSU Classifier. Intermediate ranks of subphylum and subclass have been omitted as the LSU classifier does not does not report accuracies at these ranks. Note that the values presented for the ITS classifier are identical to those presented for Training Set v2 in Figure 4.3; they have been added here again for ease of comparison.

themselves, accuracy measurements for a classifier will be misguided. Whilst the reference set of sequences that form the training set were carefully processed to remove sequences with taxonomic errors, the assumption that the reference taxonomy is indeed 'correct' is applied when performing accuracy calculations.

To ascertain how our new ITS classifier performs in comparison to existing fungal classification tools, the results of the LOOCV were compared with the LOOCV results from the fungal LSU training set as published by Liu et al. (2012). Note that the accuracy percentages published in their paper do not agree with the numbers reported in the supplementary information, which we had difficulty reconciling. The comparisons have therefore been performed using the reported accuracies in Figure 4B of Liu et al.'s paper.

Figure 4.4 illustrates that for higher taxonomic ranks from domain to order, our ITS classifier achieves accuracies comparable to the LSU classifier. For lower taxonomic ranks however, the ITS classifier produces superior results, with a 1.2% gain in accuracy observed at the family level from 97.8% in the LSU classifier to 99.0% in the ITS classifier, and an increase of 6.8% from 92.0% in the LSU classifier to 98.8% in the ITS classifier. The use of the LSU gene as the basis of classification by Liu

et al. limited the resolving power of the classifier to genus rank. We show here that a Naïve Bayes classifier built using a training set comprised of high-quality ITS sequences can achieve an accuracy of 90.2% on the training set at the **species** level. Note that while our ITS training set comprised of mainly full ITS sequences of length up to 1500 bp (Figure 4.2), the training sets tested in the LSU classifier ranged from 75 to 400 bp sequences to simulate the lengths of sequence reads produced by the common sequencing technologies. The results of the LSU classifier using 400 bp sequences, which gave the highest accuracies, are presented in Figure 4.4). The superior accuracies attained by the ITS classifier is partly due to the longer lengths of the ITS sequences in the training set. For a fairer comparison with the LSU classifier, we also tested the classifier using 400 bp sequences from a validation set (Section 4.4), which had similar accuracies to the full length sequences in the training set, and thus still performs better than the LSU classifier. The performance of the validation set is discussed in Section 4.5.

The results of the LOOCV performance of the training sets presented in this section provide a clear example of how diverse taxonomic composition and greater taxonomic representation of a training set greatly benefit a classifier. ITS Training set v2 outperforms its predecessor ITS Training Set v1 and the LSU training set (Liu et al., 2012) in terms of accuracies from the LOOCV evaluation method (Figures 4.3 and 4.4). This is attributed to the greater number of sequences in ITS Training Set v2, which contains 1.5 times the number of sequences in ITS Training Set v1 (Table 4.1), and almost 3 times the number of sequences in the LSU training set (Table 4.2). This observation reaffirms the claim made by Porter and Golding (2012) that as the number of sequences in the training set increase, so should the number of correct classifications made by the classifier.

4.4 Characteristics of the Validation Set

The high accuracy of the ITS reference training set constructed in this study was demonstrated in the previous section, with our classifier based on these sequences achieving an accuracy of 90.2% at the species level. To enable deeper analysis of the classifier, a validation set was created as described in Section 3.4 that contained 1,400 GenBank sequences from families which the classifier performed poorly on during the LOOCV testing of the training set. As with the training sets, the quality of sequences in the validation set was firstly investigated using the same metrics of sequence length and taxonomic composition.

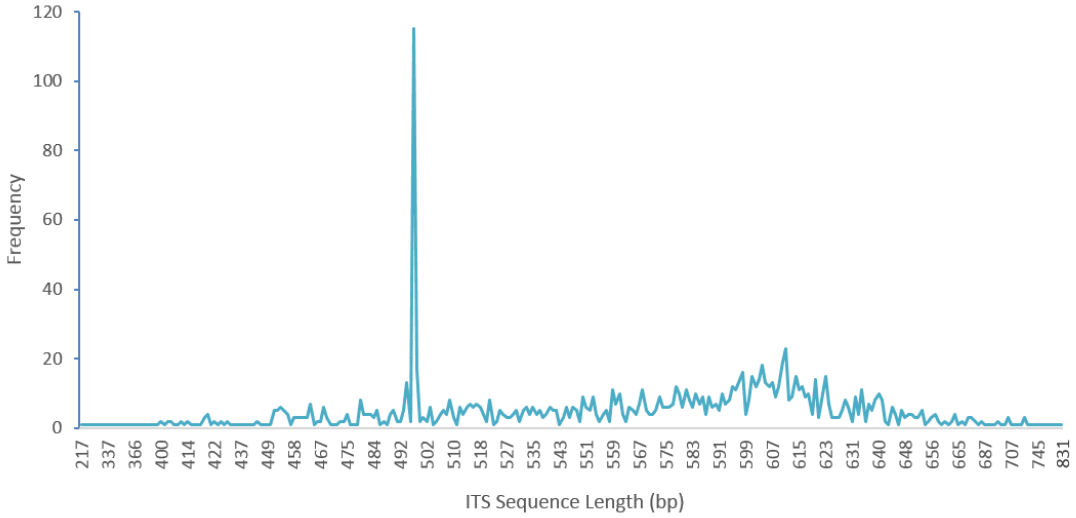


FIGURE 4.5. Length Distribution Histogram for the Validation Set.

4.4.1 ITS Sequence Lengths

The length frequency distribution histogram (Figure 4.5) for the validation set displays a similar pattern to that observed for both ITS training sets in Figure 4.1 and Figure 4.2 respectively. The lengths of the ITS sequences in the validation set range from 217 bp to 831 bp, with an average length of 542 bp. 96.79% of sequences were of a length between 400 and 700 bp, with a peak length of 492 bp. These values are coherent with those from the length distributions of both ITS training sets with the exception of the maximum length which is 500 bp shorter in comparison. The validation set was constructed to specifically target certain families, and was not required to be a representative sample like the training set, which accounts for the difference in maximum lengths. The larger proportion of shorter sequences in the validation set suggests the classifier performed worse on these shorter sequences during the training set evaluation. This is to be expected as shorter sequences are composed of fewer k-mers and hence contain lesser information to base the classifications on.

4.4.2 Taxonomic Composition

The phylum distribution of the validation set reflects the sequences that were specifically targeted as the classifier had difficulty classifying these. In contrast to the ITS training sets (Table 4.1), the majority of

TABLE 4.3. Taxonomic composition, at the phylum rank, of the validation set used to test the ITS classifier.

Phylum	# Sequences	% of Total
Ascomycota	322	23.00
Basidiomycota	1,007	71.93
Blastocladiomycota	13	0.93
Chytridiomycota	43	3.07
Fungi *Incertae sedis*	0	0.00
Glomeromycota	1	0.07
Neocallimastigomycota	0	0.00
Zygomycota	14	1.00
Total	1,400	100.00

the validation set is comprised of Basidiomycota sequences (71.9%), followed by sequences from the Ascomycota phylum (23.0%). Like the training sets, these two phyla form 94.93% of the total sequences. The validation set does not contain any sequences from Fungi *Incertae sedis* or Neocallimastigomycota, as sequences in these phyla seemed to be accurately classified.

4.5 Validation Set Performance

Figure 4.6 shows the performance of the ITS classifier on the validation set as measured in terms of the number of correct classifications, where 'correct' is determined with respect to the labelled taxonomy of the sequence. In the training set assessment, the nature of all misclassifications were characterised to inform what changes needed to be made for the next iteration of the training set, so accuracy measures were calculated using the entire set of training sequences. In contrast, the accuracy of the validation set was calculated using only the sequences for which the confidence value at the species level was 90% or greater, to gain a true understanding of its performance. Of the 1,400 sequences classified, 834 sequences satisfied this criteria and were used in the accuracy calculations.

As shown in Figure 4.6, from domain down to class, the accuracies are above 99%. The accuracy at the order level is 97.96% while accuracies at family and genus were 98.08% and 98.20% respectively. At the species level, a reasonable accuracy of 63.19% was attained. It should be noted that the sequences in the training set were annotated with UNITE taxonomies, while the sequences in the validation set had

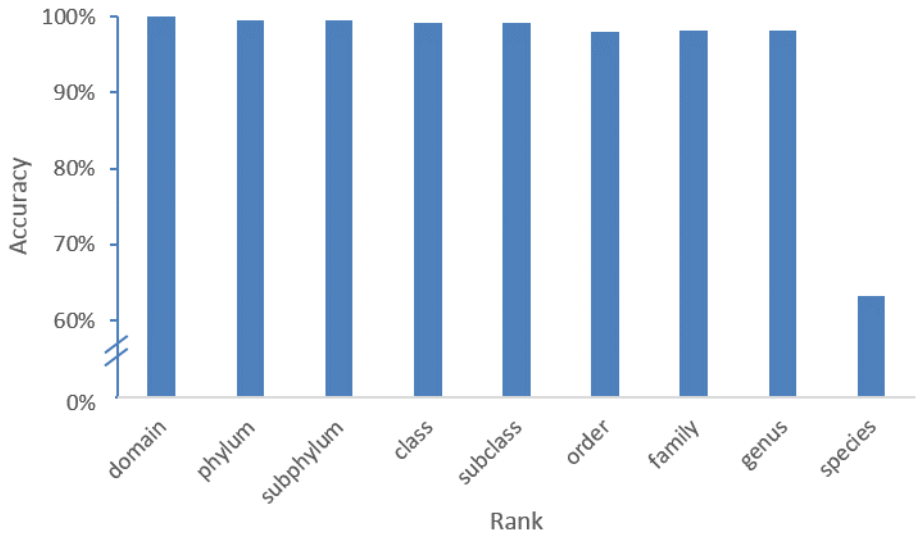


FIGURE 4.6. Accuracies of ITS Classifier on the Validation Set.

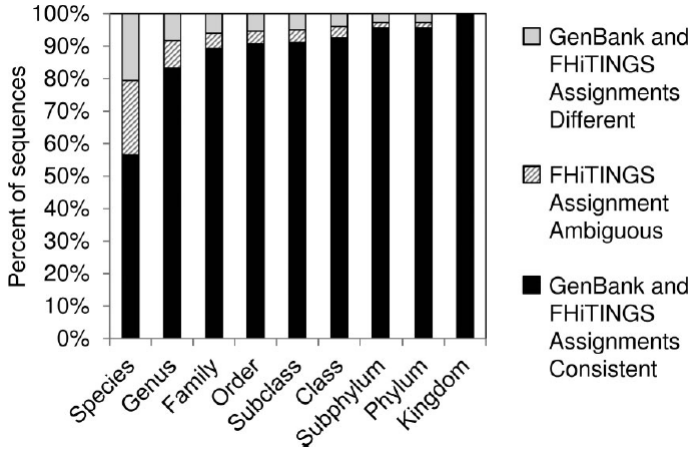


FIGURE 4.7. Test results of the FHiTINGS classifier in comparison to GenBank annotations for ITS sequences (Dannemiller et al., 2013). Note that the taxonomic ranks along the x-axis are in reverse order to those in Figure 4.6.

GenBank taxonomies. This difference between the taxonomies meant that the misclassifications needed to be manually inspected to verify whether they were true 'errors', similar to the process in the training set evaluation (Section 4.3).

These results are comparable or better to the performance of FHiTINGS (Dannemiller et al., 2013) (Figure 4.7), which was tested via *in silico* analysis of 515 ITS sequences, obtained from GenBank (`http://www.ncbi.nlm.nih.gov/genbank/`), with species level identifications. An important distinction to be made is that FHiTINGS was evaluated using 515 sequences, whereas the ITS classifier was tested using 1400 sequences. FHiTINGS was used to classify each of the 515 sequences, and the classification was compared with the original annotations from GenBank. Classifications were considered 'consistent' if the FHiTINGS classification matched the GenBank classification. If after applying the LCA algorithm, the classification still does not match, it is marked as 'ambiguous'. And finally, inconsistent classifications are marked as 'different'. FHiTINGS achieves accuracies of 57% at the species level and 83% at the genus level. If ambiguous classifications are not treated as incorrect, the accuracies then increase to 79% and 91% for species and genus ranks respectively. For a true comparison between our ITS classifier and FHiTINGS, both should be re-evaluated using the same validation set.

In our ITS classifier, the time taken for the classification of all 1,400 sequences was approximately 1 minute, reiterating the advantage of composition based classifiers over similarity-based classifiers and phylogeny-based classifiers as stated in Section 2.4.3. The only bottleneck therefore is the LOOCV method; once the training set is finalised and the classifier built, the latter only taking a couple of minutes for the 24,447 sequences in Training Set v2, the actual time for classifications is much faster. The speed of classifications is an important consideration as large microbial studies, in particular metagenomic studies that aim to characterise all species present in an environmental sample, will often submit a large number of sequences, in the order of thousands, to be classified simultaneously.

4.6 Amplicon Sequencing Simulation Performance

Fungal biologists will commonly sequence the ITS region by firstly amplifying the ITS genome sequences using the Polymerase Chain Reaction (PCR) (Figure 4.8). This results in a set of DNA fragments in the forward direction, and a set of DNA fragments in the reverse direction. These sets of amplified DNA fragments are then passed to a sequencing technology that returns the actual sequence of DNA in terms of its ordered, constituent nucleotides. The Sanger sequencing technology returns sequences as 'reads' of length up to 1100 bp, so often the entire ITS region can be captured in a single read,

FIGURE 4.8. PCR Sequencing of the ITS region. Two primers (short sequences of DNA) are used; (red) one that binds a conserved DNA site upstream of the ITS1 (in the 18S SSU gene) and another (orange) that binds a conserved DNA site downstream of the ITS2 (in the 28S LSU gene). The primers are then extended via the addition of nucleotides in the direction showed by the dashed arrow, to produce two DNA fragments that are replicas of the original sequence. This process is repeated many times to produce several copies of the DNA sequence. Image not to scale.

in either the forward or reverse direction. Nowadays, 'pyrotagging', or 'amplicon sequencing' which is a type of Next-Generation Sequencing technology, is more commonly used which produces short reads up to 400 bp in length, so the reads correspond to either the first 400 bp of the ITS1 region or the last 400 bp of the ITS2 region depending on which direction is sequenced.

An experiment was designed to simulate the sequencing reads produced by amplicon sequencing, to investigate how our classifier performs on these shorter reads. For each of the 1,400 sequences in the validation set, the first 400 bp and the last 400 bp were extracted and passed through the ITS classifier. Sequences of length shorter than 400 bp were excluded from this analysis. For consistency, only sequences where the classification confidence was 90% or greater were included in the accuracy measurements, which led to a total 772 sequences representing the first 400 bp reads and 822 sequences representing the last 400 bp reads. Misclassifications were resolved in a similar manner as described in Section 4.5 above.

The results displayed in Figure 4.9 show that for all ranks, the accuracy gained from the first 400 bp reads and last 400 bp reads are highly similar to the accuracy attained from the full ITS sequence. In fact, from domain to genus, the difference in accuracies between all three types of tests are too small to be distinguishable, with only a 0.02% difference in genus level accuracies of 98.20%, 98.19% and 98.18% between the full ITS sequence, the first 400 bp and last 400 bp respectively. This difference is more pronounced at the species level, with accuracies of 63.19% for the full ITS, 64.77% for the first 400 bp and 60.10% for the last 400 bp. These differences however, are not large enough to be significant. The presence of the conserved 5.8S region between the ITS1 and ITS2 may also explain

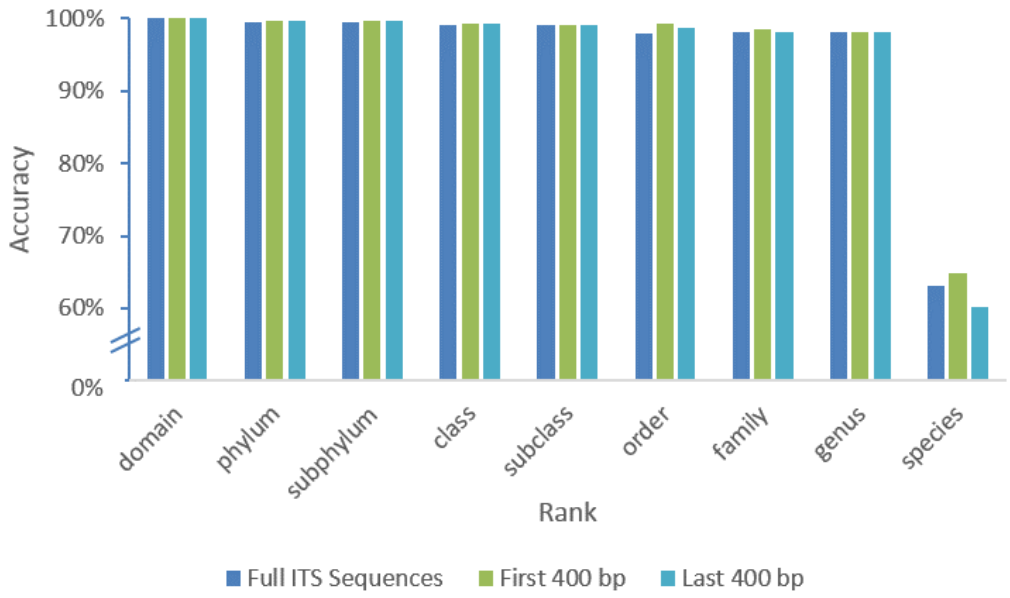


FIGURE 4.9. Comparison of validation set accuracies of amplicon sequence reads (400 bp) against full ITS sequences. Note that the values for the full ITS sequences are identical to those in Figure 4.6; they have been included here again for ease of comparison.

these observations, which suggests that exploring a training set containing ITS sequences with the 5.8S region excluded may be a worthwhile endeavour.

Therefore the results clearly indicate that the entire ITS region need not be sequenced; the ITS1 and ITS2 sequences contain sufficient variation such that short 400 bp sequences produced by amplicon sequencing of either of these regions can be classified with high accuracies comparable to classifying the full ITS sequence.

Two other classification tools discussed in Section 2.4.1, MEGAN (Huson et al., 2007) and FHiTINGS (Dannemiller et al., 2013), require the user to perform a BLAST search of their sequence reads, obtain the output and convert it to a specified format from which these tools will then perform the classification. In contrast, Wang et al. developed the Naïve Bayes classifier so that raw sequence reads can directly be passed to the classifier, eliminating an extra step and reducing time the total time for analysis. This highlights another strength of this classifier in its ability to efficiently cater for short reads produced by various sequencing technologies.

4.7 Effects of Sequence Composition on Confidence Values

A highly useful feature of the classifier is the calculation of confidence values for each assignment made. To complement the analysis of the effects of sequence composition on classifier accuracies (Section 4.6), we now investigate the effects of sequence composition on the confidence values. Here sequence composition refers to which genes or regions of the rRNA locus (e.g. 18S SSU, ITS1, 5.8S, ITS2, 28S LSU) are present in the sequences.

For each of the sequences in the validation set, four sets of sequences were generated: (i) the raw sequences downloaded from GenBank (Raw), (ii) the full ITS region (ITS1-ITS2), (iii) the ITS1 region only (ITS1) and (iv) the ITS2 region only (ITS2). Each type of sequence was classified using the classifier, and the confidence values compared.

The graphs in Figure 4.10 are representative of the three types of sequence compositions observed in the validation set. As expected, a general trend observed is the decrease in confidence values from higher taxonomic ranks to lower, more specific taxonomic ranks. Graphs a) to d) were produced from raw sequences containing the full ITS region and a part of the adjoining 18S SSU and 28S LSU genes. Graph a) shows that all four types of sequences have the same confidence value of 1.0 until the class rank, after which the ITS1-only sequence drops in confidence value. At family and genus ranks, the ITS1-ITS2 sequence has the highest confidence, followed by ITS2, ITS1 and the raw sequence. At the species rank however, all confidence values seem to converge for this particular sequence. In graph b), all four types of sequences yield the highest confidence value of 1.0 until the order rank, after which the raw sequence shows a steep descent down to 0.4. ITS1-only and ITS2-only decrease slightly to 0.8, while the full ITS1-ITS2 maintains a confidence value of 1.0 until the genus rank. At the species rank, ITS1 has the highest confidence value of 0.7, followed by ITS1-ITS2, ITS2 and finally the raw sequence with a confidence of 0.4. Graph c) displays a similar pattern to graph b), with a difference in confidence values only occurring after the order rank. The full ITS1-ITS2 sequence had the highest confidence values of 0.9 from family to species, while ITS1 had the lowest values in this range, with a final confidence value of 0.24 at the species level. The raw sequence and ITS2 had confidence values of 0.74 and 0.68 at the genus rank respectively, and 0.44 and 0.68 respectively at the species rank. In

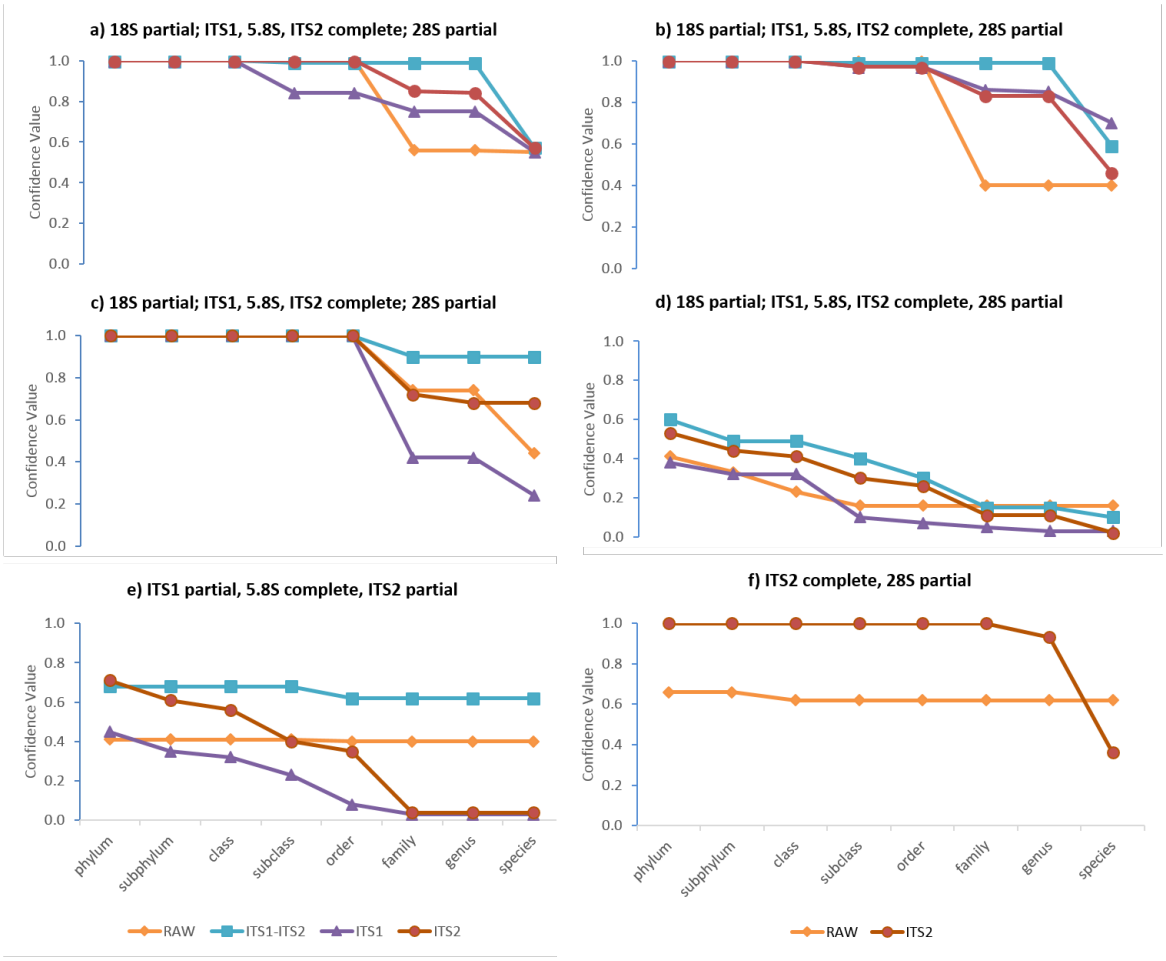


FIGURE 4.10. Confidence values for different types of raw sequences, as indicated by the titles of each graph. Note that confidence values at the domain rank are not shown as they all have a value of 1.0.

graph d), the ITS1-ITS2 sequence generally has higher confidence values across all rank, however at the species rank, the raw sequence has a slightly higher confidence value.

Graph e) displays the trend in confidence values for a sequence containing the partial ITS1, complete 5.8S and partial ITS2, with no report of the SSU or LSU genes. Therefore it is expected that the results of the raw and ITS1-ITS2 sequences should coincide. The only similarities between the confidence values of these two sequences is the uniformity across all ranks, with confidence values almost constant around 0.4 and 0.65 for raw and ITS1-ITS2 sequences respectively. The lower confidence values of the raw sequence may be caused by the presence of partial SSU and/or LSU sequence and incorrect labelling of

the sequence. For this sequence, the confidence values of ITS2 were greater than ITS1 until order rank, after which the two sets of values converge.

The sequence represented by Graph f) only contained the complete ITS2 region and part of the LSU gene. The confidence values of the raw sequence are consistent around 0.6 for all ranks. For ITS2, the confidence values are 1.0 from domain to family, and dip to 0.93 at the genus level. At the species level however, there is a drastic decrease to 0.36, which falls below the value of the raw sequence.

Piecing these patterns together, it can be inferred that classifications of the ITS region are allied with higher confidence values, which is consistent with the results obtained from the amplicon sequencing performance (Section 4.6). In cases where the confidence values of the raw sequences were better, this is most likely caused by k-mers from LSU sequences rather than from the highly conserved 18S SSU gene, which alludes to the potential of using the LSU gene as a secondary marker in conjunction with the ITS region (Schoch et al., 2012).

4.8 Future Work

The results and analysis presented for the fungal ITS Classifier suggest opportunities for further work to overcome the limitations of the current classifier and techniques for further investigations and improvements.

At least one more iteration of the training set is required, in which misclassifications in the current Training Set v2 should be rectified and more sequences added from the Blastocladiomycota and Neocallimastigomycota phyla as these are still under-represented in current training set (Table 4.1).

The LOOCV of the training set (Section 4.3) could be tested with different values for the parameters of k-mer length (word size), the minimum bootstrap value, etc. A longer k-mer length can provide better specificity, but results in a larger feature space which adds computational complexity. Therefore a tradeoff value of k needs to be found; for the purposes of this study, the k-mer length was set to 8 which is the default value, which still produced accurate results. This could not be done in the time allocated as the LOOCV took a couple of days for each iteration, due to the underlying code which is designed to run on a single core. This process therefore could have been made much faster by modifying the code

to be multi-threaded such that different iterations of the LOOCV can be assigned to different threads, since each iteration is independent of the others. A different approach would be to utilise different validation methods, for example 10-fold Cross Validation, and compare the results with the LOOCV. For the purposes of this study, the accuracy values obtained from the LOOCV were deemed reasonable and sufficiently accurate.

For the validation set performance, amplicon reads of size 400 bp were simulated and used to further examine the classifier's performance (Section 4.6). This experiment can be extended to cover a range of metagenomic reads of different sizes, as conducted by Wang et al. in their 16S classifier and Liu et al. in their LSU classifier. This range should extend down to 100 bp (to cater for Illumina sequencing reads) or even 50 bp which is the minimum sequence length required by the classifier. This would have provided another effective test for evaluating the classifier, and relevant to the types of sequences potential end users would be submitting as queries.

The analysis of sequence composition and its effect on confidence values was investigated using a representative sequence for each of the composition types. Averaging the confidence values of all the sequence of each composition type would aid this analysis.

The development of a new classifier that uses a combination of genes, rather than a single gene such as the ITS region used in the current classifier and the LSU gene used in Liu et al.'s classifier, would also be a worthwhile investigation. The LSU gene has proven to be accurate for classifying to genus rank, and has the potential to act as a secondary marker to supplement species level classifications that an ITS-only classifier performs poorly in. This in particular applies to the groups of species outlined in Section 2.3 in which the ITS sequences are almost identical, therefore lacking sufficient variability.

Finally, a completely different type of machine learning technique, such as clustering methods, could be developed. As justified in Section 2.4.4, the RDP Naïve Bayes Classifier was chosen due to its extensive testing and proven success among biologists, who are the target end users of this tool.

4.9 Conclusions

We have developed a high quality training set of ITS sequences and used this as the basis of a new Naïve Bayes classifier. Using a series of experiments and tests, we have demonstrated the ability of the classifier to rapidly and accurately classify down to the species level.

From the LOOCV evaluation of the latest training set (Section 4.3), the accuracy at genus level is 98.8% for the ITS classifier, which is about a 6% improvement over the LSU classifier (Liu et al, 2011), and an accuracy of 90.2% at the species level.

On the validation set, the classifier also achieves accuracies of 98% at the genus level, and 64% at the species level for full length ITS sequences. The results of the amplicon sequencing reads (Section 4.6) are encouraging as similar accuracies across all ranks were attained by using these shorter sequences, eliminating the need for biologists to sequence the entire ITS region (using Sanger sequencing for example) as a 400 bp sequence provides comparable accuracy. This highlights a desirable feature of the classifier in its amenability to be used directly for classifying short sequence reads obtained from the latest sequencing technologies, which are commonly used by biologists today, without requiring further processing of the raw reads. As a consequence, the time taken between sequencing an unknown fungal isolate and classifying it to determine its species identity is greatly reduced.

Part 2

Evolutionary Dynamics of Aflatoxin Genes

CHAPTER 5

Introduction to the Evolutionary Dynamics of Aflatoxin Genes

5.1 Overview

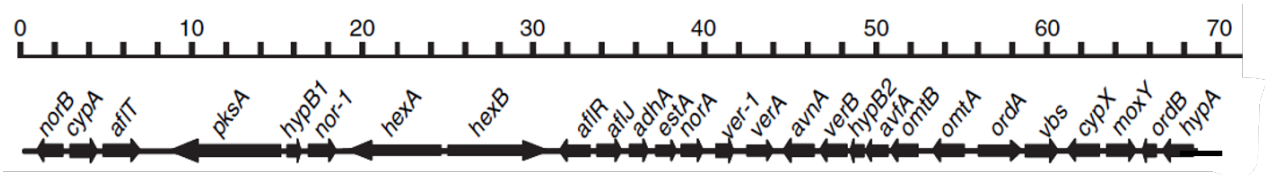
The construction of the classifier which entailed an investigation of fungal taxonomy and relationships as described in Part 1, provides a platform for focusing on a particular group of organisms in the *Aspergillus* genus that were highlighted in Chapter 1 for their aflatoxin producing capabilities, and thus which have agricultural and medical importance. This chapter provides a thorough explanation of the metabolic pathway that contains the aflatoxin genes (Section 5.2). This is accompanied by a review of literature pertaining to evolutionary relationships between these special class of organisms (Section 5.3).

5.2 Aflatoxin Biosynthesis

The *Aspergillus* genus, containing approximately 250 known fungal species, is of great importance to humans due to the wide ranging applications and health impacts of the species. *Aspergillus oryzae* and *Aspergillus sojae* have industrial uses in food fermentation processes for the production of sake and soy sauce (Chang et al., 2007). In contrast, *Aspergillus flavus*, *Aspergillus minisclerotigenes*, *Aspergillus parasiticus* and *Aspergillus nomius* are natural producers of the carcinogenic aflatoxins, which occur widely in commodities used for human and animal food consumption and pose a significant threat of contamination in the food chain.

Accurate and reliable methods of distinguishing safe from toxic *Aspergillus* species are yet to be realised, even though the mechanism of aflatoxin production through biosynthetic pathways has been well studied and characterised, including the genes, intermediate products and enzymes, and regulatory mechanisms (Yu, 2012). The first physical map was constructed in 1995 of all the aflatoxin pathway

FIGURE 5.1. Structure of the aflatoxin biosynthesis pathway in *Aspergillus flavus* and *Aspergillus parasiticus*. Arrows indicate the length and direction of genes. The genes *hypB1* and *hypB2* are hypothetical genes. The scale bar represents the length in thousands of base pairs (bp). Image adapted from Figure 1(a) in Ehrlich et al. (2005).



genes known at the time, using genetic studies based on restriction enzyme analysis (Yu et al., 1995). This method involved comparing the overlapping regions genomic clones of aflatoxin pathway genes from *A parasiticu*s and *A flavus*. These initial results found these genes to be clustered within a 60 kb DNA region and having the same gene order in both *A parasiticus* and *A flavus*. It is now known that the pathway consists of 25 genes (Figure 5.1) that are all clustered within an 80,000 bp region of the genome close to the end of chromosome III, known as the telomeric region (Yu, 2012). The genes are ordered according to the enzymatic steps required for aflatoxin biosynthesis (Yu et al., 1995), and the protein products of these genes are highly diverse in terms of their functions, including oxygenases, reductases, dehydrogenases, methyltransferases and regulatory enzymes AFLR and AFLJ that control the activation and expression of the pathway (Yu et al., 2004).

The previous study described was based on genomic sequences and although identification of aflatoxin pathway genes was achieved, insights into expression levels and regulation of genes could not be gained. A subsequent study to verify these results involved the large scale sequencing of Expressed Sequence Tags (ESTs) from *A flavus* (Yu et al., 2004). ESTs represent gene sequences, that is, regions of the genome that are expressed. Unique ESTs were searched against a non-redundant protein database to assign putative functions based on the gene ontology (GO) functional classification. Indeed, it was found that all except four of the 25 aflatoxin biosynthetic genes identified by Yu et al. in 1995 had corresponding ESTs in *A flavus* and also *A parasiticus*, verifying the expression of these genes. The gene sequences were on average 95% homologous between the two species.

Strains of *A oryzae* are also known to contain the aflatoxin biosynthetic gene cluster homologous to *A flavus* in their genome (Payne et al., 2006; Rokas et al., 2007). To elucidate the expression patterns of these aflatoxin genes, large scale EST analysis in A oryzae was performed by Akao et al. in 2007. A

similar approach was adopted as the one used by Yu et al. (2004), by searching ESTs against a non-redundant protein database using BLASTx (`http://blast.ncbi.nlm.nih.gov`) and assigned to functional GO categories based on matches that were lower than the statistical significance threshold of 10^{-9}. Sequences corresponding to the flatoxin biosynthetic gene cluster in the reference genome *A oryzae* strain RIB 40 were compared to all the ESTs, and none of the sequences were found to occur in the ESTs except for the genes *aflJ* and *aflT*. The former is involved in regulation of the aflatoxin pathway while the latter encodes a transporter protein. In contrast to the findings made by Yu et al. (2004) for *A flavus*, 23 out of the 25 genes showed no expression in *A oryzae*.

Gibbons et al. (2012) used a different strategy to study *A oryzae* gene expression using units of reads per kilobase per million mapped reads (RPKM) was used as a quantitative measure of gene expression. When studying the growth of *A oryzae* on rice, they report the down-regulation of five secondary metabolites, including the aflatoxin pathway which was significant in terms of change in RPKM. Their findings are consistent with previously characterised mutations in the aflatoxin gene cluster including transcription binding site mutations in the *aflR* promoter, a 250 bp deletion in the *aflT* 3' coding region, a frameshift mutation in the *norA* coding region, multiple synonymous mutations in the *verA* coding region and a 40 kb intergenic region deletion between the *norB* and *norA* genes.

The two studies described above (Akao et al., 2007; Gibbons et al., 2012) both report the little or lack of expression of aflatoxin genes in *A oryzae*, however do not discuss the mechanism of how or why the aflatoxin genes are not functional. A more focused investigation on the proteins AFLR and AFLJ was conducted by Kiyota et al. in 2011. AFLR is the protein product of the *aflR* gene, which is the transcriptional factor required for the activation of the aflatoxin biosynthetic pathway. The exact function of AFLJ, the gene product of *aflJ*, is undetermined but it is known to interact with AFLR and co-activate transcription. Experiments conducted by Kiyota et al. showed that in the absence of a functional AFLR protein, AFLJ on its own could not activate expression of the aflatoxin genes. The authors suggest that DNA mutations in *aflJ* lead to mutations in the AFLJ protein, which hinders its interaction with AFLR and down-regulates the aflatoxin pathway as a result.

Therefore, despite sharing such high sequence identity with aflatoxin-producing *A flavus* and an orthologous aflatoxin gene cluster, these genes do not appear to be functional in *A oryzae* and therefore are not expressed. This is consistent with findings that no strains of *A oryzae* are known to actually produce

aflatoxin (Chang and Ehrlich, 2010; Payne et al., 2006; Machida et al., 2008) and that its regular use in fermentation processes in the food industry is considered safe.

The genome of *A sojae* was only sequenced recently in 2011 (Sato et al., 2011). Following the sequencing and assembly using Newbler (`http://454.com/products/analysis-software/index.asp`), Open Reading Frames (ORFs), which represent genes, were identified using *ab initio*-based prediction tools. ORFs longer than 300 bp were predicted to be genes. A comparative analysis was performed by searching putative protein domains with BLASTp (`http://blast.ncbi.nlm.nih.gov`), and it was found that 81.7% of the ORFs had more than 90% identity with the ORFs of *A oryzae*.

Just as *A oryzae* and *A flavus* are thought of as equivalent non-aflatoxin-producing and aflatoxin-producing strains respectively, a similar relationship exists between *A sojae* and *A parasiticus*, with *A sojae* being the non-aflatoxin-producing version (Chang et al., 2007). Evidence shows that *A sojae*, like *A oryzae*, possesses an orthologous aflatoxin gene cluster, whose gene organisation is equivalent to that of *A parasiticus* (Chang et al., 2007; Machida et al., 2008), and is incapable of producing aflatoxins. This is attributed to a single point mutation from nucleotide 'C' to 'T' in the *aflR* transcriptional regulator gene, resulting in a stop codon and inducing premature termination of the gene. The resultant truncated AFLR protein cannot interact with its transcriptional co-activator AFLJ and therefore the aflatoxin biosynthetic pathway remains inactivated. This differs from *A oryzae*, where mutations in the *aflJ* is thought to cause the lack of aflatoxin producing ability (Kiyota et al., 2011).

Therefore, it is well understood that *A oryzae* and *A sojae* are non-toxigenic forms of *A flavus* and *A parasiticus* respectively, and both contain the aflatoxin biosynthetic pathway genes orthologous to their toxigenic counterparts. Recently, concerns have been expressed regarding the potential of *A oryzae* to activate its biosynthetic gene cluster to produce aflatoxin, and much research has been directed towards this area (Amaike and Keller, 2011). Species that are able to produce aflatoxins are scattered throughout phylogenetic trees, suggesting that the ability to produce aflatoxin was lost and gained several times during the course of evolution (Varga et al., 2011). It is reported that 30-80% of *A flavus* isolates may be non-aflatoxigenic, which may be accounted for by loss of selection pressure for secondary metabolite production as agricultural practices were developed (Chang and Ehrlich, 2010). It has also been proposed that non-aflatoxin producing strains retain the inactive aflatoxin genes in their genome

rather than removing them, to provide the organism with the benefit of being competitive within its ecological niche (N Tran-Dinh and D Midgley, personal communication). Determining the variation of aflatoxin genes between strains and the selective pressure of these genes as compared to the rest of the genome may be a useful starting point for pursuing this investigation.

5.2.1 Aflatoxin Types B and G

There are four major types of aflatoxins that are produced by *Aspergilli*: B1, B2, G1 and G2, each differing slightly in structure (Amaike and Keller, 2011). Toxigenic *A flavus* isolates produce aflatoxins B1 and B2 whereas *A minisclerotigenes*, *A parasiticus* and *A nomius* produce all four types (Yu, 2012), as summarised in Table 5.1. The aflatoxins have been named upon the basis of fluorescence colour – blue (B) or green (G) – under ultraviolet light. Of these toxins, aflatoxin B1 is the most toxic and carcinogenic to humans and animals.

Ehrlich et al. (2004) showed for the first time why *A flavus* isolates are incapable of producing G-type aflatoxins. DNA sequences from the 5' terminal (start) of the aflatoxin biosynthetic gene cluster of *A flavus* and three B- and G-aflatoxin producing species, including *A parasiticus* and *A nomius*, were extracted and sequenced. Alignment of these sequences using the DNAMAN software (`http://www.lynnon.com/`) revealed a 800 – 1500 bp deletion in *A flavus* spanning genes *norB* and *cypA* (Figure 5.1) that were predicted to encode a cytochrome P450 monooxygenase and an aryl alcohol dehydrogenase respectively. The lack of expression of these genes was experimentally confirmed. The role of the *cypA* gene in aflatoxin G production was also examined in *A parasiticus* by disrupting the cypA gene and measuring its expression. It was found that only B-type aflatoxins were produced, with no formation of G-type aflatoxins, demonstrating that the *cypA* gene is required in the biosynthesis of aflatoxin G. When the *cypA* gene is present in the genome, it is located directly adjacent to the aflatoxin gene cluster. *A flavus* lacks the complete gene, rendering it unable to produce G-type aflatoxins. The same deletion in the aflatoxin gene cluster is also present in *A oryzae*, the non-toxigenic variant of *A flavus*. The authors conclude by suggesting that B-type aflatoxin-producing species diverged from B and G-type aflatoxin producers.

TABLE 5.1. Summary of the aflatoxin producing ability of major species of *Aspergillus* (Varga et al., 2011).

Species	Aflatoxin B1	Aflatoxin B2	Aflatoxin G1	Aflatoxin G2
Aspergillus flavus	✓	✓	×	×
Aspergillus minisclerotigenes	✓	✓	✓	✓
Aspergillus nomius	✓	✓	✓	✓
Aspergillus oryzae	×	×	×	×
Aspergillus parasiticus	✓	✓	✓	✓
Aspergillus sojae	×	×	×	×

Therefore, among the *Aspergilli*, non-toxigenic strains have been found to contain the aflatoxin biosynthetic pathway genes orthologous to toxigenic strains that occur in this genus. The question of why some strains of *Aspergillus* produce aflatoxin and other strains do not, remains an imperative and unresolved problem. As mentioned previously, this distinction is important to ensure food safety. Furthermore, why these non-functional genes have persisted in the genomes of non-toxigenic isolates, and the exact mechanisms by which these genes were attained, are still unclear. These problems form the motivations driving the current study, in which we aim to investigate the evolutionary dynamics of *Aspergillus* fungi to shed light on their evolutionary relationships, and the behaviour of the aflatoxin genes at the evolutionary level. The concept of evolutionary dynamics is defined in the following section.

5.3 Evolutionary Dynamics among Fungal Organisms

Evolutionary dynamics provides an avenue to decipher evolutionary processes driving change in genome sequences to give rise to new variations or novel sequences. This can reveal which regions of the genome have high rates of mutations, in which bases in the genome are substituted with other bases, and which regions are under higher selective pressures than others to undergo change.

A common method of studying evolutionary dynamics is via the construction of phylogenetic trees at the gene, protein or species levels, and is a hypothesis of the assumed order of evolutionary events (`http://en.wikipedia.org/wiki/Phylogenetics`). The two main techniques for inferring phylogenetic trees are Maximum Likelihood and Maximum Parsimony. Maximum Likelihood methods seek to find the tree with the maximum probability of producing the observed data, while Maximum Parsimony methods seeks the tree that requires the least evolutionary change (the fewest changes in

character states) to explain the observed data (Sober, 2004). This section reviews studies published in literature that investigate such evolutionary dynamics of *Aspergillus* fungi at the species level.

5.3.1 Classification of *Aspergillus* Species

The highly contrasting effects of *Aspergillus* fungi on foods and human health necessitates accurate methods of distinguishing safe from toxic forms. Traditional molecular-based methods such as mitochondrial DNA restriction fragment length polymorphism (RFLP) can separate *A flavus*, *A parasiticus* and *A nomius*; while polymerase chain reaction (PCR) of rRNA internal transcribed spacer (ITS) regions and amplified fragment length polymorphism (AFLP) can separate *A flavus* / *A oryzae* from *A parasiticus* / *A sojae* but not distinguish the individual species (Chang et al., 2007).

A comparative whole genome study between *A flavus* and *A oryzae* revealed 99.5% sequence identity at the genome level and 98% identity at the protein level (Rokas et al., 2007). Similarly, *A flavus* and *A parasiticus* share 97-99% nucleotide identity (Chang et al., 2007) while 82% of protein-coding genes of *A sojae* share more than 90% identity with corresponding genes in *A oryzae* (Sato et al., 2011). This high degree of similarity and close relatedness among these *Aspergillus* species presents a challenge to the accurate and unambiguous classification of each species.

In using phylogenetic methods, Nakamura et al. (2011) state that commonly used genetic evolutionary markers such as 18S rRNA, 28S rRNA and the ITS regions are almost identical across this group of fungal organisms, which limits their ability to successfully differentiate between them. It was stated in Section 2.3 that *A flavus* and *A minisclerotigenes* for example, differ at only 1 base in their ITS sequence. Instead, the authors exploit the more variable sequences in aflatoxin biosynthetic genes to perform multi locus sequence analysis (MLSA) which uses multiple gene loci (in this case the aflatoxin genes *aflR*, *aflT*, *norA*, *vbs*), as opposed to just single genes, and increased sequence length to improve phylogenetic resolution power. This technique was used in a comparative study of 22 strains from the *Aspergillus* genus. These gene sequences were used to construct estimated phylogenetic trees using the Neighbour-Joining method (Saitou and Nei, 1987; Studier and Keppler, 1988). The resulting multi-gene phylogenetic tree shows a clear separation between *A flavus* and *A oryzae*, and *A sojae* and *A parasiticus*, whereas trees based on individual gene sequences do not. In addition, the branches are supported by high bootstrap values mostly greater than 97%, whereas the single gene trees have low bootstrap support, thus

lacking reliability. It is notable that the MLSA tree shows the strains of *A oryzae* and *A flavus* as each others' closest relatives. The reasons as to why this particular set of four genes was chosen is unclear. Furthermore, this study only included 3 strains of *A oryzae*, and Nakamura et al. recognise that further study with a greater number of strains for all species is required.

Phylogenetic analysis was used by Varga et al. (2011) to classify 2 new recently identified *Aspergillus* species, now called *Aspergillus pseudocaelatus* and *Aspergillus pseudonomius*. The authors used ITS sequences, along with the calmodulin and β-tubulin genes. Following the sequencing, editing and alignment of these gene sequences, phylogenetic analysis based on maximum parsimony was performed using PAUP (`http://paup.csit.fsu.edu/`). The robustness of the topology of the trees was assessed by running 1000 bootstrap replicates. The three phylogenetic trees corresponding to the three genes appear to have different topologies.

The studies described thus far have been based on the nuclear genomes of *Aspergillus* fungi. This DNA, contained within the nucleus of cells, encodes the majority of the entire genome. A small portion of the genome is located is another cellular component called the mitochondria. The mitochondrial genome was the focus of Joardar et al.'s recent study into the classification and evolution of fungi by focusing on mitochondrial genomes. The authors reason that mitochondrial sequences may provide new and vital insights not gained by studying the nuclear genome alone. Six *Aspergillus* and three *Penicillium* mitochondrial genomes were sequenced and assembled using available *de novo* contigs and mapping to a known mitochondrial reference genome. The quality threshold for selecting contigs was defined as those with 2-fold high quality coverage of each nucleotide and showing evidence of circularity, as mitochondrial genomes are circular as opposed to nuclear genomes which are linear. The genomes were annotated for large-scale rearrangements including insertions and deletions, and Open Reading Frames (ORFs) predicted via searching using BLASTp (`http://blast.ncbi.nlm.nih.gov`) to assign functional classifications based on sequence similarities. ORFs longer than 100 amino acids were considered putative genes. This threshold length appears to be standard for prediction of ORFs and has been used in other fungal genome studies (Machida et al., 2005; Sato et al., 2011).

To examine the evolutionary relationships among the fungi, Joardar et al. performed phylogenetic analysis of 14 core concatenated proteins that were found from the annotation of all mitochondrial genomes

sequenced. These genes, involved in major mitochondrial processes such as oxidative phosphorylation, ATP synthesis and mitochondrial protein production, share a high degree of sequence conservation as well as synteny, which refers to the conservation of sequence blocks within sets of chromosomes. Sequences were first concatenated and aligned using Muscle (Edgar, 2004), and the PHYLIP package (`http://evolution.genetics.washington.edu/phylip.html`) was used to generate a Maximum Parsimony tree which was compared with a Maximum Likelihood tree generated using the Randomised Axelerated Maximum Likelihood (RAxML) program (Stamatakis et al., 2005). The two mitochondrial protein trees were consistent, and also had similar topologies to trees built using multiple nuclear protein trees, thus validating the results obtained (Joardar et al., 2012).

This study was complemented by conducting phylogenetic analysis of individual mitochondrial genes using ITS sequences (Section 2.3). These individual gene trees, however, did not have high bootstrap support, compatible topologies between trees, nor high resolving power to differentiate species. These findings are consistent with those made by Nakamura et al.. Therefore the authors propose that the use of concatenated core mitochondrial proteins can be valuable for phylogeny construction at the species level. This reiterates the claim made by Galagan et al. (2005) that single gene phylogenies can conflict with whole-genome trees, and that the latter provides greater resolving power on the basis of concatenated genes. Figure 5.2 presents a phylogenetic tree based on trees in literature.

5.3.2 Evolution of *Aspergillus* Species

Determining the evolutionary mechanisms of Aspergillus genomes has been the focus of many studies as this has the potential to shed new light on genome architecture, population biology, secondary metabolism including aflatoxin production, and virulence mechanisms (Gibbons and Rokas, 2013).

As described previously, members of this genus, namely *A flavus*, *A parasiticus*, *A oryzae* and *A sojae* are closely related and share high sequence similarity (Chang et al., 2007; Rokas et al., 2007; Sato et al., 2011). A noteworthy observation is that the genomes of *A flavus* and *A oryzae* are similar in size with each other (36.8 Mb and 37.2 Mb respectively, where 'Mb' stands for 'mega base pairs' which is 1,000,000 base pairs) and significantly larger than the genomes of *Aspergillus nidulans* (30.1 Mb) and *Aspergillus fumigatus* (29.4 Mb) (Figure 5.2). Upon sequencing the genome of *A oryzae* with 9x coverage, Machida et al. (2005) performed analysis to find syntenic blocks shared with *A fumigatus*

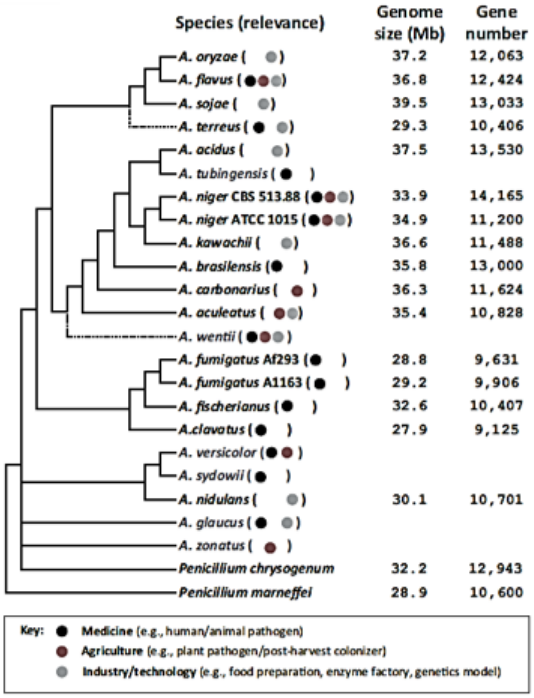


FIGURE 5.2. Phylogenetic tree and relevance of various Aspergillus genus organisms (Gibbons and Rokas, 2013).

and *A nidulans*, as well as blocks specific to *A oryzae*, that are particularly enriched for genes encoding secondary metabolites. This syntenic analysis was performed by defining a region of conserved synteny as the longest contiguous block of at least one orthologue followed by another orthologue or homologous region. Orthologues between the 3 species were identified through bi-directional BLASTp (`http://blast.ncbi.nlm.nih.gov`) searches that had a bit score greater than 200, while homologous regions were predicted using tBLASTx (`http://blast.ncbi.nlm.nih.gov`) matches that had a bit score greater than 100. The analysis revealed a mosaic structure of the genome, believed to be characteristic of horizontal gene transfer, an evolutionary process involving the transmission of genes between species by means other than from parents to offspring via reproduction. Machida et al. report the larger genome size of *A oryzae* is a consequence of lineage-specific sequence acquisition as opposed to sequence loss in *A nidulans* and *A fumigatus*. There is evidence these genes did not originate from whole-genome or segmental duplication (Chang and Ehrlich, 2010), and that horizontal gene transfer has shaped the *Aspergillus* genome where it has served as both the donor and recipient lineage (Gibbons and Rokas, 2013).

As described previously, most studies have centred on the nuclear genomes of fungal organisms. In contrast, Joardar et al. (2012) instead focused on mitochondrial genomes as a means of studying mechanisms of fungal evolution. The *Aspergillus* and *Penicillium* genomes sequenced in this study were subject to annotation using BLASTp and BLASTn searches (`http://blast.ncbi.nlm.nih.gov`) to identify intron insertion boundaries, where introns are non-coding regions between genes. These predicted boundaries were verified by searching putative protein products in tBLAST (`http://blast.ncbi.nlm.nih.gov`). Although the core mitochondrial genes were found to be highly conserved, the high level of variation in intron distribution between species supports experimental evidence for horizontal gene transfer and recombination in mitochondrial genomes. Furthermore, Joardar et al. attribute this intron variability as the primary source of difference in genome size, and state that horizontal gene transfer may be common in *Aspergillus* and *Penicillium* mitochondrial genomes.

A flavus and *A oryzae* share 99.5% nucleotide sequence identity, with the number of unique genes between the two fungi representing only 2.1% and 2.7% of total protein coding genes in the two organisms respectively (Rokas et al., 2007). This has led many to believe that *A oryzae* is domesticated from an ancestor of *A flavus* (Machida et al., 2005), or that *A oryzae* is a domesticated ecotype of *A flavus* (Cleveland et al., 2009; Payne et al., 2006), where the domestication was driven by significant remodelling of metabolism genes and pathways (Gibbons et al., 2012). In particular, the parsimony inferred phylogenetic tree constructed by Gibbons et al., using '100,084 high-quality genome-wide variant sites', shows all 8 *A oryzae* strains separate to all 8 *A flavus* strains, suggesting that *A oryzae* is a different species to *A flavus*. Although this tree is supported by high bootstrap values, the choice of *A oryzae* strains may have been a factor for this observation. Similarly, it is proposed that A sojae is domesticated from A parasiticus (Chang et al., 2007). Although there is genetic and molecular evidence in favour of this hypothesis (Rokas et al., 2007), some researchers believe that *A flavus* and *A oryzae* and *A parasiticus* and *A sojae* are in fact the same species, just different isolates. It is claimed that *A oryzae* originated from a domestication event rather than being a domesticated species of *A flavus*, and that these species names were assigned by the food industry for convenience to differentiate between safe and toxic species (N Tran-Dinh and D Midgley, personal communication). Therefore the lack of agreement on this issue is a reflection on the complexity and challenges facing the elucidation of evolutionary mechanisms as well as the taxonomy of the *Aspergillus* genus.

5.4 Summary

The shortcomings and unanswered questions highlighted in this review have directed and informs the objectives of this study. We aim to use evolutionary dynamics as a comparative tool to understand the differences between aflatoxin producing and non-aflatoxin producing species, and gain more insight into the nature of the highly debated relationship between *A flavus* and *A oryzae*.

CHAPTER 6

Methods for Analysing Evolutionary Dynamics

6.1 Overview

It was established in the literature review in Chapter 5 that one of the objectives of this study is to perform a comparative analysis of the evolutionary dynamics of the aflatoxin genes. The elucidation of these evolutionary dynamics is facilitated by phylogenetic analysis, which involves the construction of phylogenetic trees. This will be achieved by investigating the aflatoxin genes from a group of novel fungal genomes that were sequenced by the mycology team led by Dr Nai Tran-Dinh and Dr David Midgley at CSIRO Animal, Food and Health Sciences.

This chapter provides a detailed explanation of our analysis pipeline, from assembling the raw sequence reads of each of the 39 fungal genomes (Section 6.2), to extracting and characterising the aflatoxin genes, and analysing their evolutionary dynamics (Section 6.3).

6.2 Fungal Genome Assembly

This section outlines the steps taken for the assembly of the fungal genomes. A detailed description of the sequencing process and assembly algorithm are provided, followed by an explanation and justification of the different types of assemblies we produced.

6.2.1 Next-Generation Sequencing

Sequencing is defined as "the process of determining the order of nucleotide bases (A, C, T, and G)" of DNA (Illumina, 2013b). The advent of Next-Generation Sequencing (NGS) technologies has revolutionised this process to be massively parallel, scalable and high-throughput, with the newest instruments rapidly producing terabases (Tb) of sequence data from a single run.

The 39 novel fungal genomes were sequenced in two stages using Illumina NGS technologies. It is noteworthy that the life cycle of fungi alternates between a haploid and diploid stage. During the haploid stage, there is only one set of chromosomes, whereas in the diploid stage, there are two identical copies of each chromosome. To reduce the computational complexity and challenge of dealing with ambiguities from two copies of the genome during assembly (refer to Section 6.2.2), the haploid genome of the 39 fungal organisms was extracted. Each of the DNA samples was then passed through Illumina MiSeq machines as a rapid pre-sequencing step to assess the quality of the DNA preparation, which was followed by sequencing using Illumina HiSeq machines.

The sequencing process (Figure 6.1) starts by sharding the genomic DNA into a collection of random fragments. Each fragment is sequentially resynthesised using a DNA template strand, and the nucleotide bases that are incorporated into the newly synthesised fragment are identified by a laser that captures the fluorescence emitted (Illumina, 2013a). The strings of bases identified from each fragment are known as 'reads'. This approach has been termed "sequencing-by-synthesis" by Illumina. The fragments can be uniformly sequenced using millions of parallel reactions, leading to the high-throughput capabilities of NGS. Furthermore, each base in a read is assigned a Phred quality score Q as a metric of the accuracy of the sequencing platform (Illumina, 2011). This score measures the estimated probability p that a base is called incorrectly, that is, identified as the wrong nucleotide, and is calculated using the formula:

$$Q = -\log_{10} p \tag{6.1}$$

Therefore, smaller values of p are associated with higher quality scores Q. The sequencing-by-synthesis approach described is claimed to have the "highest percentage of error-free reads" (Illumina, 2011).

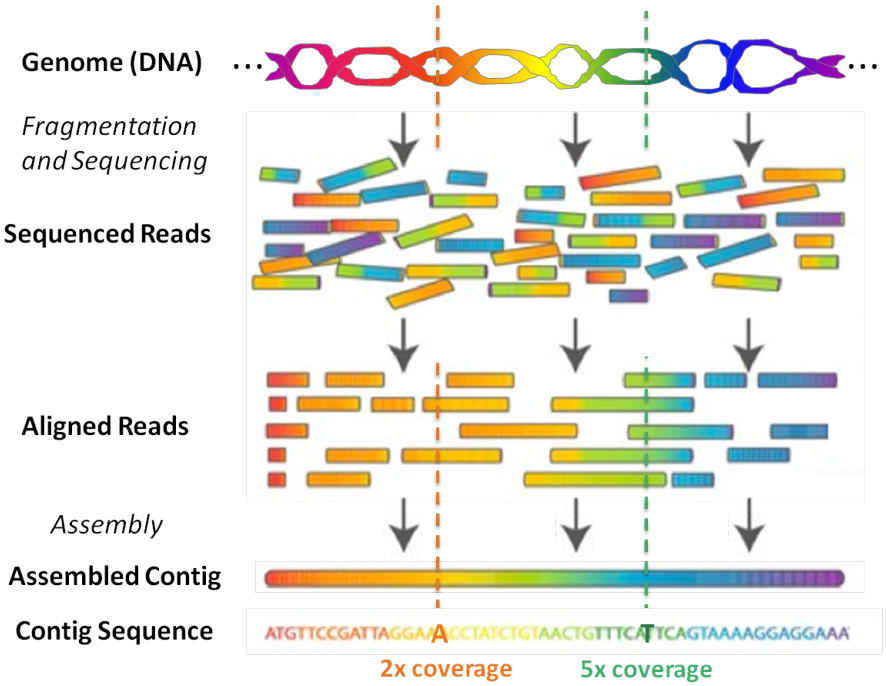


FIGURE 6.1. Overview of the assembly of Next-Generation Sequencing reads. Image adapted from `http://upload.wikimedia.org/wikipedia/commons/b/bd/Whole_genome_shotgun_sequencing_versus_Hierarchical_shotgun_sequencing.png`.

Each base (now referring to position in the original genome), is sequenced with a given depth of coverage – a measure of the average number of times the base has been sequenced – and is reflected in the number of reads containing that base after alignment. The concept of coverage is illustrated in (Figure 6.1), where the base 'A' (denoted by the orange dashed line) has a coverage of 2 fold as this position is contained in 2 separate reads, while the base 'T' (denoted by the green dashed line) has a coverage of 5 fold. The sequence reads produced, however, are not perfect and often contain errors that arise as artefacts of the sequencing process. Therefore, higher coverage is desired as we can exploit the redundancy in the reads to correct errors using consensus bases amongst all the reads. A more detailed explanation of the nature of sequencing errors and techniques to resolve them is presented in Section 6.2.3.

The MiSeq machine produced sets of 250 bp reads for each genome that were then trimmed from the ends to remove regions with low Phred quality scores as part of the sequencing protocol. The reads were

FIGURE 6.2. HiSeq and MiSeq reads produced using paired-end sequencing.

therefore of variable length, however the majority were 250 bp in length. On the other hand, the output of the HiSeq sequencing was sets of shorter, 100 bp reads. While both techniques produce sequence reads of similar quality, the power and throughput of the sequencing machines is the distinguishing feature. MiSeq machines produce up to 15 million reads (8.5 Gb) whereas HiSeq machines can produce up to 1.2 billion reads (180 Gb) in approximately the same amount of time (Illumina, 2013c), enabling HiSeq reads to be sequenced with a much greater coverage.

The reads produced from both these sequencing protocols are in the form of paired-end reads, where each DNA fragment is sequenced at both ends (Figure 6.2). This is in contrast to single-read sequencing where only one end of the DNA fragment is sequenced, resulting only in the reads labelled ‘Read 1’ in Figure 6.2.

The goal of assembly is to construct reads into longer, contiguous sequences known as contigs, using overlapping regions from the aligned reads (Figure 6.1). The longer the contigs, the fewer contigs required to represent the genome, and the fewer gaps that are present in the assembly. It has been observed that assemblies of the short reads produced by NGS often contain a high number of gaps in regions of no overlaps between reads, leading to a highly fragmented assembly of poor quality contigs (Illumina, 2013a). In particular, this is true for regions of the genome composed of repetitive elements. The potential and benefit of using paired-end reads is realised in resolving such gaps that cause contigs to break across repetitive regions and other ambiguous regions that are difficult to sequence. As demonstrated in Figure 6.3, the approximately constant length of the distance between the pairs provides information about their relatedness, which can be exploited to bridge long gaps and map reads over repetitive regions with higher precision (Illumina, 2013a). The resulting enhanced alignments and superior assemblies thereby validate the use of paired-end sequencing.

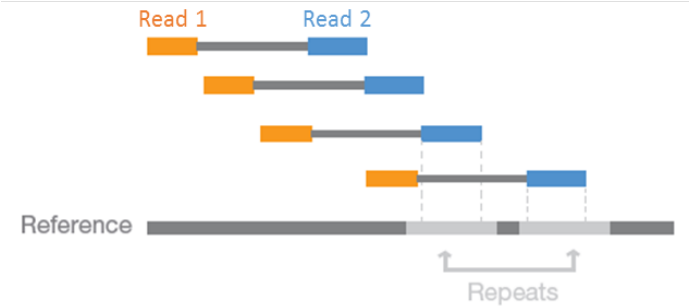


FIGURE 6.3. Alignment of paired-end reads to resolve ambiguous regions such as repetitive regions. The orange and blue blocks represent paired-end reads. Image adapted from 'An Introduction to Next-Generation Sequencing Technology', Figure 4 (Illumina, 2013a).

6.2.2 Velvet Assembler

The decision of whether or not assembly of reads is required is dependent of the purpose or application of the research. Studies examining gene expression or Single Nucleotide Polymorphisms (mutations at single bases) do not always require full assemblies as the genetic information sought after is contained within the reads themselves. As an example, the tool Blue (Greenfield et al., 2013) discussed in Section 6.2.3 does not require an assembly for error correction, and can directly process sequence reads. On the other hand, studies centred around whole genome and *de novo* sequencing of novel organisms (where no reference sequence is available) such as in the present study, assembly is critical in order for the downstream genome analysis to be meaningful.

Our assembly pipeline utilised the Velvet assembler (Zerbino and Birney, 2008), which was chosen due to its high popularity and wide use, reflected in its 2,509 citations, for the *de novo* assembly of short reads that are produced by sequencing technologies such as Illumina and 454 Sequencing. The underlying algorithm of Velvet is based on a de Bruijn graph model, and proceeds in three stages. The first stage involves constructing the graph, and begins by converting the sequence reads into a set of overlapping k-mers, where a k-mer is defined as a DNA sequence of length k bp and specified by the user. Due to the double-stranded nature of DNA, which consists of two strands that are reverse complements of each other, reads could be sequenced from either strand and the difference in directionality must be taken into account. Velvet handles this by generating the reverse complement for each k-mer observed in the reads, and both k-mers are added as keys to a hashtable and mapped to the ID of the first read the k-mer

was found in as well as its position in the read. A second hashtable is created which stores each read as the key mapped to its original k-mers that are overlapped or partly shared in subsequent reads. Each uninterrupted sequence of original k-mers is then created as a node. The graph is then traversed using the information in the hashtables about overlapping k-mers to join nodes via the creation of directed edges.

The structure of the de Bruijn graph is displayed in Figure 6.4, where nodes, represented as rectangles, are a series of k-mers in which adjacent k-mers overlap by $k-1$ nucleotides. Each node is attached to a twin node that represents the k-mers of the reverse complement read, which creates a super node called a 'block'. A directed edge from node i to node j connects the nodes if the last k-mer of node i overlaps with the first k-mer of node j. This is illustrated in Figure 6.4 where the last 5-mer of the node labelled i, 'GATTG' overlaps with the first 5-mer of node j, 'ATTGA' by the 4-mer 'ATTG'. A directed edge in the reverse direction from $\overline{j}$ to $\overline{i}$ is also added, where $\overline{j}$ and $\overline{i}$ are the twin nodes of i and j respectively. All operations performed on a node are accordingly applied to its twin automatically. Likewise, modifications to an edge are propagated to its reverse edge symmetrically. Reads are mapped as paths along the nodes and edges of the graph.

The second stage of the algorithm performs a simplification step on the constructed graph to lower the cost of computation and memory by reducing the number of nodes and edges. The start and end of blocks coincide with the start and end of reads, which results in chains of blocks that form linearly connected subgraphs. These chains of blocks are iteratively merged into single blocks whenever there occurs a node k that only has one outgoing edge to a node i that only has one incoming edge, as shown in Figure 6.4 by the orange boxes. As mentioned previously, all operations on nodes and edges are implicitly performed on twin nodes and edges, hence the merging step collapses nodes along with their corresponding twin nodes. This simplification step is feasible as there is no loss of information.

The final step of Velvet following simplification of the graph is removal of errors in the reads, which involves correcting and merging paths where appropriate. The approach taken uses topological features in terms of the location of these errors within the read, such as internally or close to either end, and the resulting anomalous structures in the graph, to guide the detection and removal of these errors.

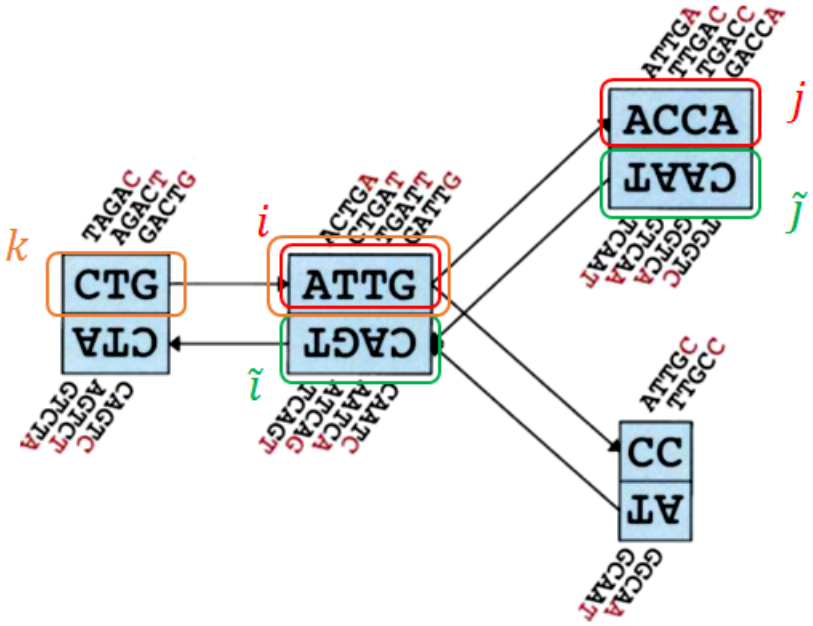


FIGURE 6.4. de Bruijn graph structure that are used by Velvet for assembly. Image adapted from Zerbino and Erwin (2008).

As a consequence, the Velvet algorithm requires four parameters to be specified for each genome assembly:

- `k`: the k-mer length.
- `cov_cutoff`: the minimum k-mer coverage depth, that is, the frequency with which a k-mer has been observed among the set of reads. Because the coverage of sequence reads is not constant and suffers from low coverage areas, the `cov_cutoff` parameter can be used to exclude contigs containing k-mers with coverages below this cutoff from the final assembly.
- `exp_cov`: the expected k-mer coverage depth.
- `ins_length`: the length of the sequenced DNA fragment. For our genomes, this was set to a constant 300 bp.

It is imperative that the values of these parameters are chosen carefully for each genome in order to achieve a high quality assembly, that is, with longer and therefore fewer contigs. Zerbino discusses the choice of k as often a trade-off between sensitivity and specificity. Setting a k-mer length that

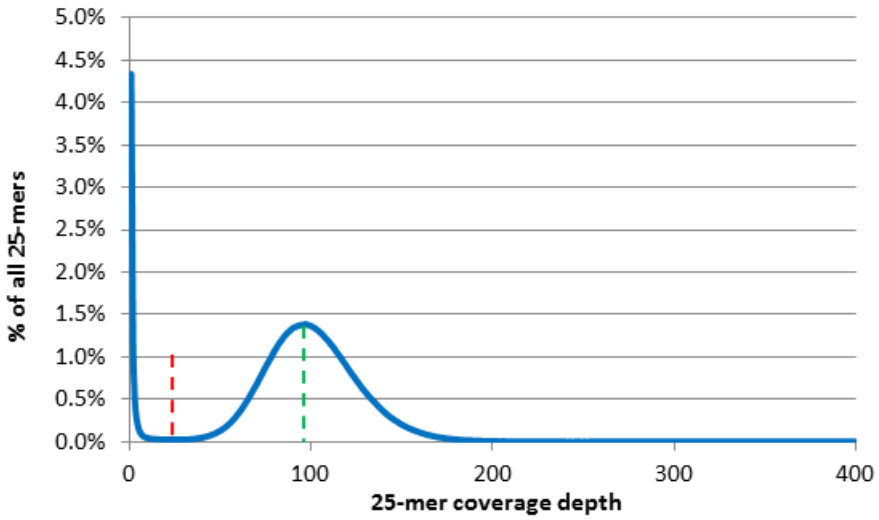


FIGURE 6.5. k-mer frequency (coverage) distribution histogram produced using Tessel with k = 25 for the haploid genome of *Aspergillus flavus* strain 342.

is too short results in more non-specific matches when finding overlapping regions, while longer k-mers provide higher specificity but overlapping regions may come from lower coverage areas. Previous experience with genome assemblies of several different types of organisms informs us that k-mer lengths between 41 and 57 are highly effective (P Greenfield, personal communication), and hence assemblies were carried out using both these k-mer lengths in order to evaluate their efficacy in the assembly of fungal genomes.

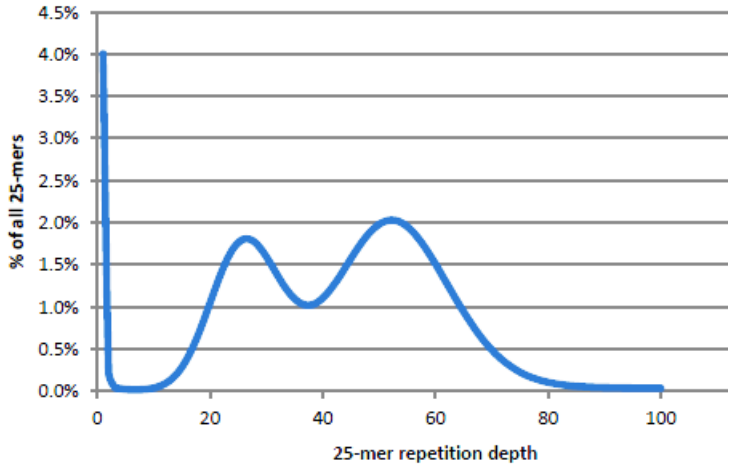


FIGURE 6.6. k-mer frequency distribution produced using Tessel with k = 25 for a diploid genome. Image from Greenfield et al. (2013).

The choice of suitable values for the minimum coverage and expected coverage for each genome were guided by an examination of k-mer frequency distribution histograms and determined empirically. These histograms, which plot k-mer coverages among the set of reads, were produced using Tessel (to be published), a fast and scalable k-mer counting tool. I helped to implement this multi-threaded program with my supervisor Paul Greenfield as part of a previous research project. The parameters required for Tessel are the k-mer length and the expected size of the genome, which in this case were 25 and 40,000,000 bp respectively, and the number of threads to use. An example of such a distribution of 25-mers for an *Aspergillus flavus* genome is presented in Figure 6.5 below. We recall that sequencing of all fungi in this study was performed for the haploid genome, hence only one peak is observed in the histogram. A 25-mer frequency distribution histogram for a diploid genome is presented in Figure 6.6 (Greenfield et al., 2013). The differences in distribution between the haploid, containing one copy of the total DNA, and diploid, which contains two copies, are clear. Figure 6.6 displays two overlapping peaks; one for the 25-mers found in both copies (homozygous), and the other for k-mers uniquely found in only one copy (heterozygous). It is these heterozygotic differences, where at these positions in the genome half the reads will contain one nucleotide and the other half will contain another, that create ambiguity and thus make diploid genome assembly more challenging.

From Figure 6.5, the frequency distribution histogram of 25-mers reveals a plateau between the first spike and second peak for all genomes. For the histogram presented, the middle of this plateau (denoted by the red dashed line) corresponds to a 25-mer coverage depth, or frequency, of 26. 25-mers with coverages lower than 26 occur infrequently and are mostly likely caused by the presence of sequencing errors, and thus should be omitted from the assembly. This particularly applies to the singleton spike representing 4.3% of 25-mers that only have a frequency of 1, indicating they are unique in the entire set of reads for the genome. The coverage value of 26 therefore represents the coverage cutoff parameter. Disregarding the singleton spike of low coverage 25-mers, the peak of 1.4% occured at a coverage of 97 (denoted by the green dashed line), which formed the expected coverage parameter.

It is important to note that while our calculations for `cov_cutoff` and `exp_cov` were performed using $k = 25$, Velvet requires these parameter values in terms of the k that the user specifies for the assembly. The complexity of this conversion is highlighted in the Equation (6.2), where R = read length, K is the actual k-mer length used in the calculation, L is the desired k-mer length in terms of

TABLE 6.1. Conversion of coverages from 25-mer lengths to desired 41-mer and 57-mer lengths for HiSeq reads where read length R = 100.

Parameter	R	K	L	C_K	C_L
`cov_cutoff`	100	25	41	26	21
`exp_cov`	100	25	41	97	77
`cov_cutoff`	100	25	57	26	15
`exp_cov`	100	25	57	97	56

the k specified to Velvet, C_K is the actual coverage in terms of K-mers and C_L is the desired coverage in terms of L-mers.

$$C_L = C_K \times \frac{R - L + 1}{R - K + 1} \tag{6.2}$$

Table 6.1 demonstrates the conversion of coverages from the k = 25 used in the actual calculation to the k = 41 and k = 57 that were used in the assembly. As the value of k increases, the length of the k-mer increases and therefore we expect the coverages C_L to decrease relative to C_K. Indeed, as we go from 25-mers to 41-mers, the `cov_cutoff` and `exp_cov` decrease from 26 to 21 and 97 to 77 respectively. This reduction is even larger as we go from 25-mers to 57-mers, where `cov_cutoff` and `exp_cov` become 15 and 56 respectively.

6.2.3 Sequence Error Correction and Assembly of Sequence Reads

The first Velvet assemblies were carried out using the set of original, raw HiSeq reads. Despite the high quality sequence reads produced by Illumina HiSeq, it has been reported that a set of 6 billion reads can still contain 1-2% sequencing errors including substitutions, insertions, deletions and ambiguous or uncalled bases (Minoche et al., 2011). The underlying biochemical principle upon which the sequencing is based impacts the types of errors that are generated. For Illumina machines, the polymerase-based sequencing-by-synthesis technology makes the sequence reads susceptible to substitution errors or mismatches, where a base is incorrectly identified as another base (Greenfield et al., 2013). The presence of these errors can detrimentally impact the quality of results of assemblies and other forms of sequence analyses such as alignments and genome annotations. Consequently, for each genome, the raw HiSeq

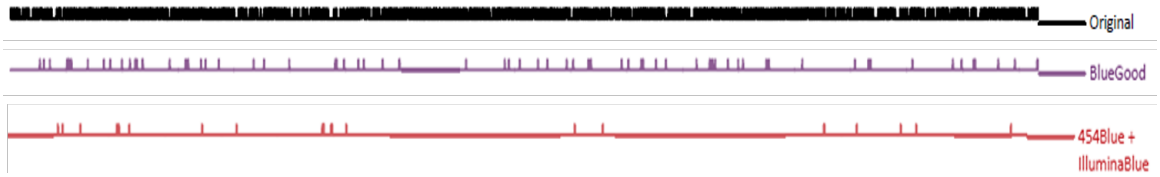


FIGURE 6.7. Effects of error correction on a bacterial genome. Image adapted from Greenfield et al. (2013).

sequence reads were corrected for these errors using the tool Blue (Greenfield et al., 2013) which focuses on correcting the errors in k-mers with coverage values to the left of the red dashed line in Figure 6.5 for reasons discussed above. The set of corrected reads were then subject to a filtering process that retained reads only if 80% of the constituent k-mers had coverages above the minimum coverage threshold. These 'good' reads were then passed on to Velvet for assembly.

To investigate how error correction and the longer length reads can improve the quality of our fungal genome assemblies, the genomes were also assembled using error corrected HiSeq reads in conjunction with error corrected MiSeq reads. While the MiSeq sequence reads are generally of lower coverage than the HiSeq reads, the longer read length means that each read may provide additional information, supplementary to the shorter HiSeq reads, that can help Velvet join together contigs that would otherwise be separated by repetitive and hence ambiguous regions (P Greenfield, personal communication). The diversity resulting from combining the short HiSeq reads with the longer MiSeq reads, together with the 'pairedness' of the reads, provides more powerful capabilities for resolving large gaps and repetitive regions to produce higher quality contigs. The importance and effects of error correction is clearly portrayed in Figure 6.7, which compares a raw, uncorrected bacterial genome ('Original') with an error corrected genome using only HiSeq reads ('BlueGood') and an error corrected genome using both HiSeq and 454 (long) reads ('454Blue + IlluminaBlue'). Each mark on the lines represent a 1000 bp region which contains an error. It is evident that the number of errors is greatly reduced using just corrected HiSeq reads, and this reduction is even more significant when using combined corrected long and short reads, which fixes almost all the errors present in the original set of Illumina reads.

To summarise, 39 novel haploid fungal genomes were sequenced using NGS. To achieve whole genome analysis, the reads needed to be assembled into longer contigs. For each of k = 41 and k = 57, three types of assemblies were constructed using (i) raw HiSeq reads, (ii) error corrected HiSeq reads and (iii)

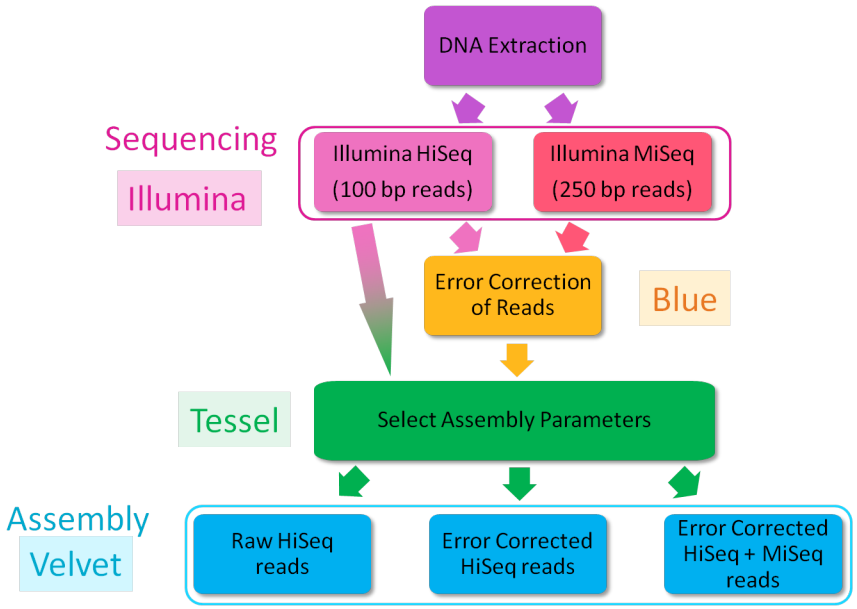


FIGURE 6.8. Summary of the sequencing and assembly workflow. The boxes represent which tools were utilised to aid that stage of analysis.

error corrected HiSeq with error corrected MiSeq reads. This genome assembly workflow is illustrated in Figure 6.8 below. The results of the various assemblies are presented and discussed in Section 7.2.

6.3 Evolutionary Dynamics of Aflatoxin Genes

With the assembled contigs available for all 39 sequenced genomes, the set of contigs corresponding to the 17 aflatoxin producing species were subject to further processing to extract the aflatoxin biosynthesis genes. Once the genes were found, comparative analyses could be performed via the construction of phylogenetic trees to elucidate the evolutionary dynamics of the genes. The methods adopted for achieving this are detailed in this section.

6.3.1 Characterisation of Aflatoxin Gene Pathway

The genes were extracted from the contigs in an iterative manner. The first iteration involved finding contigs which contained one or more of the 25 genes in the pathway, which were passed onto the second iteration that involved finding the actual genes from this set of filtered contigs. By using only a filtered

TABLE 6.2. Details of the full aflatoxin pathway reference sequences downloaded from NCBI.

Accession Number	Organism	Length (bp)	Reference
AY510451	*Aspergillus flavus* strain AF13	88,318	Ehrlich et al. (2005)
AY510452	*Aspergillus flavus* strain BN008	82,511	Ehrlich et al. (2005)
AY510453	*Aspergillus flavus* strain AF70	75,829	Ehrlich et al. (2005)
AY510455	*Aspergillus flavus* strain AF36	78,264	Ehrlich et al. (2005)
AB196490	*Aspergillus oryzae* strain RIB 40	89,641	Tominaga et al. (2006)
AY371490	*Aspergillus parasiticus*	82,081	Yu et al. (2004)
AY510454	*Aspergillus nomius* strain NRRL13137	75,352	Ehrlich et al. (2005)

subset of contigs to extract the individual genes, the search space was greatly reduced from the order of thousands of contigs (Figure 7.1) to a handful, and thus improving the computational efficiency of the process.

As we did not know what the aflatoxin gene sequences were in our novel fungal genomes, a set of 1,743 known, published reference sequences were downloaded from the NCBI Nucleotide databse (`http://www.ncbi.nlm.nih.gov/nucleotide/`) that represented the sequence of at least one gene from the aflatoxin pathway of *Aspergillus* species. This reference set also comprised of 7 sequences that spanned the entire pathway and were annotated with the start and end positions of all 25 genes, which was useful for determining the order and orientation of individual genes when they would be found in our genomes. The details of the 7 full pathway reference sequences are summarised in Table 6.2.

Each of the 1,743 reference sequences was converted to short seeds in the form of k-mers, where k was set to 25 to ensure specificity in the matching, which were then used as a filter to match contigs containing identical 25-mers. This step was accomplished using tools developed by Paul Greenfield, and was automated using custom written batch files that I wrote for this purpose.

Following this step, the contigs matched for each gene in the pathway had been identified, but the actual locations and sequences of the gene within the contigs remained to be found. Each set of matched contigs for each genome was aligned to a reference full aflatoxin pathway sequence of the same species from Table 6.2 to determine the orientation of the contigs. From the alignment, the regions in the contigs that did not match the reference sequence were trimmed, as these corresponded to regions adjacent, and

outside of the aflatoxin gene cluster. The contigs were concatenated to form a single sequence in the correct orientation, which was realigned to the reference sequence using Mauve (Darling et al., 2004). The start and stop positions of each gene annotated in the reference sequence was used to directly extract the gene sequence from our contigs, where the start of genes are marked with the sequence 'ATG'.

This procedure, in particular the alignment against the reference, enabled characterisation of the aflatoxin pathway in each genome in terms of what genes are present or missing, partially or completely, and whether there are rearrangements or deviations in the order of the genes between the organisms. This formed the first key step in understanding the structure of the aflatoxin pathway in these novel genomes. The structure of the pathway is presented in Figure 7.5.

The investigation of the whole pathway naturally led us to examine the individual aflatoxin genes in the context of how similar or dissimilar the DNA sequences are between different organisms. Similarities between sequences of each gene in each pair of organisms were compared using a k-mer comparison program written by Paul Greenfield (to be published) that generated k-mer similarity matrices. For each gene, this program compares all (overlapping) k-mers from one genome against all (overlapping) k-mers from another genome, producing a count of the number of k-mers shared between the two genomes for that gene. The program tolerates single-base substitution mismatches and only counts each of these differences once, whereas a simple exact-match tiling comparer would accumulate k mismatches for each such difference. As a result of tolerating single-base differences in this way, this program generates similarity numbers close to those produced by NCBI 'megablast' (`http://blast.ncbi.nlm.nih.gov`) when comparing similar sequences, but can do this rapidly for whole draft or finished genomes. The results are presented in Figure 7.6.

6.3.2 Phylogenetic Analysis

It was initially planned to perform phylogenetic analysis of all 25 genes to gain a complete, rounded understanding of the aflatoxin pathway. Due to time constraints, however, it was decided to only study a representative subset of the genes. Figure 7.5 presents the aflatoxin pathway structure in each of the genomes. The subset of genes was chosen such that one gene of each function was present, and the genes were spread out across the pathway. This would add another dimension to the study whereby patterns or correlations between the function and position of the gene in the pathway and its evolutionary dynamics

could be examined. The 7 genes selected to be studied are indicated by the asterisks in Figure 7.5 and are listed here in order along the pathway: *aflT*, *fasB*, *aflR*, *estA*, *avnA*, *omtA* and *moxY*.

The input to the phylogenetic tree construction program is an alignment of the sequences, hence this was the first step in the analysis. The multiple sequence alignment of each gene was carried out using Muscle (Edgar, 2004), which is highly popular as it is fast and produces high quality alignments. The alignments were visualised using CLC Sequence Viewer (`http://www.clcbio.com/products/clc-sequence-viewer/`), and the minor corrections required were manually performed by hand. This program was also used to convert the alignment file from FASTA format to the PHYLIP format as required by the tree building program.

A preliminary stage was incorporated which evaluated whether building a phylogenetic tree for each gene would be feasible, by examining the data for a 'phylogenetic signal', a measure of the ability of the data to generate a 'good' tree in which a consistent amount of evolutionary change is represented on each branch. The signal is considered useful when the amount of evolutionary change should be sufficient yet not too large such that the sequences appear random with respect to each other (M Charleston, personal communication). The strength of the phylogenetic signal present in the data is implicated in the accuracy and robustness of the tree construction (`http://en.wikipedia.org/wiki/Phylogenetics`), thus underlining the importance of testing for phylogenetic signals prior to building the tree itself. The SplitsTree program developed by Huson (1998) implements a method called 'split decomposition' to produce 'splits networks' that offers an effective means of visualising the phylogenetic signals and the 'tree-like' nature of the data, as the underlying technique does not attempt to coerce the data onto a tree (Huson, 1998). In a splits network, each edge corresponds to a bipartition or "split" that divides the taxa, or organisms, into 2 groups on the basis of a single or multiple characteristics (Morrison et al., 2012). Each split is associated with 2 metrics: the amount of support and the amount of conflict. For 'ideal data' in which there are no conflicting phylogenetic signals, the splits will be compatible and appear as single edges in the network, giving rise to structures that resemble trees. Data in which there may be conflicting phylogenetic signals are represented by a series of parallel lines that form less tree-like networks. The lengths of each edge are indicative of the amount of support for the corresponding split. This method complements the similarity matrices for the visualisation of the

data and assessment of whether downstream tree construction will be practical and informative. The splits networks for each of the 7 genes are presented in Section 7.3.2.2.

As Posada explains, statistical inference is essentially the basis of phylogenetic reconstruction. Since statistical inferences can only be made in the context of a probability model, an evolutionary model describing the changes, or substitutions that take place between the four nucleotides (A, C, G, T) in DNA sequences, becomes necessary for characterising the phylogenetic relationships between the organisms of interest (Posada, 2003). These Markov-based models of DNA evolution capture assumptions about the nature of nucleotide substitutions, by describing the different probabilities, or rates of change from one nucleotide (state) to another along the branches of a phylogenetic tree (Posada, 2008). To standardise the time scale in which the evolutionary rates of change are expressed, the models are instead expressed in terms of instantaneous rates of change between the nucleotides using the transition rate matrix Q (`http://en.wikipedia.org/wiki/Models_of_DNA_evolution`). The use of these units eliminates the need to estimate large numbers of parameters for each branch in the tree, as the branch length between an ancestor and a descendant is now a measure of the expected number of nucleotide substitutions that have occurred between these two organisms. Furthermore, the assumption is made that the sites in the alignment evolve independently and are identically distributed, which enables the rates of change to be described using a single transition rate matrix Q as shown in Equation (6.3), where μ_{ij} denotes the rate of substitution between nucleotides i and j (`http://en.wikipedia.org/wiki/Models_of_DNA_evolution`).

$$Q = \begin{pmatrix} -\mu_A & \mu_{GA} & \mu_{CA} & \mu_{TA} \\ \mu_{AG} & -\mu_G & \mu_{CG} & \mu_{TG} \\ \mu_{AC} & \mu_{GC} & -\mu_C & \mu_{TC} \\ \mu_{AT} & \mu_{GT} & \mu_{CT} & -\mu_T \end{pmatrix} \tag{6.3}$$

Over the years, a number of different Markov evolutionary models have been devised, each of which differs in the parameters used to estimate the rates of nucleotide substitution (`http://en.wikipedia.org/wiki/Models_of_DNA_evolution`). Each of these models has its own rate matrix Q which can be expressed using a 6-digit 'substitution code', where each digit corresponds to a different rate parameter. One such model is the HKY85 (Hasegawa et al., 1985) which only has 2 parameters denoted

a and b that can be described using the Q matrix in Equation (6.4). The substitution code is then obtained by horizontally traversing the 6 entries above the diagonal to get $abaaba$ which gets translated to 010010. This code, containing a '0' and a '1', suggests there are two different rate classes. The simplest substitution code is 000000 in which all parameters have the same rate as there is only one rate class. The most complex model, called the Generalised Time-Reversible (GTR) model, has all six parameters in different rate classes and is hence represented by the substitution code 012345.

$$Q = \begin{pmatrix} * & a & b & a \\ a & * & a & b \\ b & a & * & a \\ a & b & a & * \end{pmatrix} \tag{6.4}$$

It is a well known fact that the choice of a substitution model impacts the phylogenetic tree built and thus the analysis performed on it (Posada, 2008). It has been observed that an incorrectly chosen model, or one that poorly fits the data, impacts on the accuracy and reliability of phylogenetic analyses (Posada, 2003). It is therefore imperative that the most suitable model is chosen for each gene.

The evolutionary models explained thus far have only dealt with rates of nucleotide substitution, and there are additional optional components, or parameters, to describe the variation in these rate among the sites (Bazinet, 2013). They are commonly quantified as I which represents the proportion of invariable sites (sites which have zero probability of changing as opposed to sites that have not been observed to differ) and Γ which denotes the rate variation between sites. Both can be specified along with the transition rate matrix Q in defining the evolutionary model and are considered to be biologically valid and realistic models (Felsenstein, 2004). In the Γ model, a Gamma distribution described by the shape parameter α is assumed, from which each site has a rate of substitution drawn at random and independent of other sites (Liò and Goldman, 1998; Felsenstein, 2004). Distributions with large values of α begin to resemble a normal distribution with less variation and similar rates observed across most sites. For small values of $\alpha < 1$, the shape of the distribution moves away from a normal distribution and shows a high degree of variation in rates, with the majority of sites evolving slowly, and the rest very fast (Liò and Goldman, 1998).

JModelTest (Posada, 2008; Darriba et al., 2012) was designed to perform statistical-based model selection to find the most suitable evolutionary model for a given dataset, using different methods such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Decision Theory (DT) and the Likelihood Ratio Test (LRT). While the LRT performs comparisons of two models at a time, the AIC and BIC compare all competing models simultaneously (Posada, 2003) and penalise models with a greater number of parameters. The formulas for the AIC and BIC are given in Equations (6.5) and (6.6) respectively, where n is the length of the sequences, N_i is the number of parameters and L_i is the maximum likelihood in the i-th model.

$$AIC_i \;=\; 2\, log_e\, L_i \;+\; 2\, N_i \tag{6.5}$$

$$BIC_i \;=\; 2\, log_e\, L_i \;+\; N_i\, log_e\, n \tag{6.6}$$

The AIC, BIC and LRT methods generally result in similar models being selected, while the AIC and BIC tend to oversample. Despite this fact however, the BIC is commonly favoured for phylogenetic analyses (M Charleston, personal communication), and hence was the method of choice in this study. The best model is chosen as the one with the lowest negative log likelihood ($-ln\ L$), as the lower this likelihood, the better the model fits the data. Table 7.2 shows the evolutionary model chosen for each of the 7 genes.

With the models of evolution chosen, the maximum likelihood trees for each gene were then inferred using PhyML (Guindon and Gascuel, 2003) with non-parametric bootstrapping using 1000 samplings with replacement to obtain a support value for each branch. 1000 bootstrap replicates is commonly used, with an example of its use by Varga et al. as given in Section 5.3.1. From the similarity matrices (Figure 7.6) and splits networks (Figures 7.7 to 7.13) presented in Chapter 7, species of *A nomius* are clearly distantly related to the other organisms. Hence these were specified as an outgroup by placing an asterisk next to the species names in the input PHYLIP file, to guide the placement of the root during the generation of the trees. Nakamura et al. also performed a similar study with similar organisms in 2011, in which they also placed *A nomius* as an outgroup when constructing the phylogenetic trees (Section 5.3.1).

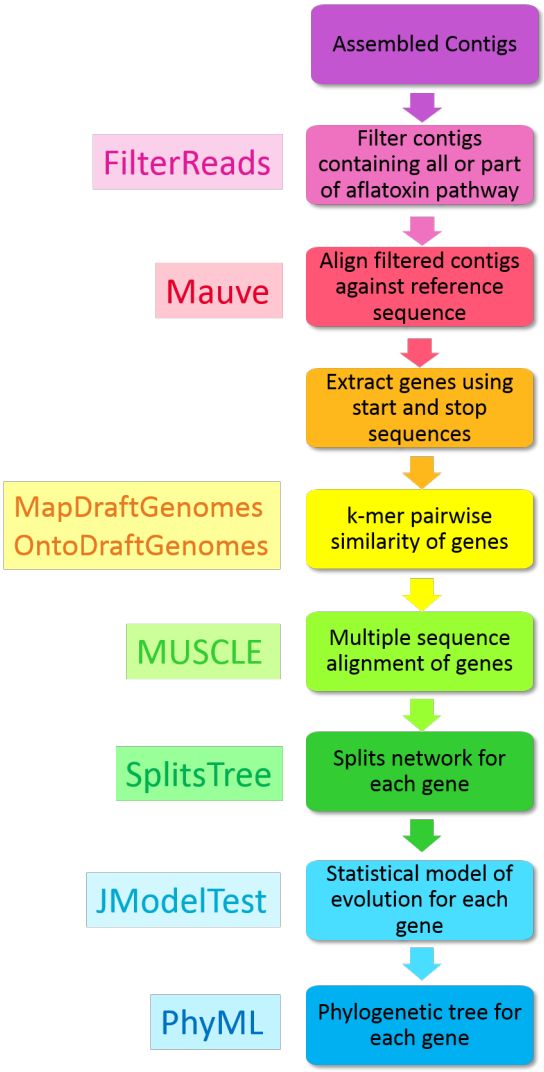


FIGURE 6.9. Workflow for the phylogenetics analysis. The boxes represent which tools were utilised to aid that stage of analysis.

The process from extracting the aflatoxin genes to performing the various facets of phylogenetic analysis is summarised in Figure 6.9.

CHAPTER 7

Results and Discussion

7.1 Overview

This chapter presents and discusses the results obtained for the second part of the study. The first section of this chapter describes and evaluates the results from the fungal genome assemblies (Section 7.2). This is followed by a discussion of each of the results obtained from the various stages in the phylogenetic analysis pipeline to study the evolutionary dynamics in Section 7.3. The concluding chapters outline further work that should be conducted to complement the work done in this study (Section 7.4) and the findings that were made about the evolutionary dynamics of the aflatoxin genes.

7.2 Fungal Genome Assembly

The objective of genome assembly is to take a set of sequencing reads and assemble them into a set of contigs such that the number of gaps in the genome is minimised. Reducing the number of gaps comes from the ability to merge contigs to form longer contigs. The longer the contigs, the fewer number of contigs in total, and thus the higher the quality of the assembly. It should be noted that it is often not necessary or worthwhile to close *all* the gaps and produce one single contig representing the full genome, as this is not only a difficult task, but often the regions of interest such as protein-coding genes and regulatory regions are already contained within the assembled smaller contigs. The set of contigs simply needs to be well-representative of the entire genome and capture all regions of interest. The genome of fungi is divided into 8 chromosomes, so even if a full assembly (with no gaps) was produced, a fungal genome assembly would comprise 8 contigs, where each contig represented one entire chromosome.

7.2.1 Assembly Statistics

The quality of genome assemblies are assessed using quantitative metrics including the number of contigs, the maximum contig length, and the contig N50. The N50 is a statistical measure defined as the length l of the shortest contig such that the collection of contigs of length $\geq l$, when summed together, represents at least 50% of the total assembled genome size (Miller et al., 2010). It can be thought of as analagous to a weighted mean or median, where more weight is given to longer contigs (`http://en.wikipedia.org/wiki/N50_statistic`). Therefore, a smaller number of contigs, larger maximum contig length and larger N50 are associated with a higher quality assembly.

For ease of explanation, we present and discuss the quality of 4 out of the 39 fungal genome assemblies with respect to these statistics. The 4 genomes (from Table 1.1) will be referred to as:

- FGS1 – *Aspergillus flavus* strain 342
- FGS7 – *Aspergillus carbonarius* strain 369
- FGS9 – *Penicillium verrucosum* strain 2940
- FGS14 – *Aspergillus minisclerotigenes* strain 4086

Figure 7.1 graphically illustrates the statistics for assemblies produced from the raw sequence reads, error corrected HiSeq reads, and error corrected HiSeq reads + error corrected MiSeq reads, for k-mer lengths of both 41 and 57. The difference between these assemblies is explained in Section 6.2.3. It is immediately evident that the MaxContigLength and ContigN50 are much higher for FGS1 and FGS14 than for FGS7 and FGS9, in both the $k = 41$ and $k = 57$ assemblies across all three types of assemblies. Likewise, the number of contigs for FGS1 and FGS14 are far fewer than in FGS7 and FGS9. For $k = 41$, the MaxContigLengths range from about 165,000 bp to 300,000 bp for FGS1 and FGS14, while for FGS7 and FGS9 the MaxContigLengths were only between 60,000 and 125,000 bp. ContigN50 values also ranged from about 30,000 to 70,000 bp for FGS1 and FGS14, which are higher than the ContigN50 values for FGS7 and FGS9 which lie between approximately 12,000 bp and 35,000 bp. The genomes of FGS1 and FGS14 could be assembled using between 1,400 and 3,600 contigs, whereas the assemblies of FGS7 and FGS9 required between 4,000 and 7,000 contigs, indicating greater fragmentation and larger proportion of gaps in the latter two genome assemblies.

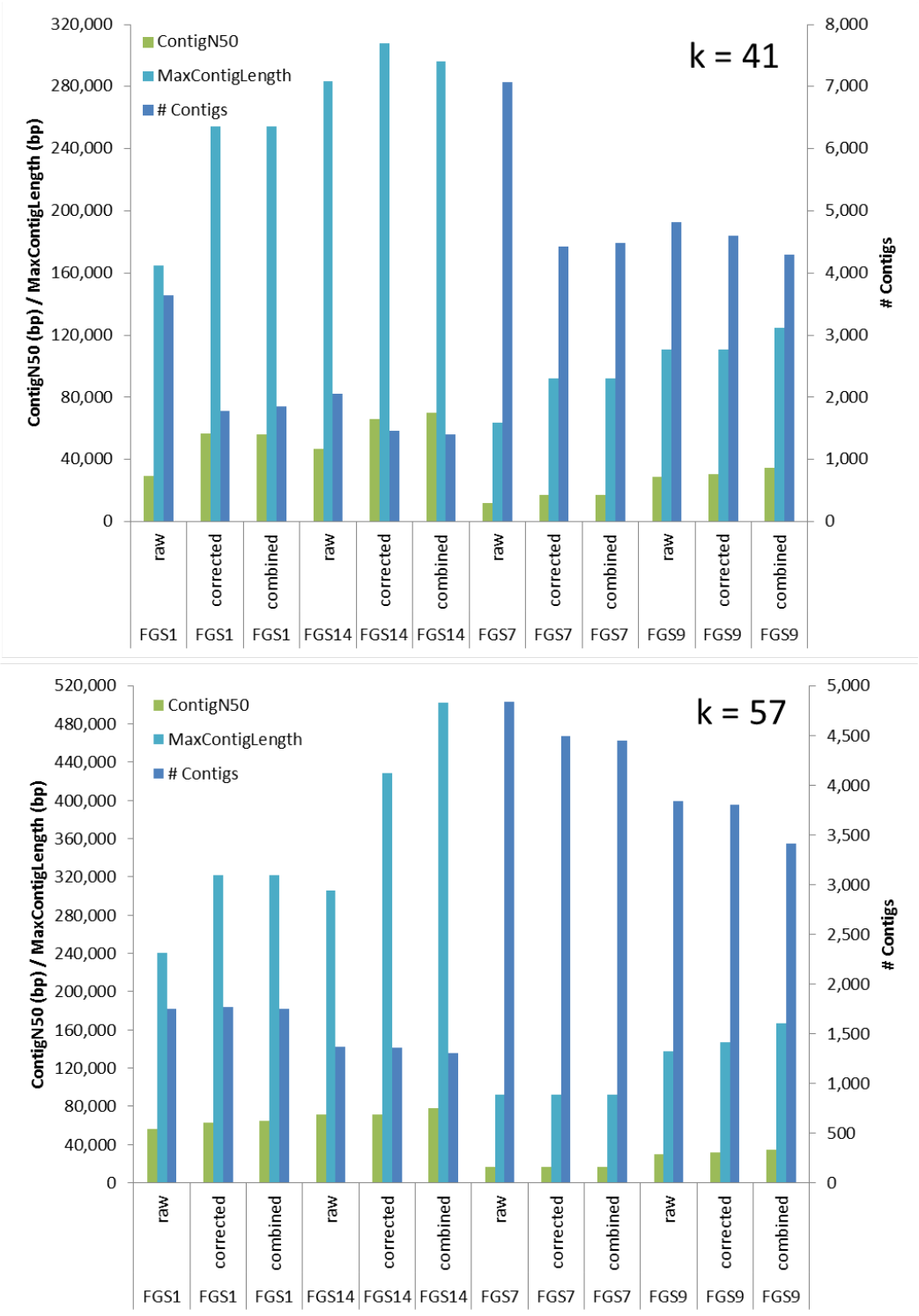


FIGURE 7.1. Assembly Statistics. The top panel shows the results obtained using a k-mer length of 41, the bottom panel shows the results obtained when using a k-mer length of 57. 'raw' refers to assembly using the raw HiSeq reads, 'corrected' refers to the assembly using error corrected short HiSeq reads and 'combined' refers to error corrected short HiSeq reads combined with error corrected long MiSeq reads.

This preliminary inspection of the assemblies produced using a k-mer length of 41 suggests that high quality assemblies were produced for the genomes of FGS1 and FGS14, while the genome assemblies of FGS7 and FGS9 were of a poorer quality. This may be attributed to a greater number of sequencing errors in the reads of FGS7 and FGS9, or the presence of more repeat regions in their genomes which the assembler had difficulty resolving. A similar pattern is observed when the same assemblies were reproduced by increasing the k-mer length to 57, with the MaxContigLengths and ContigN50 values being almost double or triple, and the number of contigs being almost half in FGS1 and FGS14 as compared to FGS7 and FGS9, thus confirming the conclusions drawn.

Assembling the genomes using k-mer lengths of 41 and 57 provided a means to evaluate which value of k resulted in superior assemblies. Each assembly produced using $k = 57$ exhibits a general trend where the MaxContigLength and ContigN50 remain similar or increase, and the number of contigs remains similar or decreases when compared to the corresponding assembly using $k = 41$. The most notable is the increase in MaxContigLength of FGS14 combined from about 300,000 bp using $k = 41$ to 500,000 bp using $k = 57$, and the decrease in the number of contigs of FGS7 raw from 7,000 contigs using $k = 41$ to 4,800 contigs using $k = 57$. This experiment confirms that in general, the longer k-mer length of 57 produces superior assemblies than using a shorter k-mer length of 41.

Comparing the three types of assemblies for each genome, it is observed that the MaxContigLength and ContigN50 increase from the assembly of raw reads to the assemblies of error corrected HiSeq reads and finally to the combined assemblies of error corrected HiSeq and MiSeq reads. Likewise, the combined assembly has fewer contigs compared to the corresponding genome assemblies of raw reads and error corrected HiSeq reads only. The $k = 57$ assemblies of the FGS14 genome clearly display this observation, with MaxContigLength starting at 305,000 bp in the raw assembly, increasing to 428,000 bp in the corrected assembly and further increasing to approximately 502,000 bp in the combined assembly. The ContigN50 increases slightly while the number of contigs displays a slight decrease as we go from the raw to corrected to combined assembly. The results obtained therefore reinforce how error correction and the combination of short and long reads can better resolve large gaps and repetitive regions to greatly enhance the quality of the assemblies produced.

Studies conducted by Greenfield et al. (2013) on assemblies of bacterial genomes further evidences this observation (Figure 7.2). Similar trends are portrayed with the MaxContigLength and ContigN50

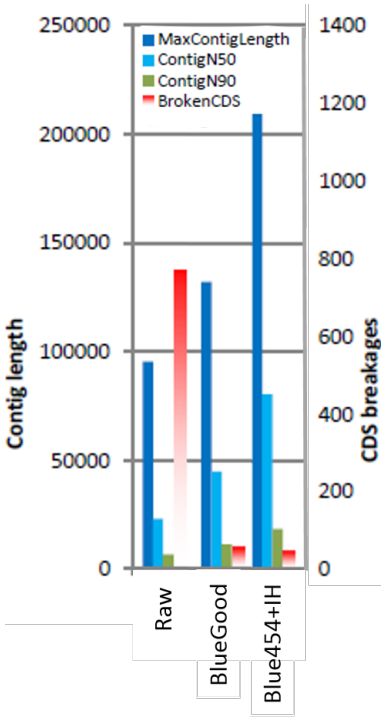


FIGURE 7.2. Bacterial Genome Assembly Statistics. The columns from left to right represent assemblies using raw reads, error corrected HiSeq (short) reads and error corrected HiSeq + error corrected 454 (long) reads. Figure adapted from Greenfield *et al*.

almost doubling from the raw assembly to the assembly of error corrected short HiSeq reads + long 454 reads. Greenfield et al. use an additional statistic called 'BrokenCDS' which counts the number of genes contain gaps where the contigs are broken, which reduces from about 750 in the raw assembly to about 50 in the combined assembly. This is an important metric as a high number of broken or incomplete genes will adversely affect studies focusing on analysing individual genes or a group of genes, such as in the current project (Section 7.3).

The MaxContigLength is not a robust metric for evaluating the quality of genome assemblies, as it may be an outlier while the remaining contigs all have much shorter lengths, and hence it does not provide a true picture of the overall contig lengths. It has also been argued that the ContigN50 statistic is not a highly useful or reliable measure (P Greenfield, personal communication). The N50 is calculated by sorting all the contigs by decreasing length and then calculating the cumulative contig lengths. The value of N50 is the size of the contig where the cumulative length is equal to or greater than half of

the total genome size (sum of all contig lengths). The problem with N50 is illustrated by the following example: suppose the total genome size is 40,000,000 bp, then half of this is 20,000,000 bp. Two of the contigs of successive lengths when sorted are c_i of length 60,000 bp and c_{i+1} of length 55,000 bp. Assume the cumulative contig length from c_1 to c_i is 19,999,999 bp, then the cumulative contig length up to and including c_{i+1} is 19,999,999 + 55,000 = 20,054,999 bp. Half the genome size of 20,000,000 bp is actually closer to the cumulative contig length of 19,999,999 bp that occurs at c_i, so the N50 value should be 60,000 bp. However half the genome size of 20,000,000 bp is only covered by the cumulative contig length at c_{i+1}, hence the N50 is assigned the length of c_{i+1} of 55,000 bp, which is 5,000 bp shorter than what it should be. This problem was observed in our genome assemblies, and demonstrates how these border cases can underestimate the true N50 value and therefore the quality of the assembly.

Figure 7.3 plots the contig number on the x-axis and the cumulative contig lengths along the y-axis. This type of plot is motivated by 'Rarefaction Curves' that are used in ecology as a technique to estimate the number of species in an environmental sample. Once the the number of samples, plotted on the x-axis, plateaus, this indicates the point where taking more samples will not yield many more species that have not already been sampled (`http://en.wikipedia.org/wiki/Rarefaction_(ecology)`). This analogy also applies to contigs; once the cumulative contig length flattens out, the addition of more contigs does not add much to the total genome size. This also reiterates the previously stated claim about not needing to produce a full assembly (by closing all gaps) to obtain a representative set of contigs. From such rarefaction curves, we propose an alternative, robust metric to the N50 statistic as a quality measure, such as the area under the curve or the gradient of the first steep ascent. Higher areas under the curve or steeper gradients are associated with higher quality assemblies. Indeed, Figure 7.3 shows the $k = 41$ raw assemblies for each genome as having the smallest area under the curve and the most flattest ascent, which causes it to require a much higher number of contigs to reach the total genome size as compared to the other assemblies. For FGS1 and FGS9, the difference between the raw assemblies and the corrected and combined assemblies for $k = 41$ is not well-distinguished, however is well pronounced in FGS14 and FGS7. For a given k, the combined assembly in general has the largest area under the curve while the raw assembly has the smallest area. The assemblies using $k = 57$ have a greater area under the curve than the corresponding assemblies using $k = 41$. These findings are consistent with results obtained from assembly statistics graphs in Figure 7.1 above, which serves to further underline the importance of error correction and the combination of short with long reads in

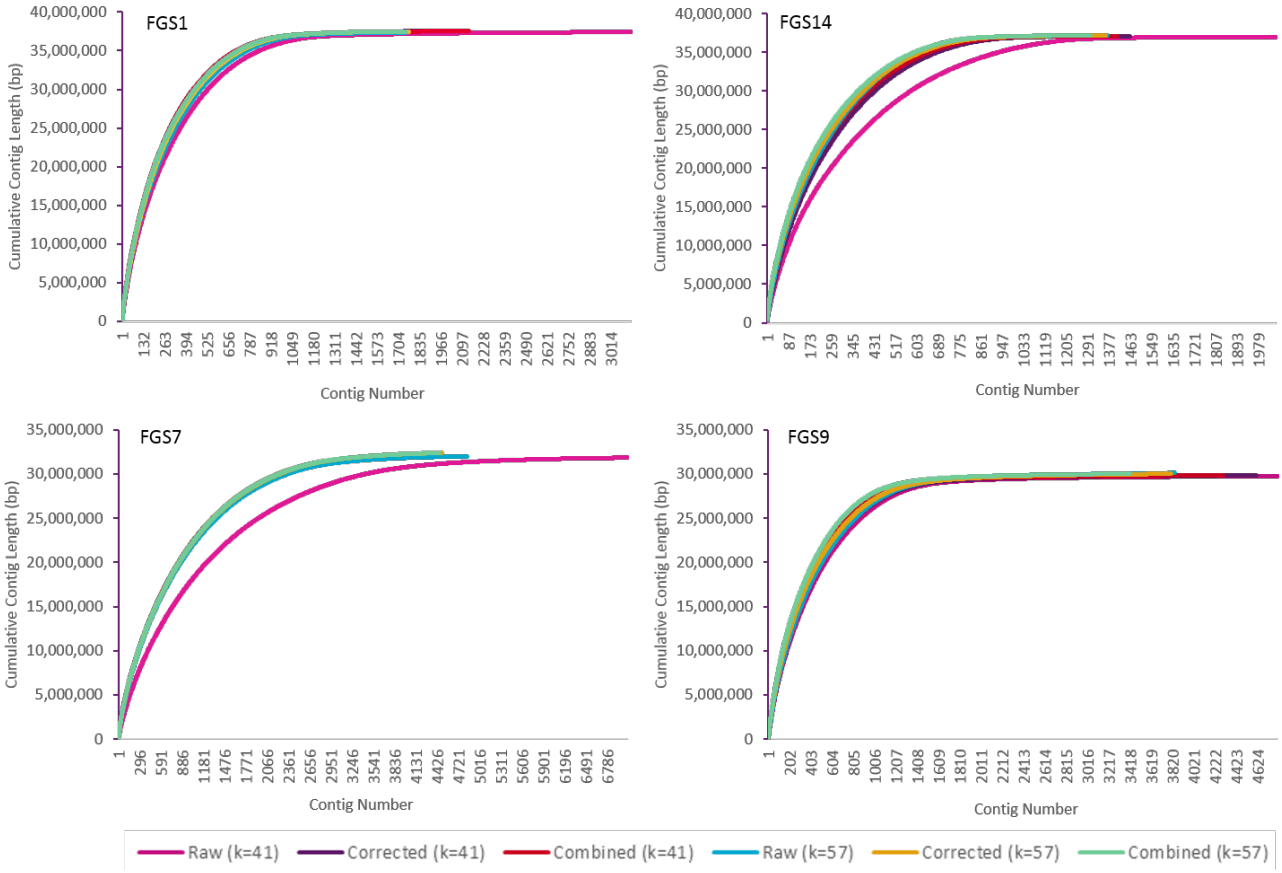


FIGURE 7.3. Rarefaction Curves of all three types of assemblies for each of $k = 41$ and $k = 57$.

producing high quality assemblies. Devising a method to efficiently compute the area under the curve or gradient is out of the scope of this project but is a highly recommended investigation, which would facilitate much easier comparisons and evaluations of the quality of assemblies.

Based on the metrics used, it is evident that the highest quality assemblies were obtained using the combined error corrected HiSeq and MiSeq reads, and a k-mer length of 57, which in general had fewer contigs, a larger N50 and larger maximum contig length. We then compared these assemblies with assemblies of the same species that have been published in literature (Table 7.1).

TABLE 7.1. Comparison of statistics between our assemblies and those published in literature. Assembly statistics for *Aspergillus flavus NRRL3357* were obtained from `http://www.ncbi.nlm.nih.gov/assembly/250208/`, and assembly statistics for *Aspergillus carbonarius ITEM 5010* were obtained from `http://genome.jgi-psf.org/Aspca3/Aspca3.info.html`.

Statistic	FGS1	*A flavus NRRL3357*	FGS7	*A carbonarius ITEM 5010*
# Contigs	1,750	958	4,448	1,346
ContigN50 (bp)	64,596	103,582	16,836	80,000
Total Genome Size (bp)	37,507,693	36,892,344	32,414,634	34,200,000
Genome Coverage	97	5	112	18

For both FGS1 and FGS14, the number of contigs is about 2 times and 3 times than in the corresponding published assemblies respectively. The ContigN50 value of FGS1 of 64,596 bp is only about half the N50 of 103,582 bp in the *A flavus NRRL3357* genome, while the ContigN50 of FGS7 had a value of 16,836 bp which is only about one-fifth of the ContigN50 of 80,000 bp attained in the assembly of *A carbonarius ITEM 5010*. These metrics are indicative of the higher quality of the published assemblies, however, this is due to the published assemblies based on reads from pure Sanger sequencing or a mix of Sanger and 454 sequencing. As previously explained in Section 4.6, Sanger sequencing reads are up to 1100 bp in length, which are much longer compared to the 100 bp Illumina reads obtained for our novel genomes. The longer length and higher quality of the input reads themselves results in the assemblies of these Sanger sequence reads of a greater quality.

From Table 7.1, there is less discrepancy between the total genome sizes with our FGS1 assembled genome being slightly larger than its published counterpart, while our FGS7 genome was slightly smaller than the published *A carbonarius*. In terms of genome coverage, FGS1 has been sequenced with a coverage about 19.5 times greater than the published *A flavus NRRL3357*, and FGS7 had a 6.2 times greater coverage than the published *A carbonarius ITEM 5010*. The deeper sequence coverages are desirable for resolving ambiguities from uncalled bases and for correcting sequencing errors, as explained in Section 6.2.

Thus far, no genome assemblies for *Aspergillus minisclerotigenes* (FGS14) have been reported in literature (D Midgley, personal communication), hence no assembly statistics are available for comparison. Published assembly statistics for a genome of *Penicillium verrucosum* (FGS9) could not be found in literature either. From our assemblies, the total genome size of FGS14, which was sequenced with 132

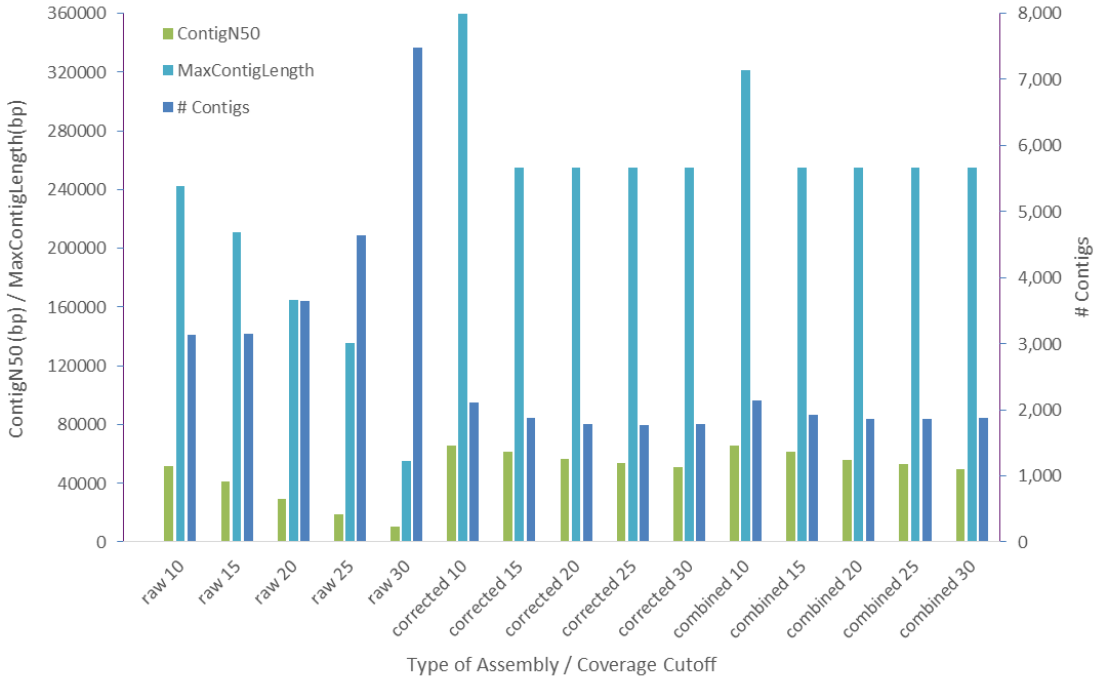


FIGURE 7.4. Assembly Statistics using different coverage cutoff values for each of the three types of assemblies for FGS1 using a k-mer length of 41.

times coverage, is 37,649,034 bp and the total genome size of FGS9 was found to be 30,068,918 bp, which was sequenced with 57 times coverage. Although no comparison with literature can be made, *Aspergillus minisclerotigenes* (FGS14) is known to be a very close relative of *Aspergillus flavus*. The almost identical genome sizes of 37,649,034 bp (FGS14) and 37,507,693 bp of *Aspergillus flavus* (FGS1) is therefore coherent with this observation.

7.2.2 Effects of Coverage Cutoff Parameter

As explained in Section 6.2.2, the coverage cutoff (`cov_cutoff`) parameter is used by Velvet to filter out contigs which have low k-mer coverage depth due to potential sequencing errors. An experiment was conducted to investigate the extent of the effects of this parameter on the quality of assemblies produced. The results obtained from assemblies with coverage cutoffs ranging from 10 to 30, at intervals of 5, for the $k = 41$ FGS1 genome are presented in Figure 7.4.

For the assemblies using the raw reads, the MaxContigLength, ContigN50 and the number of contigs varies significantly as the coverage cutoff increases. The MaxContigLength decreases from 242,616 bp at a coverage cutoff of 10 to 54,875 bp at a coverage cutoff of 30, while the number of contigs almost doubles from 3,140 to 7,472. The ContigN50 also decreases 5-fold from 51,888 bp at a coverage cutoff of 10 to 10,407 bp at a coverage cutoff of 30. The values of these metrics remain fairly constant in the corrected and combined assemblies, with a general decrease exhibited for all 3 metrics as the coverage cutoff increases from 10 to 30.

It was expected that as the minimum coverage threshold increases, the quality of the contigs should increase. The patterns observed, however, display the contrasting result. This can be explained as higher coverage cutoffs results in Velvet discarding a greater number of poor coverage contigs, however some of these may be required in the assembly to join contigs and close gaps. Because these contigs are no longer available, there are more gaps that cannot be resolved, and hence there are a larger number of separate contigs that cause lower MaxContigLength and ContigN50 values, as observed. Therefore, the coverage cutoff parameter highly influences the quality of the resulting contigs and an appropriate value must be determined depending on the type and nature of the sequence reads that are to be assembled.

7.3 Evolutionary Dynamics of Aflatoxin Genes

The assemblies of combined error corrected HiSeq and MiSeq reads using a k-mer length of 57 (Section 7.2) were judged to produce the highest quality assemblies. The resulting assembly contigs of 17 of the 39 sequenced genomes known to contain the aflatoxin genes (5 *Aspergillus flavus*, 3 *Aspergillus oryzae*, 3 *Aspergillus parasiticus*, 3 *Aspergillus minisclerotigenes* and 3 *Aspergillus nomius*) were used for the *in silico* aflatoxin gene phylogenetic analysis presented in this section.

7.3.1 Characterisation of Aflatoxin Biosynthesis Pathway

Prior to performing analysis of the aflatoxin genes, preliminary investigations were conducted to elucidate the structure of the aflatoxin pathway in terms of what genes are present, and whether there are any obvious structural differences in the pathway between the organisms being studied.

7.3.1.1 Structure of Aflatoxin Gene Pathway

As mentioned in Section 5.2, aflatoxins are produced from an intricate metabolic pathway consisting of 25 genes (Yu et al., 2004). The complexity of the pathway is not only evident from the large number of genes involved in the production of aflatoxin, but the diversity in the functions of the genes involved (Figure 7.5).

The schematic in Figure 7.5 shows the order of genes in each of the novel genomes. Note that there are intergenic regions separating each of the genes; the genes are shown as contiguous in the diagram for simplicity.

There are no structural rearrangements of the aflatoxin pathway between the sequenced organisms, that is, the order of the genes has been conserved across all organisms, and is consistent with aflatoxin pathway structures shown in Figure 5.1 as well as others reported in literature (Yu et al., 2004). Indeed, the 3 strains of *A flavus-oryzae*, which are non-toxigenic, contain the full aflatoxin pathway that is identical to their toxigenic *A flavus* variants, as was established in the literature review in Section 5.2.

While there are no rearrangements, there are however large deletions in the aflatoxin pathways of FGS1 and FGS20 that have resulted in almost half to three-quarters of the entire pathway going missing in these *A flavus* genomes. In FGS1, the deletion spans the region from *norB* to *norA*, while in FGS20 the deletion is larger, spanning the *norB* to *omtA* genes. Since both these deletions span the *aflR* transcription factor gene which is required to activate the pathway, it is expected that these two strains of *A flavus* cannot produce aflatoxin.

A BLAST search (`http://blast.ncbi.nlm.nih.gov/`) of the contig in the FGS1 genome containing the aflatoxin genes was conducted, and the top match obtained is an *A flavus* strain V5F-13 (GenBank accession number: JQ435497) that contains a translocation breakpoint junction before the *ver-1* gene. The 100% identity and E-value of 0.0 provide strong evidence that this match is true and not a result of random chance, as explained in Section 2.4.1. A similar translocation event is expected to have occurred in the FGS20 genome.

A chromosomal translocation is a genetic event resulting in the exchange of genetic material, which can be balanced or unbalanced (`http://en.wikipedia.org/wiki/Chromosomal_translocation`).

The exchange of material in a balanced translocation is equal, hence there is no loss of genetic information. In contrast, the exchange of material is unequal in an unbalanced translocation, which can lead to missing or additional genes. Under the hypothesis that the aflatoxin region in FGS1 and FGS20 have undergone balanced translocation events, the missing *norB* to *norA* genes would be relocated to another region of the respective genomes, perhaps to a different chromosome, and would have still been found by the methods used to extract the genes (Section 6.3.1). The fact that the *norB* to *norA* genes could not be found anywhere in the FGS1 or FGS20 genomes is suggestive of an unbalanced translocation event where the genes have been completely removed from the genomes. The region upstream of the genes in the two genomes, that is, corresponding to the region where the rest of the aflatoxin genes should have occurred if not missing, were searched against the reference sequence of *A flavus* strain NRRL 3357 using BLAST (`http://blast.ncbi.nlm.nih.gov/`). Genes that were matched with an E-value of 0.0 included a choline dehydrogenase, ceramidase and gulonolactone oxidase for FGS1, and a mitochondrion biogenesis protein and *Sat1* intracellular transport protein for FGS20. These genes do not correspond to any of the missing aflatoxin genes that occur upstream in the pathway, thus providing more support in favour of an unbalanced translocation event in both these genomes.

Similar findings of a 40,000 bp deletion between the *norB* and *norA* genes have also been reported in literature as a cause of the inability to produce aflatoxins (Gibbons et al., 2012). Furthermore, the location of the aflatoxin biosynthesis pathway genes in the telomeric region, or towards the end, of chromosome 3 may have a positional effect whereby this region is susceptible to mechanisms causing gene loss, recombination and other genomic rearrangements such as inversions, deletions and translocations (Carbone et al., 2007). Experimental procedures should be conducted to verify that these strains of *A flavus* are indeed non-toxigenic.

It was previously stated during the discussion of the genome assembly results that contig breaks or gaps do not generally occur in regions such as genes (Section 7.2.1). In all 3 *A parasiticus* genomes (FGS5, FGS6, FGS37) and 2 *A minisclerotigenes* genomes (FGS14, FGS18), the entire 80,000 bp aflatoxin region is contained within a single contig. In all other genomes however, the pathway has been broken across a number of contigs, as indicated by the vertical bars in Figure 7.5. Most of these contig breaks occur in intergenic regions between genes, and the majority appear to be localised between the *fasB* and *aflR* genes, the *aflJ* and *adhA* genes, and between the *avfA* and *omtB* genes. The locations of these

breaks occur in regions of single nucleotide repeats, for example 'CCCCCCC…'. In challenge to the claim made, there are contig breaks that occur within the genes *cypA* in FGS3, *aflJ* in FGS2, AF4351 and FGS11, and *hypA* in FGS1 and FGS20. The ends of the two contigs covering the *aflJ* gene in FGS2 were examined, and a large repeat region of the nucleotide 'A' adjoined by regions of unbalanced coverage (caused by the presence of 'adapters' that bind to the DNA in the Illumina sequencing process) was observed. Due to these repeated regions of 'A's, the Velvet assembler was unable to determine how to join the contigs, and instead generated a break in the contigs. It is expected such repeat regions also caused the contig breaks that occurred within the *cypA* and *hypA* genes.

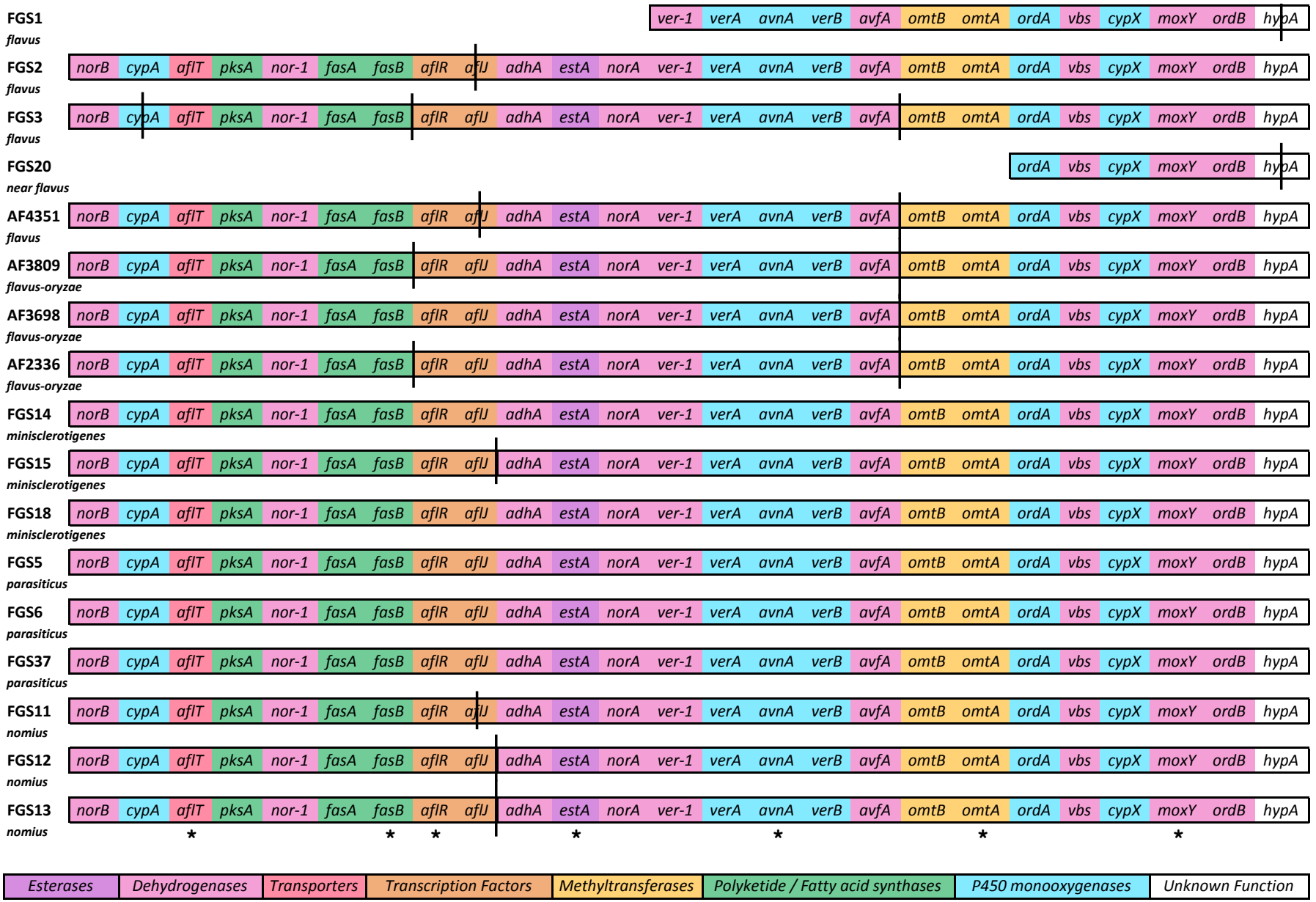


FIGURE 7.5. Structure of the aflatoxin gene pathway in our sequenced fungal genomes. The genes are coloured according to their function, which is given by the legend at the bottom. Vertical bars correspond to where contig breaks have occurred. Asterisks indicate which genes were selected for phylogenetic analysis.

Another finding obtained from the characterisation of the aflatoxin pathways is that the first 2 genes of the pathway, *norB* and *cypA*, contain deletions in the genomes of *A flavus* (FGS2, FGS3, AF4351), *A flavus-oryzae* (AF3809, AF3698, AF2336) and *A minisclerotigenes* (FGS18). These deletions are approximately 500 – 800 bp in length and occur at the end of the *norB* genes, and 80 – 1200 bp deletions that occur at the beginning of the *cypA* genes. The lengths of the deletions are similar to those reported by Ehrlich et al. in 2004, who experimentally confirmed an 800 – 1500 bp deletion spanning the *norB* and *cypA* genes in *A flavus*. Their experiments were the first to show that the *cypA* gene is required for the production of G-type aflatoxins (Section 5.2.1). All the *A flavus* and *A flavus-oryzae* genomes in this study are either missing or contain a partial *cypA* gene, which is consistent with Ehrlich et al.'s findings and evidences their inability to produce aflatoxin G (Table 5.1). Unlike *A flavus* that only produces B-type aflatoxins, a distinguishing feature of its close relative *A minisclerotigenes* is its ability to produce both B- and G-type aflatoxins. It is therefore surprising and interesting that one of the *A minisclerotigenes* genomes (FGS18) contains the same deletions in *norB* and *cypA* that are observed in the *A flavus* genomes, thus questioning the ability of this particular strain to produce G-type aflatoxins as previously believed. Experimental analyses in the laboratory to complement this study should be conducted for validation of the findings obtained.

7.3.1.2 Pairwise Similarity between Aflatoxin Genes

The sequences of each gene from each organism were compared by calculating the number of k-mers that are shared, or similar, between each pair of organisms. The notion of 'similarity' was defined in Section 6.3.1. The number of shared k-mers were converted to percentages and are represented in the form of a 'similarity matrix' for each gene, where each row and column corresponds to an organism, that are ordered as: *A flavus*, *A flavus-oryzae*, *A minisclerotigenes*, *A parasiticus* and *A nomius*. The percentages values are coloured to cover the range of 48% (red) to 100% (green), which was decided based upon the lowest similarity of 48.7% found across all the genes. As each diagonal entry is comparing the pairwise similarity of an organism with itself, the number of shared k-mers is 100% and hence we see a pattern of green unit boxes along the diagonal of the similarity matrix for each gene. The similarity matrices for all 25 gene are presented in Figure 7.6. The original file containing all similarity matrices with actual percentage values is included in the Supplementary Files folder submitted with this thesis.

The generation of these coloured similarity matrices offers an effective visual means for detecting the presence of phylogenetic signals in the data, thereby complementing the analysis performed using the splits networks (Section 7.3.2.2), which forms the subsequent stage in the analysis pipeline. The patterns of the matrices changes drastically as we move along the pathway. Using manual inspection, we have assigned the genes into groups or clusters upon the basis of the pattern of the matrix, which are indicated by the symbols for each matrix. The constituent genes of each cluster appear to be localised to particular regions of the aflatoxin pathway, however because these clusters were formed based on observations, this may be subjective and vary from person to person. More sophisticated and robust techniques should be devised to form these gene clusters. Regardless, the group of genes from *pksA* to *adhA* which are located towards the beginning of the pathway, appear to be clustered into one group as indicated by the blue circles. This similarity pattern is characterised by two reasonably solid green square boxes; the large box in the top left corner represents the group of *A flavus*, *A flavus-oryzae*, *A minisclerotigenes* and *A parasiticus* and the small bottom right hand corner represents four species of *A nomius*. The gene sequences between these two groups are highly different, but are highly similar within each group, with similarity values mostly greater than 90%.

The cluster labelled with the violet diamond, containing the genes *avnA* to *omtB* and *ordA*, is characterised by the four large dark green boxes that occur along the diagonal. These boxes correspond to the *A flavus / A flavus-oryzae*, *A minisclerotigenes*, *A parasiticus* and *A nomius* species respectively. This suggests that the sequences of these genes are quite similar within species, but highly variable between species.

Towards the end of the pathway, the genes that form the cluster labelled with the pink triangle, that is, *vbs* and *moxY* to *hypA*, display a different pattern in their similarity matrices. There is effectively one large top left hand corner box and one small bottom right hand corner box as observed in the cluster containing the genes *pksA* to *adhA*. Unlike this cluster however, the top left box representing all non-*A nomius* species, appears broken and patchy rather than as a solid green box. This pattern implies that for these genes, there is high divergence in the sequences even within the same species.

There are common patterns in the similarity matrices for all genes in all clusters. Each similarity matrix has a small green box in the bottom right corner that comprises of four *A nomius* strains, which are clearly separated from the rest of the organisms. Within this box, the first row and column (shown by

the label 'FGS11' in the *aflJ* matrix), corresponding to the FGS11 genome, appears to be divergent from the other three *A nomius* strains in this group which are almost 100% similar to each other for all genes. All similarity matrices also show that *A flavus* strain BN008 (shown by the label 'BN008' in the *aflJ* matrix) is quite different from its *A flavus* and *A flavus-oryzae* relatives as it causes a 'cross-hairs'-like break in the otherwise contiguous green boxes that represents this group of species. The correct identity of the BN008 species is not clear, as it most similar to *A parasiticus* in the *vbs* similarity matrix, but most closest to *A minisclerotigenes* in the *cypX* similarity matrix. This anomaly will be better explained by the phylogenetic analysis conducted in following sections.

One of the aims of this study into the evolutionary dynamics was to enhance our understanding of the nature of the relationship between *A flavus* and *A flavus-oryzae* (Section 5.3.2). In all similarity matrices, the strains of *A flavus-oryzae* are highly similar in their gene sequences to the strains of *A flavus*, with close to 100% similarities. Within the *A flavus / A flavus-oryzae* clade, all the *A flavus-oryzae* strains appear to be most similar to the other *A flavus-oryzae* strains. These initial investigations suggest that *A oryzae* is not a distinct species from *A flavus*, rather it is a group of organisms within the *A flavus* species. It is expected from the construction of the splits networks and phylogenetic trees in the subsequent sections, that more insights will be gained either in favour or disagreement with these observations.

We now focus our attention to the similarity matrices of the *norB* and *cypA* genes, the first two genes in the aflatoxin pathway. The patterns of the matrices for these two genes are yet again different to the other genes, and thus have been placed into a separate cluster. The top left hand box containing *A flavus* and *A flavus-oryzae* do not appear similar to the sequences of any other species. Similarly, *A minisclerotigenes* strain 3382 (shown by the label 'FGS18' in the *norB* and *cypA* matrices) appears divergent from the all other species, including the other two *A minisclerotigenes* strains that are its closest relatives of the organisms in this study. This provides further verification of the observation made in Section 7.3.1.1 where all the *A flavus* and *A flavus-oryzae*, and FGS18 are either missing or contain an incomplete *norB* and *cypA* gene, suggesting their inability to produce G-type aflatoxins due to this deletion.

It is important to note that these similarity calculations check for approximate matching of k-mers rather than exact matching. This is essential for better comparison of the sequences to allow for changes that can occur between sequences, such as substitutions, insertions and deletions of nucleotides. Although

the method used currently handles substitutions only, it is expected that after insertions and deletions are accommodated into the calculation of shared k-mers, the overall patterns and clusters of the genes based on the similarities will still be maintained. The percentage similarity values presented here are aligned with the values reported by Ehrlich et al. (2005), who calculated the percentage identities (exact matching) for each gene between *A flavus* and *A parasiticus*, and *A flavus* and *A nomius*, and found that the genes between *A flavus* and *A parasiticus* are much more similar than with *A nomius*.

FIGURE 7.6. Pairwise k-mer similarity matrices of each of the 25 genes in the aflatoxin pathway, ordered from left to right in each row. Coloured from low similarity (red) to high similarity (green). The symbols associated with each gene represents the group or cluster based on the pattern of the similarity matrices.

7.3.2 Phylogenetic Analysis

The characterisation of the overall structure of the aflatoxin pathway in the newly sequenced genomes in the previous Section 7.3.1 now enables us to further the analysis one level deeper to study the evolutionary dynamics of the individual genes themselves. This analysis is facilitated by the generation of phylogenetic trees (Section 7.3.2.4), prior to which preliminary analysis was conducted to assess whether the building of phylogenetic trees is feasible (Sections 7.3.2.1 and 7.3.2.2), and if so, which evolutionary models the trees should be based on (Section 7.3.2.3).

7.3.2.1 Multiple Sequence Alignment

The extracted sequences of each gene from all the organisms, including genes from the 7 reference aflatoxin pathway sequences, were aligned using Muscle (Edgar, 2004) as explained in Section 6.3.2. Careful inspection informed that the multiple sequence alignments for each gene were of a high quality, requiring only minor corrections that were performed manually. The original files containing the alignments for each gene are included in the Supplementary Files folder submitted with this thesis.

7.3.2.2 Splits Network

In order to construct a meaningful phylogenetic tree, the 'tree-like' nature of the data needs to be assessed. This can be achieved by producing a splits network as a means of visualising the phylogenetic signals in the data, which provide an indication of the accuracy of the trees generated. As explained in Section 6.3.2, the SplitsTree program (Huson, 1998) was used to generate splits networks for each of the 7 genes, which are presented in Figures 7.7 through 7.13 in the order they occur in the pathway.

Upon visualising the 7 splits networks, it is immediately evident that the *fasB* gene has strong, non-conflicting phylogenetic signals as evidenced by each split being almost perfectly represented by a single edge (Figure 7.8). The *fasB* network, which most closely resembles a tree, shows 2 separate clades or monophyletic groups in which the constituent species have evolved from a common ancestor. These clades can be thought of as comprising of *A nomius* species and non-*A nomius* species. This pattern observed agrees with the pattern in the similarity matrix of the *fasB* gene (Figure 7.6), in which there are essentially 2 groups as indicated by the two green boxes; the top left hand square consisting of *A*

flavus, A oryzae, A minisclerotigenes and *A parasiticus*, and the bottom right hand square consisting of *A nomius* strains. Similarly, the tree-like nature of the network for *avnA*, which has few conflicts, can be inferred from the majority of parallel edges that overlap to form single edges (Figure 7.11) in a tree-like manner. In this network there are 4 distinct clades which coincide with the 4 green boxes in the *avnA* similarity matrix (Figure 7.6), further strengthening the phylogenetic signal of this gene.

While the *fasB* and *avnA* genes display the strongest phylogenetic signals from their tree-like networks, the networks of all other genes shows some conflict in the clade containing the non-*A nomius* species, as illustrated by separated parallel lines that form parallelograms, signifying the lack of resolution and incompatibility of the splits. These contradictory patterns are most evident in the *omtA* and *moxY* networks (Figure 7.12 and Figure 7.13 respectively), which is in accordance with the patchy and dispersed top left hand corner region in their respective similarity matrices (Figure 7.6).

A comparison of the 7 networks reveals features that are common across all networks. Like in all similarity matrices where the strains of *A nomius* form a separate green box in the bottom right hand corner (Figure 7.6), these *A nomius* strains also form their own clade in all the networks, which are distantly separated from all other *Aspergillus* species by almost single edges. The high compatibility of these splits coupled with the long edges shows strong support for this separation in the networks of all genes. Furthermore, just as the *A flavus* strain BN008 does not appear to be similar to any of the other *A flavus* and *A oryzae* species in the similarity matrices for all genes, this strain is not a member of the *Aspergillus flavus-oryzae* clades, nor is it a close relative in any of the splits networks, reinforcing its high divergence in all genes.

Furthermore, all strains of *A flavus-oryzae* are grouped together with the *A flavus* species, rather than on their own separate branches. This again supports the observation made in the similarity matrices (Figure 7.6) that the *A flavus-oryzae* strains possibly form a sub-clade of *A flavus* rather than being a separate species.

Therefore, despite the presence of some conflict in phylogenetic signal in some of the networks, most networks do form tree-like structures due to the presence of distinct clades, hence the feasibility of generating phylogenetic trees from this data has been established.

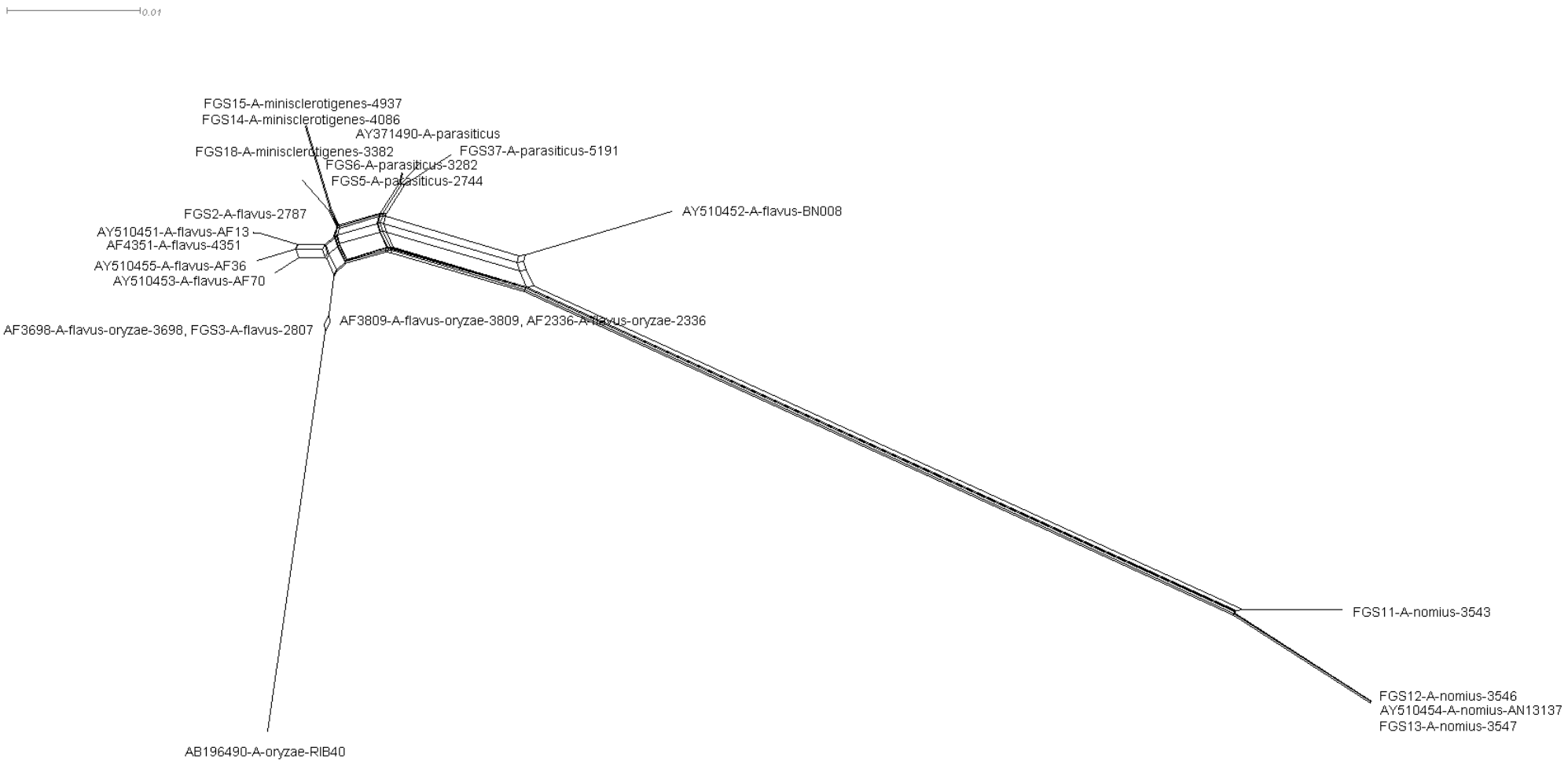


FIGURE 7.7. Splits Network for the *aflT* gene generated using SplitsTree.

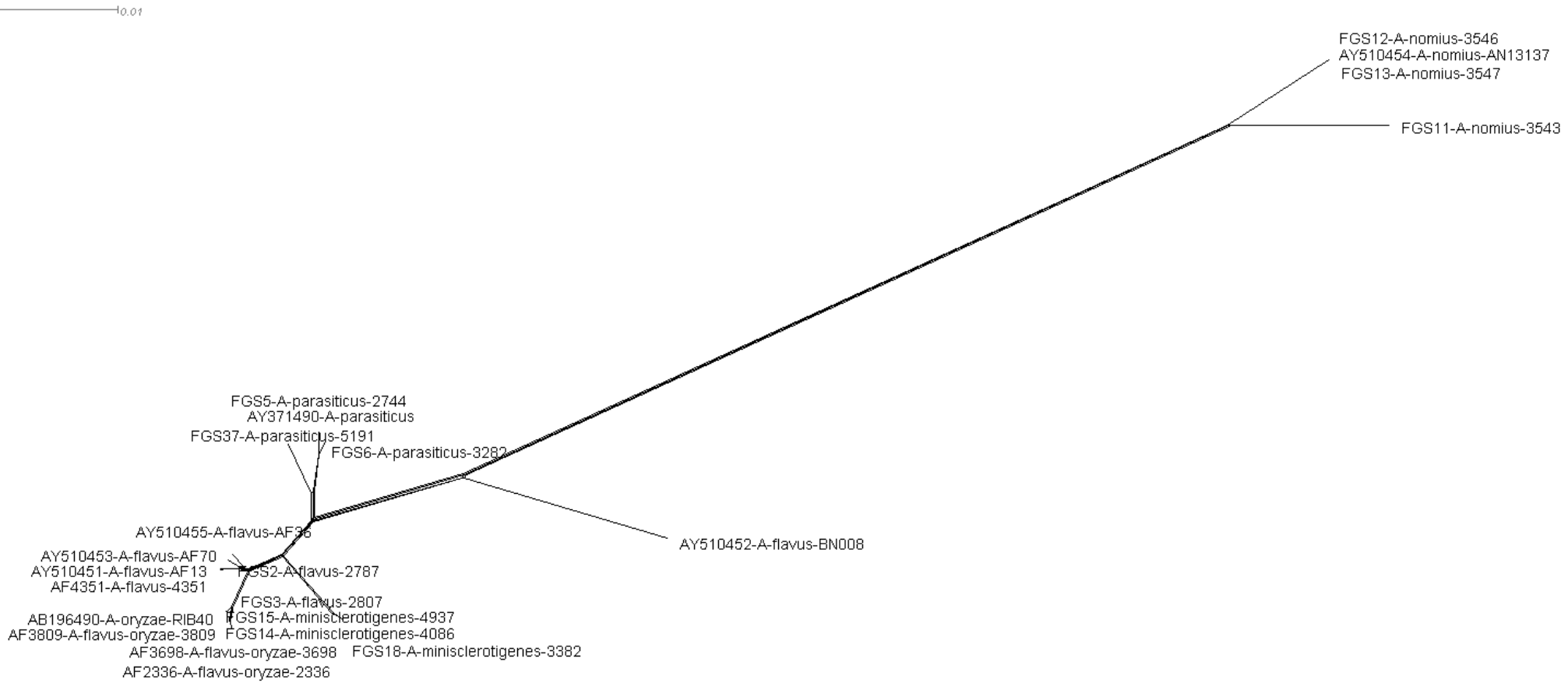


FIGURE 7.8. Splits Network for the *fasB* gene generated using SplitsTree.

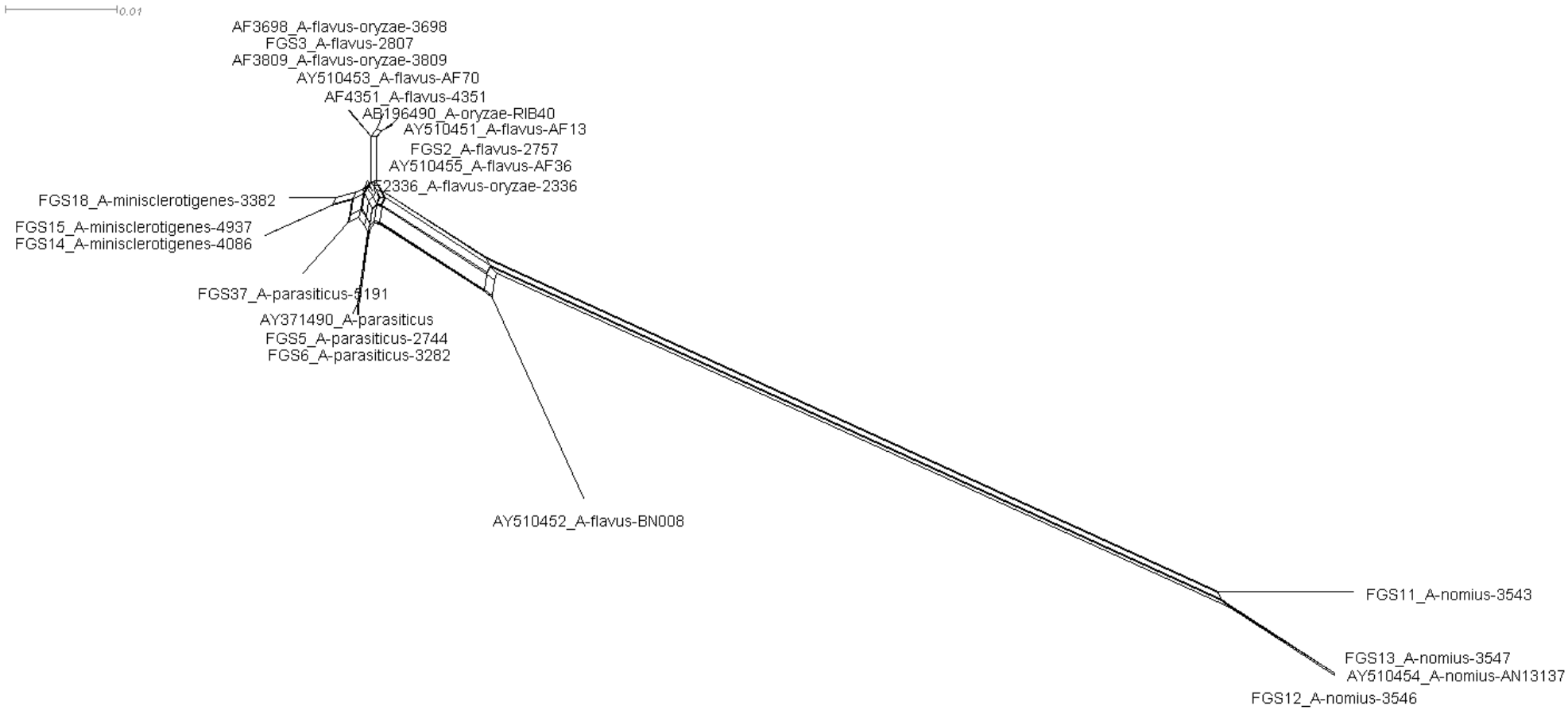


FIGURE 7.9. Splits Network for the *aflR* gene generated using SplitsTree.

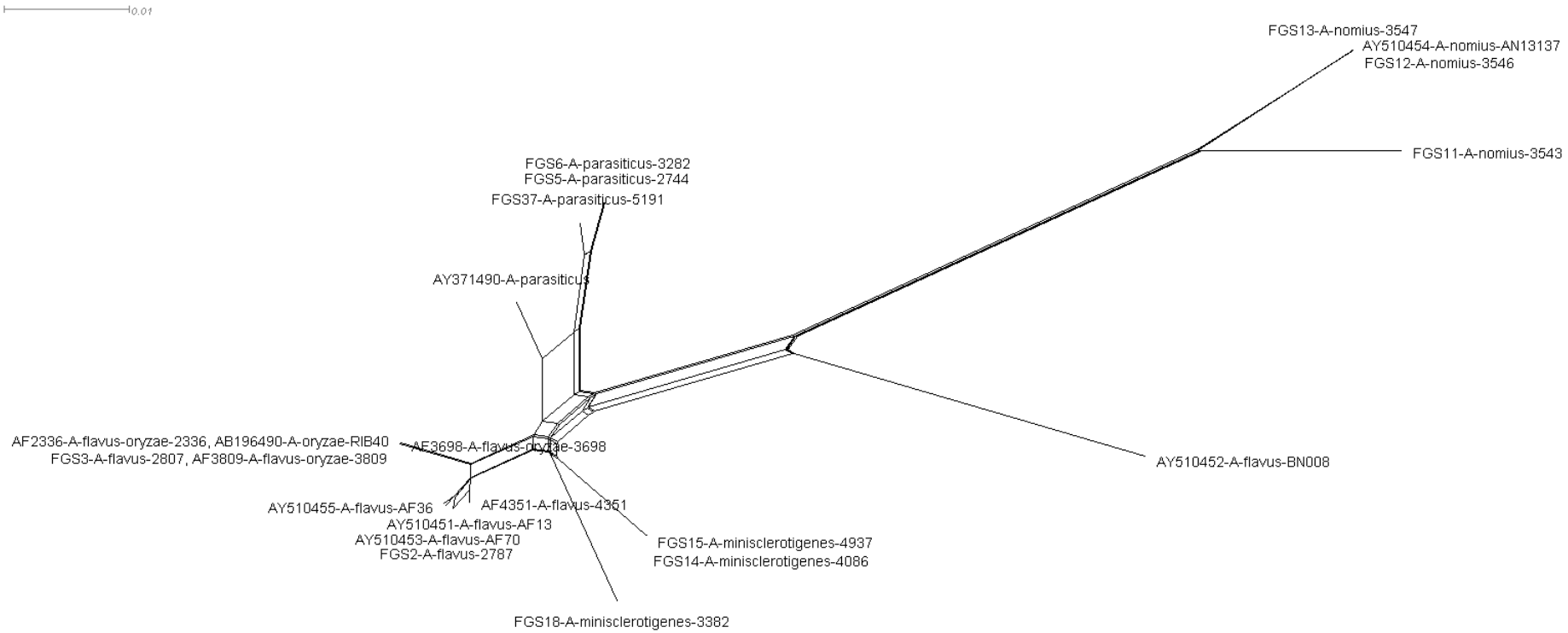


FIGURE 7.10. Splits Network for the *estA* gene generated using SplitsTree.

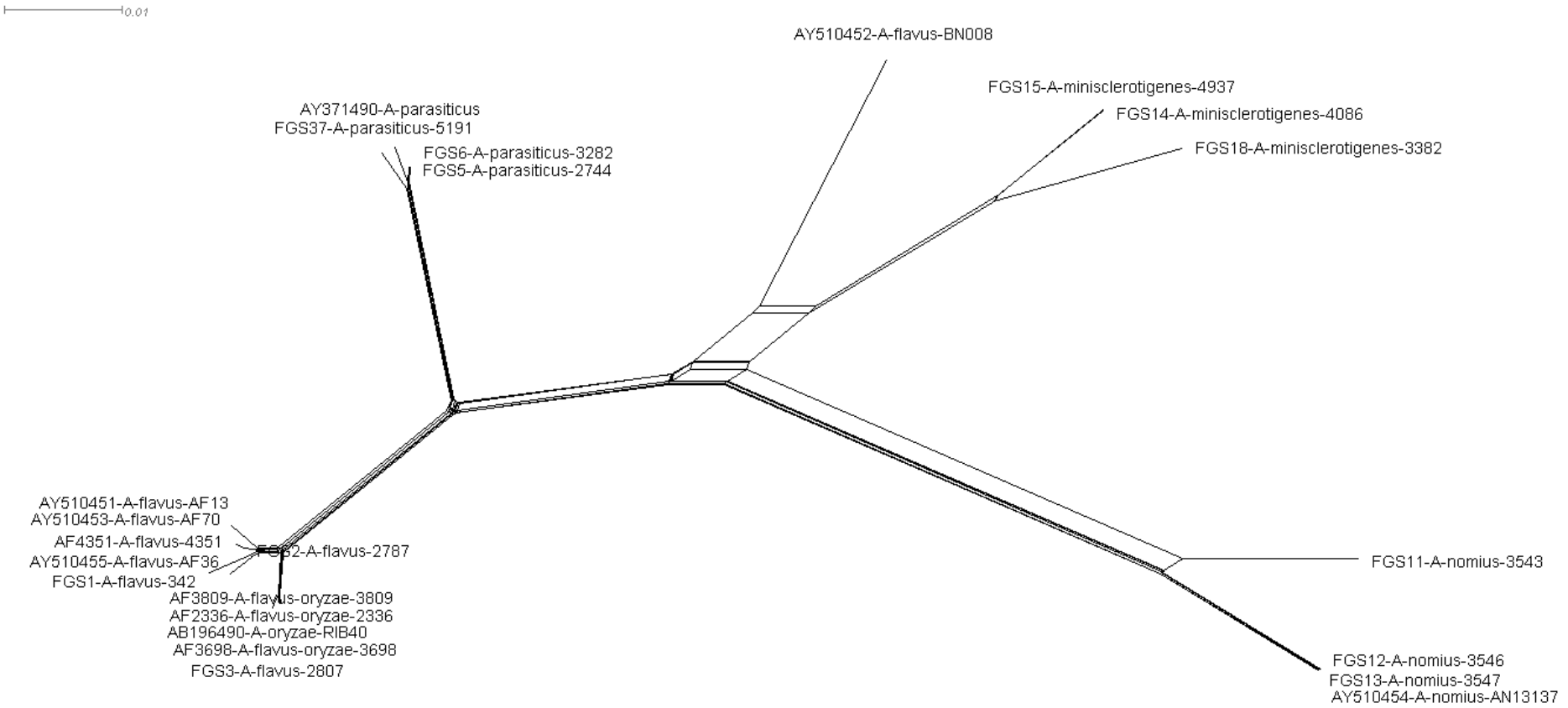


FIGURE 7.11. Splits Network for the *avnA* gene generated using SplitsTree.

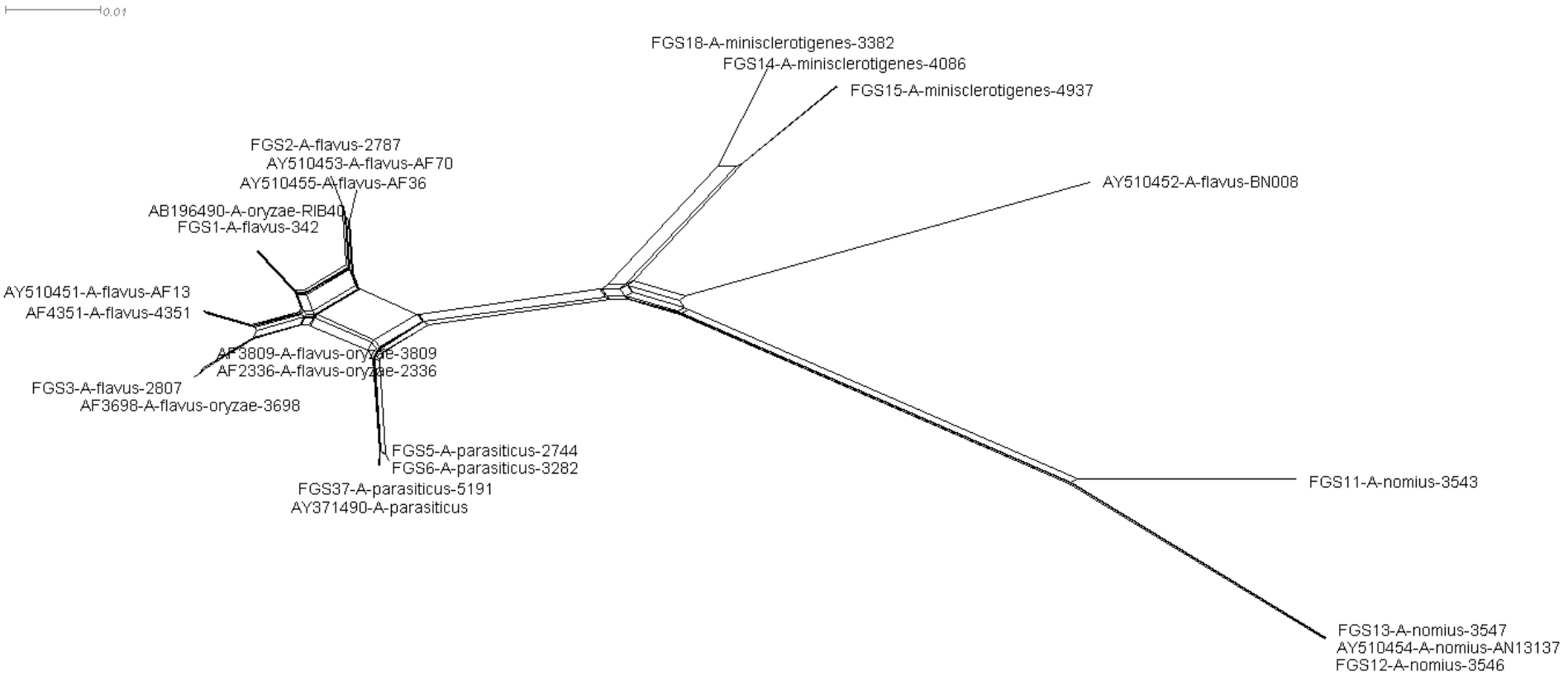


FIGURE 7.12. Splits Network for the *omtA* gene generated using SplitsTree.

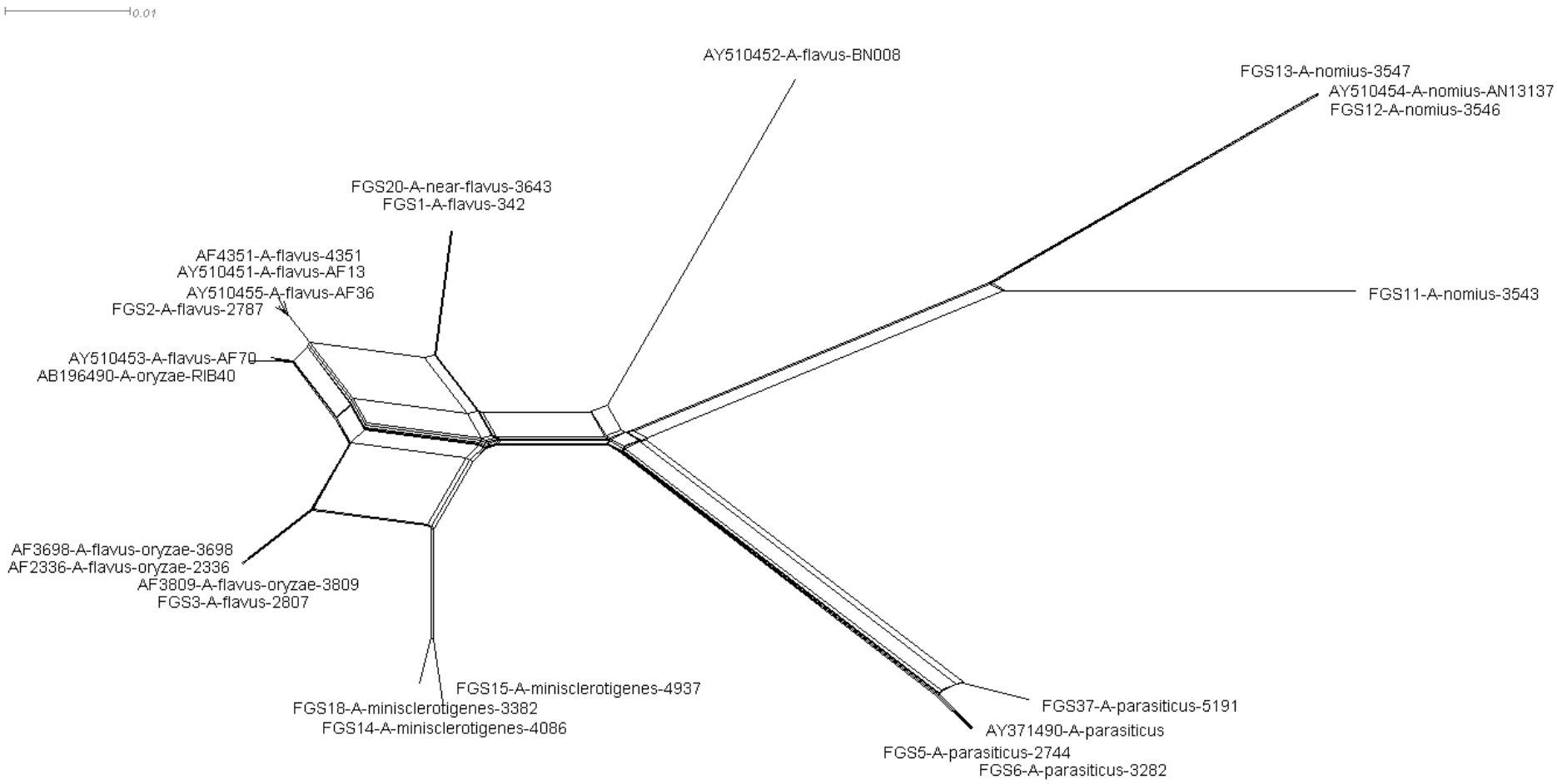


FIGURE 7.13. Splits Network for the *moxY* gene generated using SplitsTree.

7.3.2.3 Statistical Models of Evolution

The investigation carried out in Section 7.3.2.2 established the general 'tree-like' nature of the sequences for each gene, providing assurance that the construction of phylogenetic trees will be feasible. The trees generated in this study are based on Maximum Likelihood (ML) algorithms as described in Sections 5.3 and 6.3.2, which require a model of evolution to be specified to describe the patterns of DNA base substitution (Bazinet, 2013). The underlying statistical model can impact the topology of the tree that is generated (Posada, 2008), therefore impacting analyses and conclusions drawn from it. Table 7.2 summarises the models, along with the associated parameters, chosen by the JModelTest program (Posada, 2008; Darriba et al., 2012) as the best fit model for each gene.

All models chosen quite similar and are relatively simple, containing only a few parameters. While it has been observed that complex models better fit the data than simpler models, complex models require the estimation of a larger number of parameters from the same amount of data, thus incorporating more errors into each estimate, and are also more computationally expensive (Posada, 2003). It is therefore recommended to only incorporate as much complexity as is deemed necessary into the model.

Of the selected models, the simplest model is the K80 model (Kimura, 1980), which was chosen for the *avnA* and *omtA* genes. This model consists of only 2 parameters for the 2 types of substitutions that can occur: one that measures the transition rate (changes from nucleotides $A \leftrightarrow G$ or $C \leftrightarrow T$) and the other measures the transversion rate (A or $G \leftrightarrow C$ or T) (Felsenstein, 2004). This model also assumes that each of the bases A, C, G and T are equally frequent, hence the frequency of these nucleotides shown in the columns 'freq (A)' to 'freq (T)' are each 0.25 for *avnA* and *omtA*. The HKY85 model (Hasegawa et al., 1985) extends the K80 model by allowing for unequal base frequencies, so freq(A) $\neq$ freq(C) $\neq$ freq(G) $\neq$ freq(T) (Felsenstein, 2004). This model was chosen for the *aflT*, *aflR* and *moxY* genes. Finally, the TPM3uf or '3 Parameter Model' (Kimura, 1981) chosen for the *fasB* and *estA* genes, has equal transition rates, two transversion rates, and accommodates for unequal base frequencies (Bazinet, 2013). The ratio of the transition rate to transversion rate is given by t_i/t_v, which ranges from 1.68 to 2.12, indicating the greater occurrence of transitions compared to transversions in all gene sequences.

TABLE 7.2. Statistical model of evolution for each gene as chosen by JModelTest using the Bayesian Information Criterion (BIC). The genes are listed according to their order in the pathway. The metrics are explained in the text.

Gene	Model	Partition	I	Γ	freq (A)	freq (C)	freq (G)	freq (T)	t_i/t_v	$-lnL$
aflT	HKY85	010010	–	–	0.1860	0.2862	0.2625	0.2653	1.6834	4892.9649
fasB	TPM3uf	012012	0.3830	–	0.2231	0.2729	0.2615	0.2426	–	13160.7553
aflR	HKY85	010010	–	0.8730	0.2052	0.3103	0.2734	0.2111	1.6861	3246.6417
estA	TPM3uf	012012	–	0.3490	0.2101	0.2446	0.3153	0.2300	–	2554.8237
avnA	K80	010010	–	0.6000	0.2500	0.2500	0.2500	0.2500	1.9422	4649.7273
omtA	K80	010010	–	0.5140	0.2500	0.2500	0.2500	0.2500	1.7773	4670.4624
moxY	HKY85	010010	–	0.2230	0.2248	0.2760	0.2876	0.2116	2.1171	5182.5132

Of these models, only the model for *fasB* has the parameter I, which is the proportion of invariant sites. This is quantified by the probability that the site is not permitted to vary (it is fixed) (M Charleston, personal communication), which for the *fasB* gene is 0.3830.

It was explained in Section 6.3.2 that the Γ distribution, which measures the variation of substitution rates among sites, is described by the shape parameter α. From Table 7.2, we see that wherever a Gamma distribution is applicable, that is in the models of genes *aflR* to *moxY*, the values of α range from 0.22 for *moxY* to 0.87 for *aflR*, and are all less than 1. This indicates that the distribution of rates is spread out and no longer follows a normal distribution (Felsenstein, 2004), with many sites evolving very slowly (with rates close to zero) and the rest evolving with rates that are quite high (Liò and Goldman, 1998). There does not appear to be any trend in the value of α as we move along the genes in the pathway, only that all $\alpha < 1$. Given the extent of diversity as shown by the similarity matrices (Figure 7.6) and splits networks (Section 7.3.2.2), it is not suprising that there is rate variation amongst the sites, and that gamma parameter is present in almost in every model.

In the group of genes studied, only the *aflT* gene does not have any I or Γ parameter associated with its model, hence for this gene it is assumed that the rates of change are equal at all sites in the alignment.

The frequencies of the bases A, C, G and T offer another means to measure the evolutionary dynamics of the genes by inferring the evolutionary rates of these bases. In particular, the the evolutionary rates of the bases C and G can be useful as genomic regions with a high proportion of C's and G's are often

targets for the molecular process of 'methylation', a regulatory mechanism that can increase or decrease the expression of associated genes. In the 7 genes studied, there does not appear to be much difference between the base frequencies between and across genes, suggesting that they are undergoing evolution at similar rates. Note that while the K80 model fixes the base frequencies to be equal, the frequencies in the two other models are still similar even though they have been calculated empirically from the data.

The last column is the negative log likelihood of the model as calculated by the BIC, where smaller values indicate a better fit of the model. Thus all the models provide a good fit to the data, with the *fasB* gene having the best fitting model as evidenced by the lowest negative log likelihood value.

Although JModelTest is commonly used for model selection, some critics argue the reasons for not incorporating it into phylogenetic analyses. In the model selection process, the data itself chooses or infers the model, which is then used to build the tree, making it prone to loss of apriori information (M Charleston, personal communication).

7.3.2.4 Phylogenetic Trees

The evolutionary models chosen from the previous section were used to construct maximum likelihood phylogenetic trees for each gene (Section 6.3.2), and are displayed in Figures 7.16 through 7.22 in the order they occur in the pathway. A separate figure with all seven trees is displayed for ease of visualisation and comparison in Figure 7.23.

A visual inspection of all 7 phylogenetic trees reveals that the majority of clades in each of the trees are well supported in terms of bootstrap values for each node, where a threshold of 70%, or 700 out of 1000, was applied as a minimum value for a branching to be considered valid. The trees built from the *fasB* (Figure 7.17) and *avnA* (Figure 7.20) genes are exemplary of this, with most nodes having perfect bootstrap values of 1000 (100%). This is consistent with the results from the splits networks (Section 7.3.2.2), where the *fasB* network (Figure 7.8) and the *avnA* network (Figure 7.11) have the fewest conflicts and are the most tree-like.

Another noteworthy and common feature in all the trees is the positioning of *A flavus* strain BN008, which is consistently placed on its own branch, closer to the *A nomius* clade, such that it is not a member nor a close relative of the *A flavus* and *A flavus-oryzae* clades, as would have been expected. Its

separation is supported by 100% bootstrap support in all trees except *avnA* and *moxY*. In all the gene trees where the bootstrap value is 100%, this strain appears to have diverged from the ancestor of the *A flavus*, *A flavus-oryzae*, *A minisclerotigenes* and *A parasiticus* species. Again, this is in accordance with the patterns seen for all genes in the similarity matrices (Figure 7.6) where the BN008 strain does not appear to be similar to any of the species in this study, and in the splits networks where it is consistently placed on its own edge (Section 7.3.2.2). Combining these results strongly suggests that strain BN008 is not a species of *A flavus*,and has perhaps been incorrectly labelled in the original GenBank entry.

In each of the trees, the four *A nomius* species form their own group as a result of being specified as outgroups during the creation of the trees (Section 6.3.2). Within these groups, the order of the strains is constant, with FGS11 being more divergent than the others and forming its own branch. All 25 similarity matrices also demonstrate that FGS11 is divergent from the other *A nomius* species as shown by the first row and column in the bottom right hand corner square.

The clades consisting of *A flavus* and *A flavus-oryzae*, despite having high bootstrap values across all genes for the clade itself, the branching pattern of the individual species within these clades is not well supported as indicated by the low bootstrap values. This lack of resolution is also evident in the all splits networks (Section 7.3.2.2), where there are a series of parallel lines and boxes at the base of these clades. Although the branching patterns within this clade are not well supported, the clade itself, and the member species of this clade are well supported by high bootstrap values. In all 7 phylogenetic trees, the *A flavus-oryzae* strains are not separated from the *A flavus* strains and appear to form a sub-clade within the *A flavus* species. This topology of the trees produced strongly disagrees with the tree generated by Gibbons et al., using sites across the whole genome rather than single genes, which clearly shows *A flavus* and *A oryzae* as distinct species on separate clades (Figure 7.14). This result may have been attained due to Gibbons et al.'s choice of *A oryzae* strains that all originated in Japan. In this study, the three novel *A oryzae* strains sequenced originate from Singapore, Korea and Thailand respectively (Table 1.1), to remove any biases the geographical location may have had. Thus our results provide strong support in favour of *A flavus* and *A oryzae* being different strains of the same species, rather than completely different species.

The trees generated from the *fasB* (Figure 7.17), *aflR* (Figure 7.18) and *estA* (Figure 7.19) genes display a consistent phylogenetic signal, as the clades appear to be similar in the constituent member species.

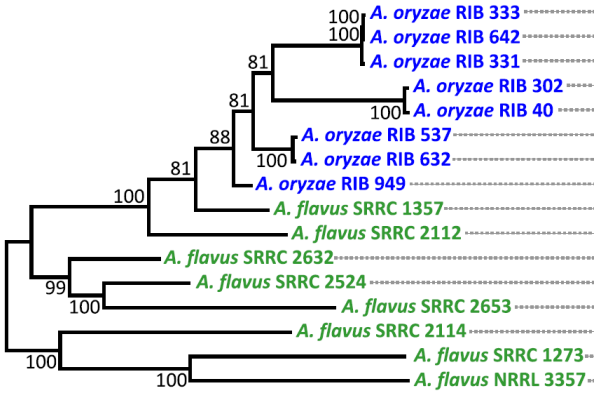


FIGURE 7.14. Phylogenetic tree presented by Gibbons et al. (2012)

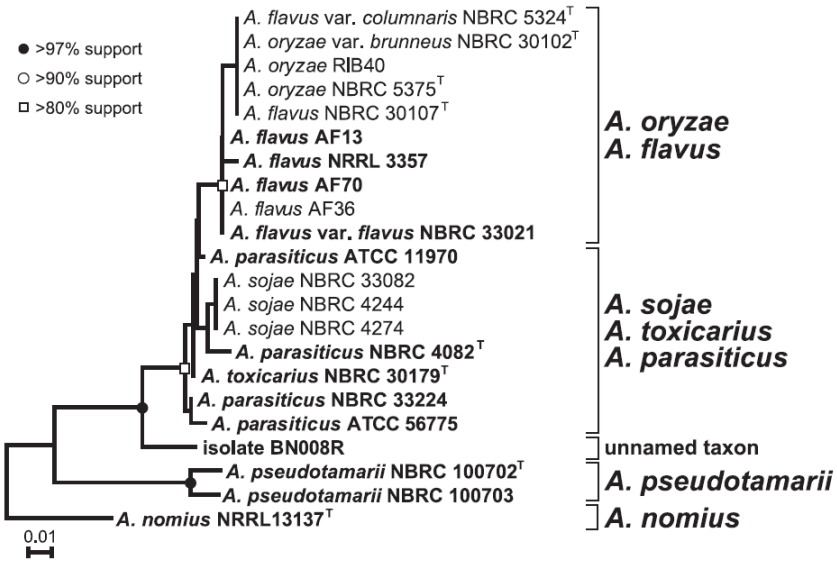


FIGURE 7.15. Phylogenetic tree on the *aflR* gene presented by Nakamura et al. (2011)

The branch lengths, or the distance between the root to the clades, are also similar. The topology of the *aflR* gene tree in Figure 7.18 is highly consistent with the *aflR* gene tree published by Nakamura et al. (2011), and is presented in Figure 7.15. Both *aflR* trees show *A flavus* and *A oryzae* strains in the same clade, and the BN008 strain forming its own separate branch. The high degree of consistency between the *aflR* tree generated here and the *aflR* tree published in literature provides validation of the methods and results attained in this study.

This pattern and topology begins to change in the trees for *omtA* (Figure 7.21) and *moxY* (Figure 7.22), where the branching pattern and clades appear shuffled. This rearrangement of clades however is backed by very low bootstrap values, and hence is not a robust branching. These low bootstrap values provide further evidence of the possible conflict present in the data that was illustrated in the similarity matrices (Figure 7.6) and splits networks of these two genes (Figures 7.12 and 7.13). In addition, the *aflT* and *moxY* trees (Figures 7.16 and 7.22 respectively) provide evidence of evolutionary rate heterogeneity, as the branch lengths appear highly variable unlike in the other gene trees. The *A flavus-oryzae* clade is separated by a long branch in the *aflT* gene tree, while the *A minisclerotigenes* clade is well separated in the *moxY* gene tree, however the separation is only well supported in the former, as in the latter the bootstrap value is 659, which falls short of the threshold of 700.

These rearranged patterns and variable branch lengths is suggestive of possible positive selection acting on these genes or the occurrence of horizontal gene transfer, an evolutionary process involving the transmission of genes between species by means other than from parents to offspring via reproduction. If neither of these reasons are valid, then the extent of mismatch between the trees is likely caused simply by the lack of phylogenetic signals. (Carbone et al., 2007) carried out investigations into the evolution of the aflatoxin cluster across five *Aspergillus* species and found only weak phylogenetic evidence in support of horizontal gene transfer. They propose that intragenomic re-organisation such as gene duplications and losses, in conjunction with vertical gene transfer are more plausible mechanisms responsible for the evolution of the aflatoxin gene cluster. This view conflicts with claims that horizontal gene transfer has played a major role in shaping the genomes of *Aspergillus* species (Chang and Ehrlich, 2010; Joardar et al., 2012; Gibbons and Rokas, 2013).

The results attained also imply that the *aflT*, *omtA* and *moxY* genes do not follow a 'molecular clock', in which the the rates of change, or expected number of substitutions per year, is constant across all the branches of a tree (`http://en.wikipedia.org/wiki/Substitution_model`). In other words, the rates of change between an ancestral species and its any of its descendants is expected to be equal as all descendants have acquired the same number of substitutions since diverging from the common ancestor (Posada, 2003). A molecular clock hypothesis is desirable as the process of phylogeny reconstruction becomes more easier and accurate (Posada, 2003). Thus further work is required to gain

a more thorough understanding of the evolutionary dynamics of this set of genes. Possible focuses and directions of further investigations are discussed in the following section.

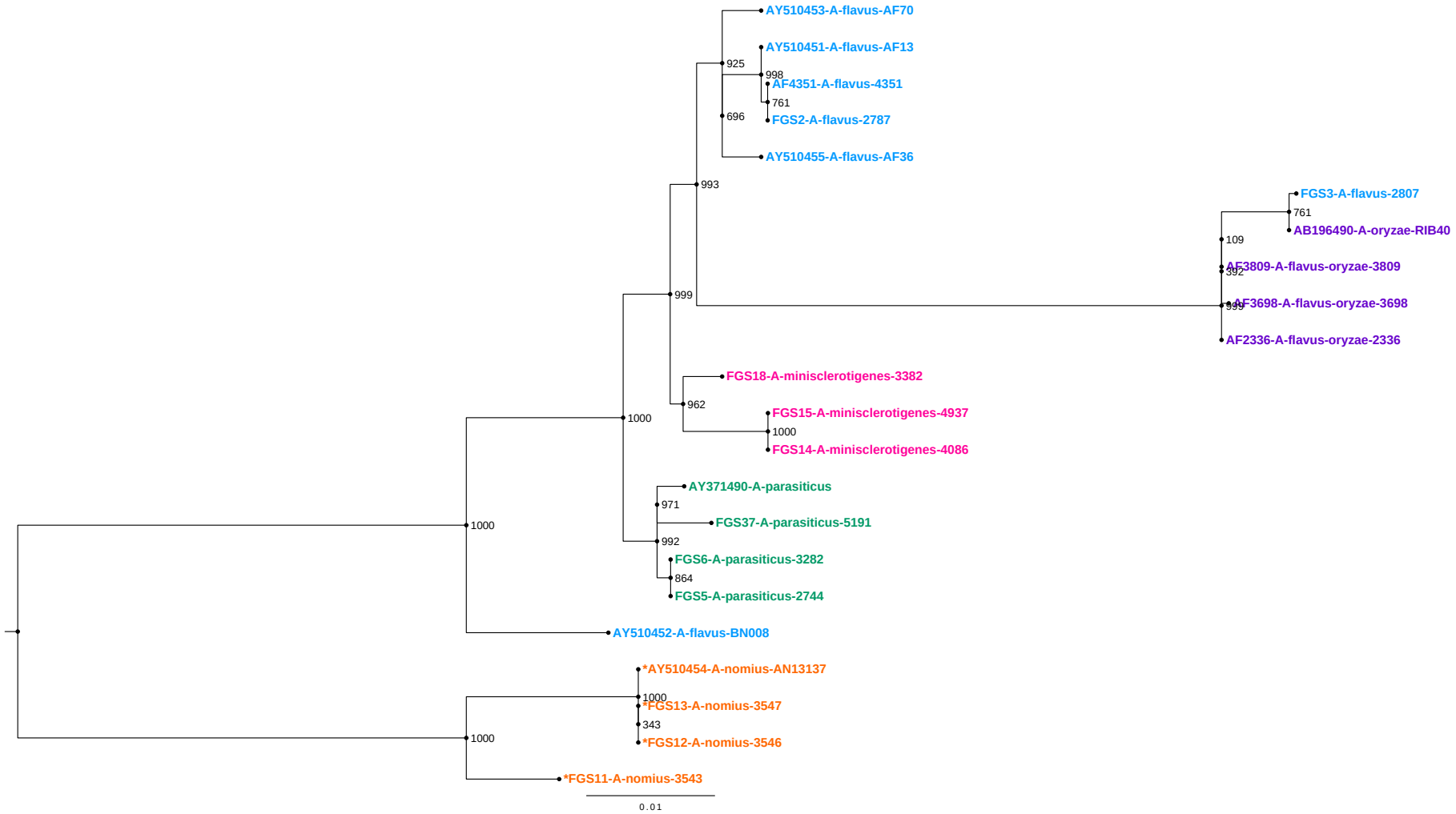


FIGURE 7.16. Maximum Likelihood Phylogenetic Tree for the *aflT* gene generated using PhyML. The numbers on the nodes represent bootstrap support values out of 1000. The scale represents the branch length, a measure of the expected number of substitutions per site.

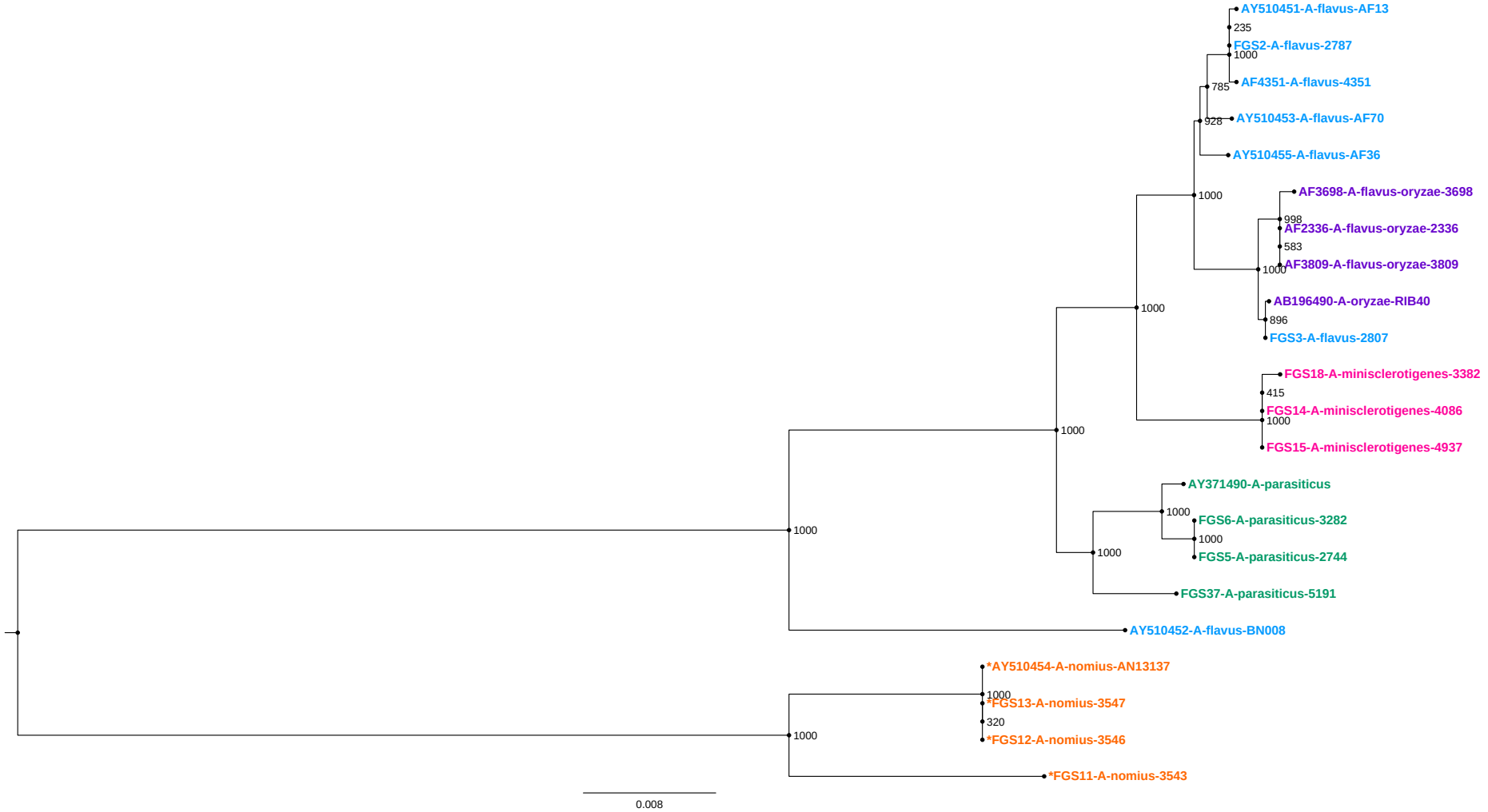


FIGURE 7.17. Maximum Likelihood Phylogenetic Tree for the *fasB* gene generated using PhyML. The numbers on the nodes represent bootstrap support values out of 1000. The scale represents the branch length, a measure of the expected number of substitutions per site.

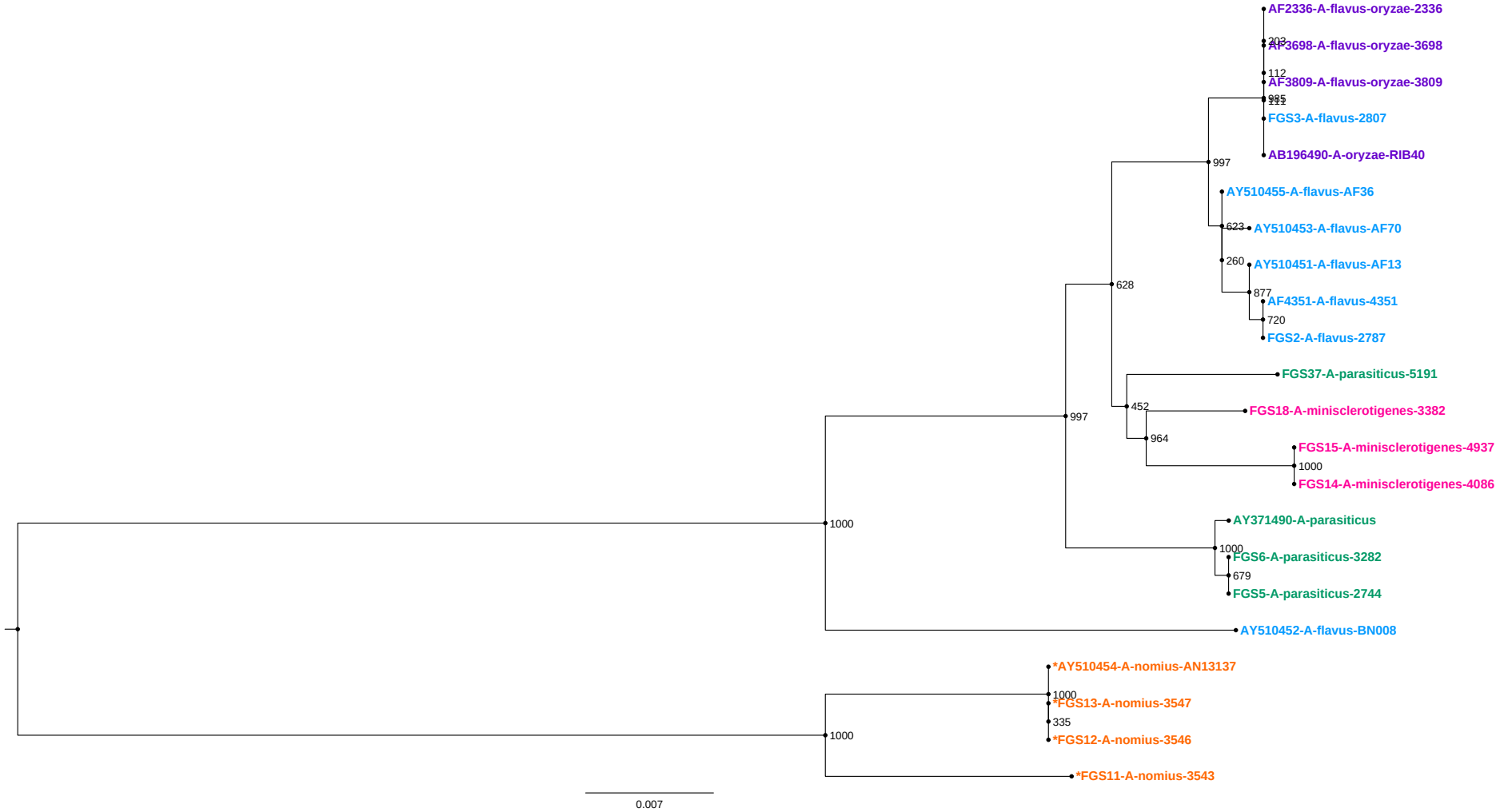


FIGURE 7.18. Maximum Likelihood Phylogenetic Tree for the *aflR* gene generated using PhyML. The numbers on the nodes represent bootstrap support values out of 1000. The scale represents the branch length, a measure of the expected number of substitutions per site.

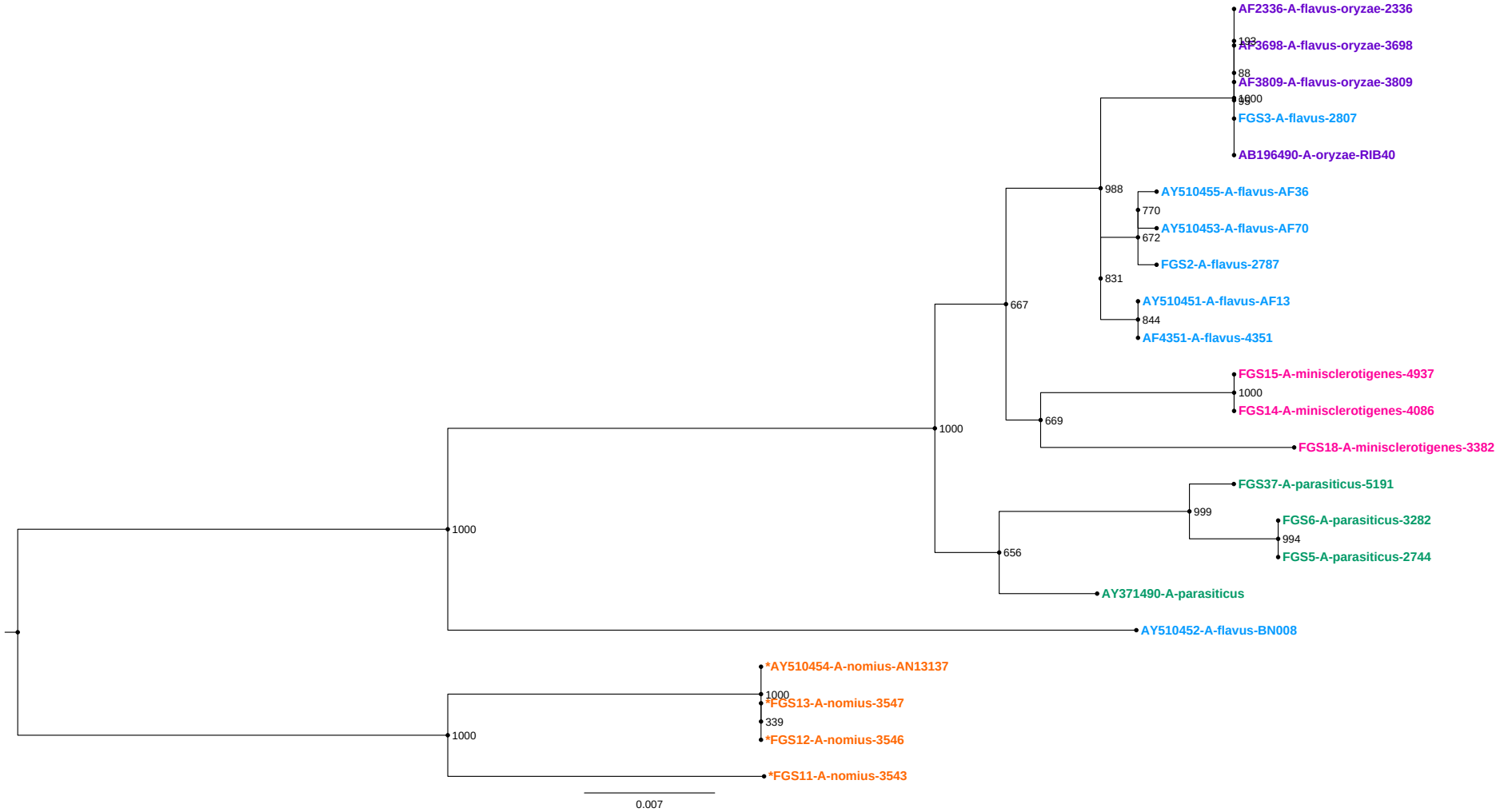


FIGURE 7.19. Maximum Likelihood Phylogenetic Tree for the *estA* gene generated using PhyML. The numbers on the nodes represent bootstrap support values out of 1000. The scale represents the branch length, a measure of the expected number of substitutions per site.

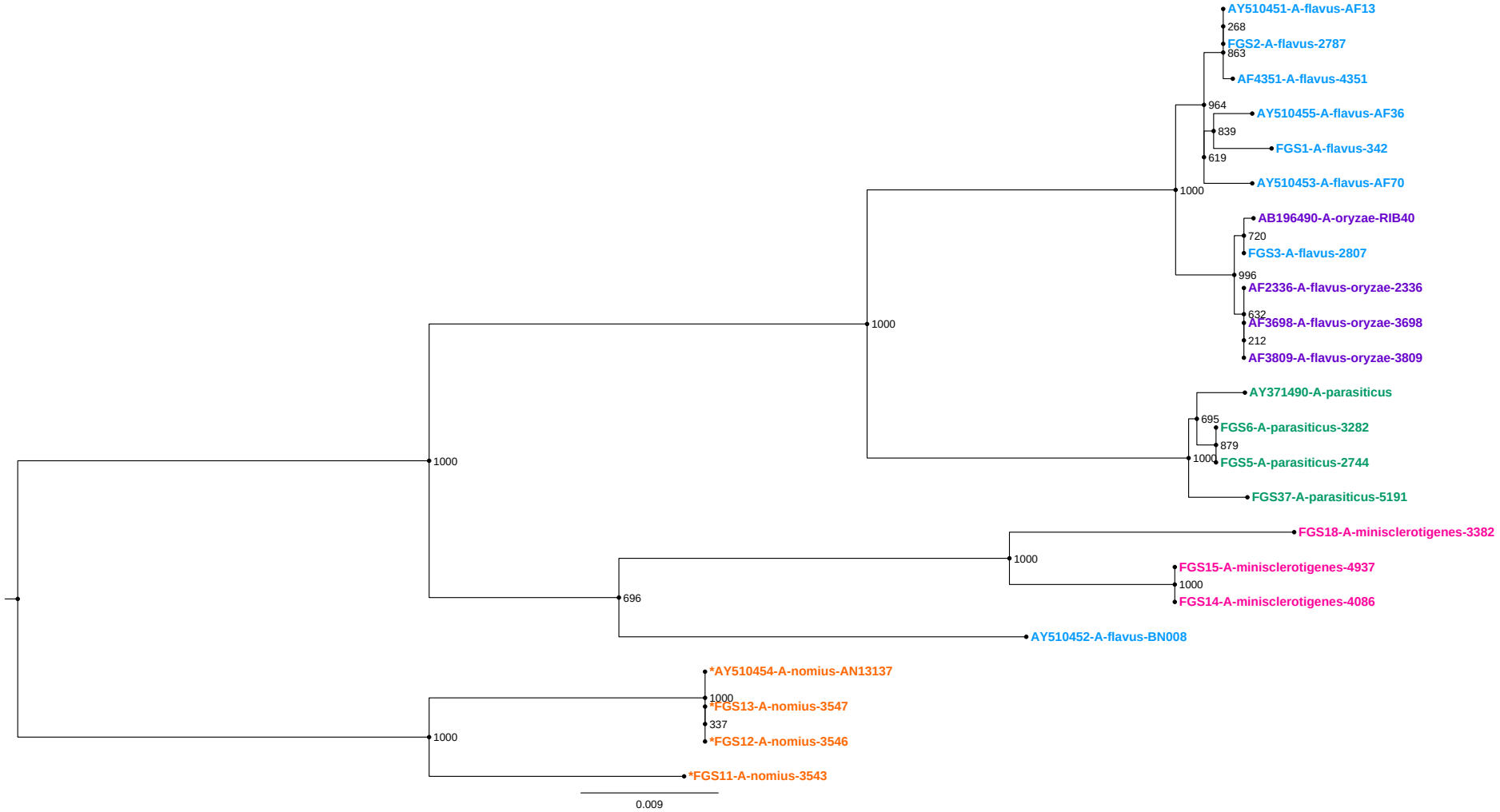


FIGURE 7.20. Maximum Likelihood Phylogenetic Tree for the *avnA* gene generated using PhyML. The numbers on the nodes represent bootstrap support values out of 1000. The scale represents the branch length, a measure of the expected number of substitutions per site.

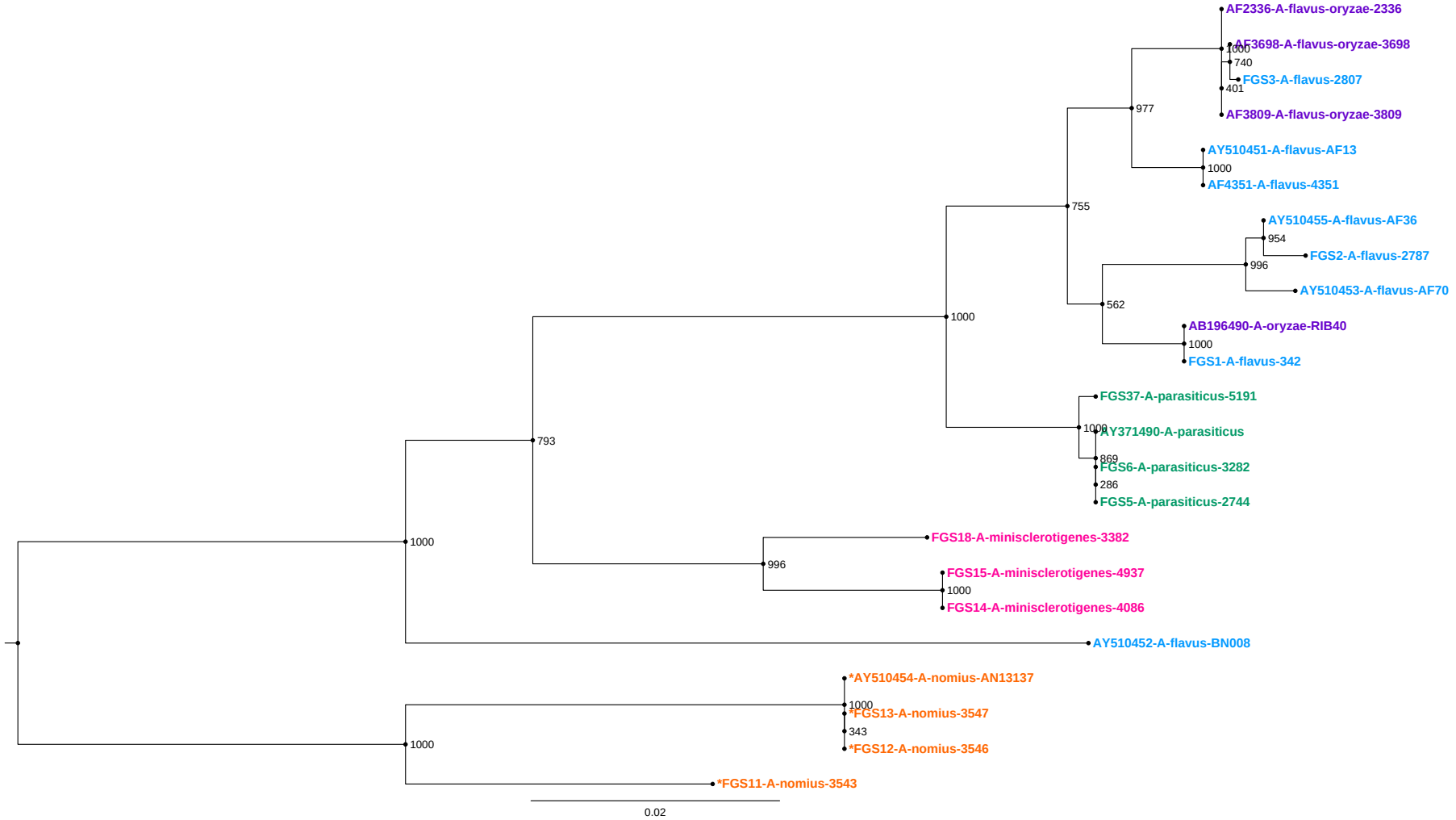


FIGURE 7.21. Maximum Likelihood Phylogenetic Tree for the *omtA* gene generated using PhyML. The numbers on the nodes represent bootstrap support values out of 1000. The scale represents the branch length, a measure of the expected number of substitutions per site.

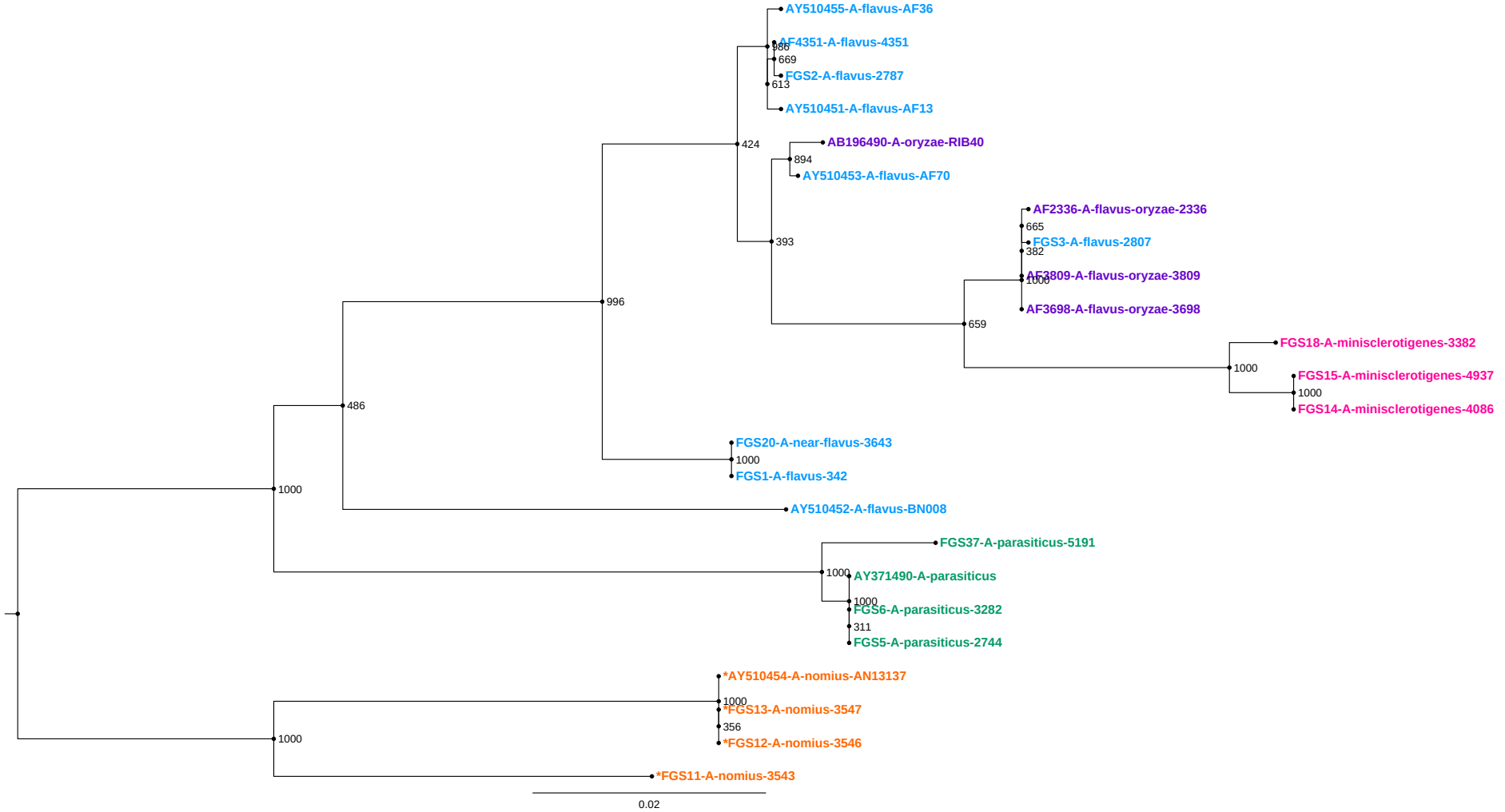


FIGURE 7.22. Maximum Likelihood Phylogenetic Tree for the *moxY* gene generated using PhyML. The numbers on the nodes represent bootstrap support values out of 1000. The scale represents the branch length, a measure of the expected number of substitutions per site.

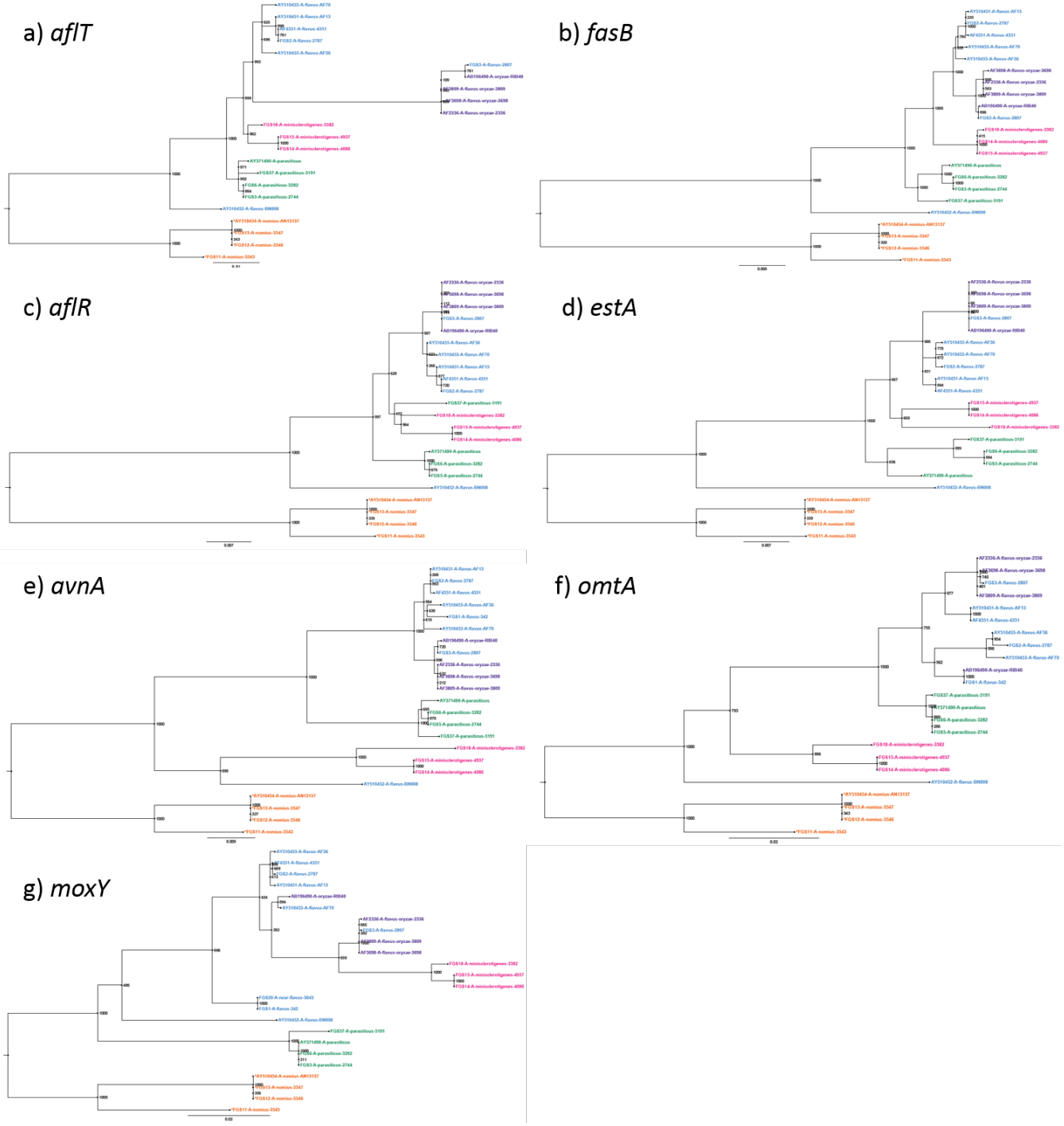


FIGURE 7.23. Comparison of Maximum Likelihood Phylogenetic Trees constructed for each of the seven genes.

7.4 Future Work

Recommendations and directions for future work to complement and provide further insight into the investigations performed in this study are discussed in this section.

From Section 7.3.1, all three strains of *A flavus-oryzae* contain the full aflatoxin pathway, including the *aflR* gene required to activate this pathway, with deletions in the *norB* and *cypA* genes. It should be investigated whether these species have the same mutations in the *aflR* and *aflJ* genes that have been published in literature (Chang et al., 2007; Kiyota et al., 2011) that diminishes its aflatoxin-producing capability.

The gene k-mer similarity matrices (Section 7.3.1.2) were analysed and the genes assigned to clusters via manual inspection. For the conclusions drawn to be more robust, less subjective techniques such as metrics to measure the difference between matrices with a defined notion of distance, or clustering algorithms, could be used to group genes that appear to be evolving under similar dynamics. These clusters could then be used to guide evolutionary model selection rather than based on the data itself, which has the potential to cause a loss of apriori information, and was highlighted as a criticism of this process of choosing statistical models (Section 7.3.2.3).

As the evolutionary dynamics were investigated for 7 genes only, extending this work to all the 25 genes in the pathway is logical and worthwhile for gaining a complete understanding of the evolutionary dynamics of the pathway. From the literature review (Chapter 5), several published works reported results where multi-gene trees were more informative than single gene trees (Galagan et al., 2005; Joardar et al., 2012). Concatenated or consensus trees could be generated for all the genes as a whole, or using the genes within each cluster, to increase the robustness of the trees. Building a consensus tree would be feasible because the branching patterns in the phylogenetic trees in Figures 7.16 to 7.22 are fairly consistent. Each individual gene tree could then be compared with the consensus trees to identify and examine differences. This could be particularly useful to resolve the conflicting phylogenetic signal for genes such as *omtA* and *moxY*, which may be more informative in a combination rather than as separate trees.

The Maximum Likelihood trees built in this study could also be compared to Maximum Parsimony trees inferred from the same genes. This method was used by Joardar et al. (2012) as a means of cross-validating both types of phylogenetic trees generated.

For each tree built, it may also be necessary to use a combination of substitution models, rather than a single model, to provide a better fit of data with different selective pressures and evolutionary constraints acting upon different genomic regions (Posada, 2003).

Another method of comparing the evolutionary models is to use the Likelihood Ratio Test (LRT) to accept or reject the hypothesis of a molecular clock, a more general model containing more parameters, in which the branch lengths from a root to any terminal node are constant, that is, all taxa are assumed to be evolving at the same rate. The null hypothesis the rates of evolution are equal across all lineages, and the alternative hypothesis is that rates are allowed to vary independently (Posada, 2003).

In addition, positive selection can be tested for using the LRT, as well as dN/dS ratios, which measure the rate of non-synonymous changes N (where a nucleotide substitution in DNA results in a different amino acid being incorporated into the protein, which may lead to an altered protein function) to synonymous changes S (where a nucleotide substitution in DNA results in the same amino acid being incorporated into the protein so no changes in its function). A ratio greater than 1 implies positive selection, a ratio less than 1 implies stabilising selection, and a ratio equal to 1 implies none or neutral selection. dN/dS ratios of less than 1 were reported by Ehrlich *et al* (2005) for all genes in the pathway except the last gene *hypA*, which had a dN/dS ratio of 1.04. A similar investigation should be conducted for the to compare the selective pressures of the aflatoxin genes in these novel fungal genomes.

In terms of the *A flavus / A oryzae* debate, Nakamura et al. concatenated the sequences of four aflatoxin genes to produce a tree in which the *A oryzae* strains were separate to the *A flavus* strains. The same test should be reproduced and applied to our fungal genomes to judge whether a distinct grouping of *A oryzae* is indeed attained. Gibbons et al. report the metabolic specialisation of *A oryzae*, where the genes in major metabolic pathways had undergone rearrangements. Thus future work would investigate for signs of domestication of *A oryzae* by focusing on these metabolic pathways. In addition, estimates of the age of the *A flavus / A oryzae* clade, and when the two 'species' were expected to have diverged, could provide further insights into the exact nature of this relationship.

7.5 Conclusions

From the work undertaken in this part of the study has shown that the evolutionary models and rates of change are similar from Section 7.3.2.3, suggesting that the whole aflatoxin cluster is under the same stabilising selection constraints and thus the genes have consistent evolutionary dynamics.

The splits networks have shown a clear, consistent phylogenetic signal from most of the genes (Section 7.3.2.2), and the phylogenetic trees show distinct grouping of species into clades that are well supported by high bootstrap values, thus making them reliable. From these analyses we can infer that there is a high degree of conservation of the aflatoxin pathway among the studied novel fungal genomes. Towards the end of the pathway, however, there may be slightly less conservation due to the different tree topologies of the *omtA* and *moxY* genes. These genes, encoding a methyltransferase and dehydrogenase respectively, may be less constrained and more free to undergo changes due to the possible presence of other genes available to perform these functions in case *omtA* or *moxY* alter or lose their function.

This study also offers important insights into the relationship between *A flavus* and *A oryzae*. The results from the similarity matrices, splits networks and phylogenetic trees all show evidence for *A oryzae* not being a separate species, but rather a derived clade of *A flavus*. *A oryzae* strains may be undergoing 'speciation' (to become a separate species) as a result of the particular set of strains that are used in the food industry for their non-toxigenic properties, and are thus being subject to selection driven by the food industry.

Finally, there is unquestionable evidence that the sequence given by accession number AY510452 is not an *A flavus* species. Firstly, it contains he full *norB* and *cypA* genes which makes it capable of producing G-type aflatoxins, which *A flavus* cannot produce. All similarity matrices, splits networks and phylogenetic trees show this species is not close to any of the *A flavus* or *A flavus-oryzae* species, or any of the other species in this study. It was not an objective of this study to determine the identity of this species; we can conclude, however, that it is a aflatoxin producing species (likely both B- and G-type aflatoxins) that diverged from a common ancestor of *A flavus*, *A flavus-oryzae*, *A minisclerotigenes* and *A parasiticus*. In the original paper, Ehrlich et al. call this species an 'unnamed taxon', however, the Genbank entry has the sequence labelled as '*Aspergillus flavus* isolate BN008' (`http://www.ncbi.nlm.nih.gov/nuccore/AY510452`). This observation reiterates the well known fact that errors

exist in Genbank entries as explained in Section 4.3. The labelling of this sequence in Genbank should be rectified appropriately and immediately.

CHAPTER 8

Conclusions

Species in the fungal kingdom are of critical biological importance to humans and the environment as they play a vital role in functions and processes required to sustain life on earth. Fungi are decomposers and recyclers of nutrients, act in symbiosis with 90% of all plant species, are used as a direct source of food or in industrial food fermentation processes to produce bread, beer and wine etc. Fungi also have applications in medicine; some species produce useful compounds that have been exploited in pharmaceutical drugs, with the most notable being the antibiotic penicillin, while species of *Aspergillus* naturally produce carcinogenic aflatoxins that contaminate agricultural crops and foods, and pose a risk to food safety.

These wide ranging impacts of fungi, that are both beneficial and harmful, necessitates fast and accurate methods of classifying novel or unknown fungi. Furthermore, a better understanding of the aflatoxin genes is required to elucidate why certain species produce aflatoxins and others do not, and for the characterisation of evolutionary relationships between such organisms. These are the reasons motivating the current study, which was divided into two parts to address each of these unresolved problems.

Part 1 of this thesis has described the processes taken to build a rapid classification tool that is based on sequence composition rather than sequence similarity or BLAST-based methods, and is accurate down to the species level. This classifier is built based on the sequences of the highly variable ITS region, which provides greater power of resolution than the LSU gene that was used by Liu et al. (2012), to enable classifications to go beyond the genus level and right down to the species level. Extensive processing and screening of the original set of ITS sequences was conducted to create a curated set of 24,447 high quality ITS sequences, spanning 9,073 species, that form the current version of the training set. This training set was evaluated using LOOCV, the same method adopted by Liu et al. to enable comparison withe their LSU classifier. An increase in accuracy of 6.2% was observed at the genus level

to 98.8%, and an accuracy of 90.2% was attained at the species level. The performance of the classifier was thoroughly tested and validated using a specially constructed validation set, which had accuracies of 98% and 64% at the genus and species levels respectively. The time taken for the classification of 1400 ITS sequences was of the order of a couple of minutes. Comparable accuracies were obtained when short, 400 bp sequences were classified, to simulate the common types of sequences that would be submitted by fungal biologists, or the end users of this classification tool. The classifier developed in this study therefore achieves the objectives of formulating fast, accurate classifications down to the species level. It is hoped that the classifier, along with the curated set of ITS sequences in the training set, will be a valuable resource for fungal biologists, and be integrated as an analysis step to aid fungal studies.

Part 2 of the thesis has explored the genes responsible for aflatoxin biosynthesis in a group of 17 novel fungal organisms, and discussed the comparative analysis performed to study the evolutionary dynamics of a representative subset of the 25 genes. The analysis pipeline devised involved firstly assembling the raw sequence reads of these novel genomes, extracting the gene sequences. k-mer similarity matrices and splits networks were created as a visual means of detecting phylogenetic signals in the data. Maximum likelihood phylogenetic trees were inferred for the genes under a particular statistical model of evolution. Our results suggest that the aflatoxin pathway is conserved across the fungal species, and there are similar evolutionary forces acting upon the genes. An important outcome was evidence supporting the notion that *A oryzae* and *A flavus* are variants of the same species, and that the relationship between these is that *A oryzae* is a subspecies, or special type of clade, of *A flavus*. These results refute previous claims made by Gibbons et al. in (2012) that *A oryzae* and *A flavus* are separate species. Thus further work is required before a consensus can be reached on the relationship between these two organisms. The knowledge gained from these results have the potential to ultimately be used in devising more effective strategies for controlling aflatoxin contamination and thereby improve food safety.

As the next step, we intend to implement the recommended work outlined in the future work section for both parts of this study. In conjunction with the findings already made, this will form the basis of publications that will be submitted to journals such as Applied and Environmental Microbiology, where papers about the RDP 16S classifier (Wang et al., 2007) and LSU classifier (Liu et al., 2012) were submitted to. Efforts are also under way to collaborate with the developers of the RDP classifiers, to

provide a platform that will enable our ITS classifier to be publicly available for utilisation by the wider fungal research community.

Bibliography

K Abarenkov, L Tedersoo, R H Nilsson, K Vellak, Irja Saar2, V Veldre, E Parmasto, M Prous, A Aan, M Ots, K Kurina, I Ostonen, J Jõgeva, S Halapuu, K Põldmaa, M Toots, J Truu, K Larsson, and U Kõljalg. 2010. PlutoFâĂŤa Web Based Workbench for Ecological and Taxonomic Research, with an Online Implementation for Fungal ITS Sequences. *Evolutionary Bioinformatics*, 6:189–196.

Vernon Ahmadjian. 2013. Fungus.

Takeshi Akao, Motoaki Sano, Osamu Yamada, Terumi Akeno, Kaoru Fujii, Kuniyasu Goto, Sumiko Ohashi-Kunihiro, Kumiko Takase, Makoto Yasukawa-Watanabe, Kanako Yamaguchi, Yoko Kurihara, Jun-ichi Maruyama, Praveen Rao Juvvadi, Akimitsu Tanaka, Yoji Hata, Yasuji Koyama, Shotaro Yamaguchi, Noriyuki Kitamoto, Katsuya Gomi, Keietsu Abe, Michio Takeuchi, Tetsuo Kobayashi, Hiroyuki Horiuchi, Katsuhiko Kitamoto, Yutaka Kashiwagi, Masayuki Machida, and Osamu Akita. 2007. Analysis of expressed sequence tags from the fungus Aspergillus oryzae cultured under different conditions. *DNA Research*, 14(2):47–57.

Stephen F Altschul, Warren Gish, Webb Miller, Eugene E Myers, and David J Lipman. 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215:403–410.

Saori Amaike and Nancy P Keller. 2011. Aspergillus flavus. *Annual Review of Phytopathology*, 49:107–33.

S A Balajee, A M Borman, M E Brandt, J Cano, M Cuenca-Estrella, E Dannaoui, J Guarro, G Haase, C C Kibbler, W Meyer, K O'Donnell, C A Petti, J L Rodriguez-Tudela, D Sutton, A Velegraki, and B L Wickes. 2009. Sequence-based identification of Aspergillus, fusarium, and mucorales species in the clinical mycology laboratory: where are we and where should we go from here? *Journal of Clinical Microbiology*, 47(4):877–84.

Scott T Bates, Steven Ahrendt, Holly M Bik, Thomas D Bruns, J Gregory Caporaso, James Cole, Michael Dwan, Noah Fierer, Dai Gu, Shawn Houston, Rob Knight, Jon Leff, Christopher Lewis, Juan P Maestre, Daniel McDonald, R Henrik Nilsson, Andrea Porras-Alfaro, Vincent Robert, Conrad Schoch, James Scott, D Lee Taylor, Laura Wegener Parfrey, and Jason E Stajich. 2013. Meeting Report: Fungal ITS workshop (October 2012). *Standards in Genomic Sciences*, 8(1):118–23.

Adam Bazinet. 2013. Nucleotide Substitution Models.

Scott W Behie, Israel E Padilla-Guerrero, and Michael J Bidochka. 2013. Nutrient transfer to plants by phylogenetically diverse fungi suggests convergent evolutionary strategies in rhizospheric symbionts.

Communicative & Integrative Biology, 6(1):e22321.

Meredith Blackwell. 2011. The fungi: 1, 2, 3 ... 5.1 million species? *American Journal of Botany*, 98(3):426–38.

Ignazio Carbone, Jorge H Ramirez-Prado, Judy L Jakobek, and Bruce W Horn. 2007. Gene duplication, modularity and adaptation in the evolution of the aflatoxin gene cluster. *BMC Evolutionary Biology*, 7:111.

Perng-Kuang Chang and Kenneth C Ehrlich. 2010. What does genetic diversity of Aspergillus flavus tell us about Aspergillus oryzae? *International Journal of Food Microbiology*, 138(3):189–99.

Perng-Kuang Chang, Kenichiro Matsushima, Tadashi Takahashi, Jiujiang Yu, Keietsu Abe, Deepak Bhatnagar, Gwo-Fang Yuan, Yasuji Koyama, and Thomas E Cleveland. 2007. Understanding nonaflatoxigenicity of Aspergillus sojae: a windfall of aflatoxin biosynthesis research. *Applied Microbiology and Biotechnology*, 76(5):977–84.

Thomas E Cleveland, Jiujiang Yu, Natalie Fedorova, Deepak Bhatnagar, Gary A Payne, William C Nierman, and Joan W Bennett. 2009. Potential of Aspergillus flavus genomics for applications in biotechnology. *Trends in Biotechnology*, 27(3):151–7.

Kenneth W Cullings and Detlev R Vogler. 1998. A 5.8S nuclear ribosomal RNA gene sequence database: applications to ecology and evolution. *Molecular Ecology*, 7:919–923.

Karen C Dannemiller, Darryl Reeves, Kyle Bibby, Naomichi Yamamoto, and Jordan Peccia. 2013. Fungal High-throughput Taxonomic Identification tool for use with Next-Generation Sequencing (FHiTINGS). *Journal of Basic Microbiology*, 00:1–7.

Aaron C E Darling, Bob Mau, Frederick R Blattner, and Nicole T Perna. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14:1394–403.

Diego Darriba, Guillermo L Taboada, Ramón Doallo, and David Posada. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, 9(8):772.

Robert C Edgar. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–7.

K C Ehrlich, J Yu, and P J Cotty. 2005. Aflatoxin biosynthesis gene clusters and flanking regions. *Journal of Applied Microbiology*, 99(3):518–27.

Kenneth C Ehrlich, Perng-kuang Chang, Jiujiang Yu, and Peter J Cotty. 2004. Aflatoxin Biosynthesis Cluster Gene cypA Is Required for G Aflatoxin Formation Aflatoxin Biosynthesis Cluster Gene cypA Is Required for G Aflatoxin Formation. *Applied and Environmental Microbiology*, 70(11):6518–24.

Joseph Felsenstein. 2004. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts, second edition.

James E Galagan, Sarah E Calvo, Christina Cuomo, Li-Jun Ma, Jennifer R Wortman, Serafim Batzoglou, Su-In Lee, Meray BaÅ§türkmen, Christina C Spevak, John Clutterbuck, Vladimir Kapitonov, Jerzy Jurka, Claudio Scazzocchio, Mark Farman, Jonathan Butler, Seth Purcell, Steve Harris, Gerhard H

Braus, Oliver Draht, Silke Busch, Christophe D'Enfert, Christiane Bouchier, Gustavo H Goldman, Deborah Bell-Pedersen, Sam Griffiths-Jones, John H Doonan, Jaehyuk Yu, Kay Vienken, Arnab Pain, Michael Freitag, Eric U Selker, David B Archer, Miguel a Peñalva, Berl R Oakley, Michelle Momany, Toshihiro Tanaka, Toshitaka Kumagai, Kiyoshi Asai, Masayuki Machida, William C Nierman, David W Denning, Mark Caddick, Michael Hynes, Mathieu Paoletti, Reinhard Fischer, Bruce Miller, Paul Dyer, Matthew S Sachs, Stephen a Osmani, and Bruce W Birren. 2005. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature*, 438(7071):1105–15.

D Ryan Georgianna, Natalie D Fedorova, James L Burroughs, Andrea L Dolezal, Jin Woo Bok, Sigal Horowitz-Brown, Charles P Woloshuk, Jiujiang Yu, Nancy P Keller, and Gary A Payne. 2010. Beyond aflatoxin: four distinct expression patterns and functional roles associated with Aspergillus flavus secondary metabolism gene clusters. *Molecular Plant Pathology*, 11(2):213–226.

Tarini Shankar Ghosh, M Monzoorul Haque, and Sharmila S Mande. 2010. DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics*, 11(Suppl 7):S14.

John G Gibbons and Antonis Rokas. 2013. The function and evolution of the Aspergillus genome. *Trends in Microbiology*, 21(1):14–22.

John G Gibbons, Leonidas Salichos, Jason C Slot, David C Rinker, Kriston L McGary, Jonas G King, Maren a Klich, David L Tabb, W Hayes McDonald, and Antonis Rokas. 2012. The evolutionary imprint of domestication on genome variation and function of the filamentous fungus Aspergillus oryzae. *Current Biology*, 22(15):1403–9.

Paul Greenfield, Konsta Duesing, Alexi Papanicolaou, and Denis Bauer. 2013. Blue: correcting sequencing errors using consensus and context. *Bioinformatics*, (submitted.

Stéphane Guindon and Olivier Gascuel. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5):696–704.

Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. 1985. Journal of Molecular Evolution. *Journal of Molecular Evolution*, 22:160–174.

David S Hibbett and John W Taylor. 2013. Fungal systematics: is a new age of enlightenment at hand? *Nature Reviews Microbiology*, 11(2):129–33.

D H Huson. 1998. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73.

Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. 2007. MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–86.

Daniel H Huson, Suparna Mitra, Hans-Joachim Ruscheweyh, Nico Weber, and Stephan C Schuster. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9):1552–60.

Illumina. 2011. Quality Scores for Next-Generation Sequencing. Technical report.

Illumina. 2013a. An Introduction to Next-Generation Sequencing Technology. Technical report.

Illumina. 2013b. Sequencing.

Illumina. 2013c. Sequencing Systems.

Vinita Joardar, Natalie F Abrams, Jessica Hostetler, Paul J Paukstelis, Suchitra Pakala, Suman B Pakala, Nikhat Zafar, Olukemi O Abolude, Gary Payne, Alex Andrianopoulos, David W Denning, and William C Nierman. 2012. Sequencing of mitochondrial genomes of nine Aspergillus and Penicillium species identifies mobile introns and accessory genes as main sources of genome size variability. *BMC Genomics*, 13(1):698.

Urmas Kõljalg, Karl-Henrik Larsson, Kessy Abarenkov, R Henrik Nilsson, Ian J Alexander, Ursula Eberhardt, Susanne Erland, Klaus Hø iland, Rasmus Kjø ller, Ellen Larsson, Taina Pennanen, Robin Sen, Andy F S Taylor, Leho Tedersoo, Trude Vrå lstad, and Björn M Ursing. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *The New Phytologist*, 166(3):1063–8.

Motoo Kimura. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.

Motoo Kimura. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 78(1):454–8.

Takuro Kiyota, Ryoko Hamada, Kazutoshi Sakamoto, Kazuhiro Iwashita, Osamu Yamada, and Shigeaki Mikami. 2011. Aflatoxin non-productivity of Aspergillus oryzae caused by loss of function in the aflJ gene product. *Journal of Bioscience and Bioengineering*, 111(5):512–7.

David Koslicki, Simon Foucart, and Gail Rosen. 2013. Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. *Bioinformatics*, 29(17):2096–102.

Lene Lange, Lasse Bech, Peter K Busk, Morten N Grell, Yuhong Huang, Mette Lange, Tore Linde, Bo Pilgaard, Doris Roth, and Xiaoxue Tong. 2012. The importance of fungi and of mycology for a global development of the bioeconomy. *IMA Fungus*, 3(1):87–92.

M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–8.

Pietro Liò and Nick Goldman. 1998. Models of Molecular Evolution and Phylogeny. *Genome Research*, 8:1233–1244.

Kuan-Liang Liu, Andrea Porras-Alfaro, Cheryl R Kuske, Stephanie A Eichorst, and Gary Xie. 2012. Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. *Applied and Environmental Microbiology*, 78(5):1523–33.

Masayuki Machida, Kiyoshi Asai, Motoaki Sano, Toshihiro Tanaka, Toshitaka Kumagai, Goro Terai, Ken-Ichi Kusumoto, Toshihide Arima, Osamu Akita, Yutaka Kashiwagi, Keietsu Abe, Katsuya Gomi, Hiroyuki Horiuchi, Katsuhiko Kitamoto, Tetsuo Kobayashi, Michio Takeuchi, David W Denning,

James E Galagan, William C Nierman, Jiujiang Yu, David B Archer, Joan W Bennett, Deepak Bhatnagar, Thomas E Cleveland, Natalie D Fedorova, Osamu Gotoh, Hiroshi Horikawa, Akira Hosoyama, Masayuki Ichinomiya, Rie Igarashi, Kazuhiro Iwashita, Praveen Rao Juvvadi, Masashi Kato, Yumiko Kato, Taishin Kin, Akira Kokubun, Hiroshi Maeda, Noriko Maeyama, Jun-ichi Maruyama, Hideki Nagasaki, Tasuku Nakajima, Ken Oda, Kinya Okada, Ian Paulsen, Kazutoshi Sakamoto, Toshihiko Sawano, Mikio Takahashi, Kumiko Takase, Yasunobu Terabayashi, Jennifer R Wortman, Osamu Yamada, Youhei Yamagata, Hideharu Anazawa, Yoji Hata, Yoshinao Koide, Takashi Komori, Yasuji Koyama, Toshitaka Minetoki, Sivasundaram Suharnan, Akimitsu Tanaka, Katsumi Isono, Satoru Kuhara, Naotake Ogasawara, and Hisashi Kikuchi. 2005. Genome sequencing and analysis of Aspergillus oryzae. *Nature*, 438(7071):1157–61.

Masayuki Machida, Osamu Yamada, and Katsuya Gomi. 2008. Genomics of Aspergillus oryzae: learning from the history of Koji mold and exploration of its future. *DNA Research*, 15(4):173–83.

Jon K Magnuson and Linda L Lasure. 2002. Fungal Diversity in Soils as Assessed by Direct Culture and Molecular Techniques. In *102nd General Meeting of the American Society for Microbiology, Salt Lake City*, pages 19–23. Pacific Northwest National Laboratory, Richland, WA.

Jason R Miller, Sergey Koren, and Granger Sutton. 2010. Assembly Algorithms for Next-Generation Sequencing Data. *Genomics*, 95(6):315–327.

André E Minoche, Juliane C Dohm, and Heinz Himmelbauer. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology*, 12:R112.

M Monzoorul Haque, Tarini Shankar Ghosh, Dinakar Komanduri, and Sharmila S Mande. 2009. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25(14):1722–30.

David Moore. 2013. Evolution and phylogeny of fungi.

David Morrison, Leo van Iersel, Steven Kelk, and Michael Charleston. 2012. How to interpret splits graphs.

Kasper Munch, Wouter Boomsma, John P Huelsenbeck, Eske Willerslev, and Rasmus Nielsen. 2008. Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, 57(5):750–7.

Hitomi Nakamura, Takashi Narihiro, Naoki Tsuruoka, Hanako Mochimaru, Rena Matsumoto, Yumiko Tanabe, Keiko Hagiya, Kiriko Ikeba, Akihiko Maruyama, and Satoshi Hanada. 2011. Evaluation of the Aflatoxin Biosynthetic Genes for Identification of the Aspergillus Section Flavi. *Microbes and Environments*, 26(4):367–369.

R. Henrik Nilsson, Vilmar Veldre, Martin Hartmann, Martin Unterseher, Anthony Amend, Johannes Bergsten, Erik Kristiansson, Martin Ryberg, Ari Jumpponen, and Kessy Abarenkov. 2010. An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for

use in high-throughput community assays and molecular ecology. *Fungal Ecology*, 3(4):284–287.

G. A. Payne, W. C. Nierman, J. R. Wortman, B. L. Pritchard, D. Brown, R. a. Dean, D Bhatnagar, T. E. Cleveland, Masayuki Machida, and J. Yu. 2006. Whole genome comparison of Aspergillus flavus and A. oryzae. *Medical Mycology*, 44(s1):9–11.

Teresita M Porter and G Brian Golding. 2012. Factors that affect large subunit ribosomal DNA amplicon sequencing studies of fungal communities: classification method, primer choice, and error. *PloS ONE*, 7(4):e35749.

David Posada. 2003. Selecting models of evolution. In Marco Salemi and Anne-Mieke Vandamme, editors, *The Phylogenetic Handbook*, chapter 10, pages 256–282. Cambridge University Press, first edition.

David Posada. 2008. jModelTest: Phylogenetic Model Averaging. *Molecular Biology and Evolution*, 25(7):1253–6.

A Rokas, G Payne, N D Fedorova, S E Baker, M Machida, J Yu, D Ryan Georgianna, Ralph A Dean, Deepak Bhatnagar, T E Cleveland, J R Wortman, R Maiti, V Joardar, P Amedeo, D W Denning, and W C Nierman. 2007. What can comparative genomics tell us about species concepts in the genus Aspergillus? *Studies in Mycology*, 59:11–7.

N Saitou and M Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–25.

Atsushi Sato, Kenshiro Oshima, Hideki Noguchi, Masahiro Ogawa, Tadashi Takahashi, Tetsuya Oguma, Yasuji Koyama, Takehiko Itoh, Masahira Hattori, and Yoshiki Hanya. 2011. Draft genome sequencing and comparative analysis of Aspergillus sojae NBRC4239. *DNA Research*, 18(3):165–76.

Conrad L Schoch, Keith A Seifert, Sabine Huhndorf, Vincent Robert, John L Spouge, C André Levesque, and Wen Chen. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, 109(16):6241–6.

Vineet K Sharma, Naveen Kumar, Tulika Prakash, and Todd D Taylor. 2012. Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin. *PloS ONE*, 7(4):e34030.

Elliot Sober. 2004. The Contest Between Parsimony and Likelihood. *Systematic Biology*, 53:644–653.

A Stamatakis, T Ludwig, and H Meier. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–63.

James A Studier and Karl J Keppler. 1988. A Note on the Neighbor-Joining Algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5(6):729–73.

Vicki Tariq. 2013. Importance of Fungi.

Mihoko Tominaga, Yun-hae Lee, Risa Hayashi, Yuji Suzuki, Osamu Yamada, Kazutoshi Sakamoto, Kuniyasu Gotoh, and Osamu Akita. 2006. Molecular Analysis of an Inactive Aflatoxin Biosynthesis Gene Cluster in Aspergillus oryzae RIB Strains. *Applied and Environmental Microbiology*,

72(1):484–490.

J Varga, J C Frisvad, and R a Samson. 2011. Two new aflatoxin producing species, and an overview of Aspergillus section Flavi. *Studies in Mycology*, 69(1):57–80.

Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–7.

James Robert White, Cynthia Maddox, Owen White, Samuel V Angiuoli, and W Florian Fricke. 2013. CloVR-ITS: Automated internal transcribed spacer amplicon sequence analysis pipeline for the characterization of fungal microbiota. *Microbiome*, 1:6.

T J White, T Bruns, S Lee, and J Taylor. 1990. Amplification and Direct Sequencing of Fungal Ribosomal RNA Genes for Phylogenetics. In *PCR Protocols: A Guide to Methods and Applications*, chapter 38, pages 315–322.

Carl R Woese, Otto Kandlert, and Mark L Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, 87(June):4576–4579.

J Yu, P K Chang, J W Cary, M Wright, D Bhatnagar, T E Cleveland, G a Payne, and J E Linz. 1995. Comparative mapping of aflatoxin pathway gene clusters in Aspergillus parasiticus and Aspergillus flavus. *Applied and Environmental Microbiology*, 61(6):2365–71.

Jiujiang Yu. 2012. Current understanding on aflatoxin biosynthesis and future perspective in reducing aflatoxin contamination. *Toxins*, 4(11):1024–57.

Jiujiang Yu, Catherine A Whitelaw, William C Nierman, Deepak Bhatnagar, and Thomas E Cleveland. 2004. Aspergillus flavus expressed sequence tags for identification of genes with putative roles in aflatoxin contamination of crops. *FEMS Microbiology Letters*, 237(2):333–40.

Daniel Zerbino. 2008. Velvet Manual.

Daniel R Zerbino and Ewan Birney. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–9.