Computational Discovery of Metabolic Gene Clusters in Yeast

Christopher Pyatt

A thesis presented for the degree of Doctor of Philosophy

University of East Anglia

Quadram Institute

September 2019

Abstract

Metabolic gene clusters are the genetic source of many natural products (NPs) that can be of use in a range of industries, from medicine and pharmaceuticals to food production, cosmetics, energy, and environmental remediation. These NPs are synthesised as secondary metabolites by organisms across the tree of life, generally to confer an ephemeral competitive advantage. Finding such gene clusters computationally, from genomic sequence data, promises discovery of novel compounds without expensive and time consuming wet-lab screens. It also allows detection of 'cryptic' biosynthetic pathways that would not be found by such screens. The genome sequences of approximately 1,000 yeast strains from the National Collection of Yeast Cultures (NCYC) were searched for both known and unknown metabolic gene clusters, to assess the NP potential of the collection and investigate the evolution of gene clusters in yeast.

Variants of gene clusters encoding popular biosurfactants were found in tight-knit taxonomic groups. The mannosylerythritol lipid (MEL) gene cluster was found to be composed of unique genes and constrained to a small number of species, suggesting a period of substantial evolutionary change in its history. The cellobiose lipid (CBL) gene cluster, conversely, was found to be assembled from members of widespread gene families and is present in at least two widely separated lineages.

The yeast genome dataset was also searched for novel gene clusters using a combination of state-of-the-art software and *ad hoc* methods. A case study of pigmented *Rhodotorula* species suggested a substantial amount of untapped metabolic potential. One software pipeline, FindClusters, has been developed as a method for high-throughput gene cluster variant discovery in assembled genomes. Another, Flagdown, aims to predict gene cluster types missed by existing methods.

The NCYC genomes prove that gene clusters can be found in diverse yeasts, if we only look, offering hope for the discovery of useful compounds.

Contents

List of Figures

List of Tables

Acknowledgements

The work presented in this thesis was supported by a Doctoral Training Partnership (DTP) PhD studentship from the Biotechnology and Biological Sciences Research Council (BBSRC). This research was also supported in part by the NBI Computing infrastructure for Science (CiS) group through use of the high performance computing cluster. I would like to thank my supervisory team, Ian Roberts, Jo Dicks, Steve James, and Adam Elliston, for their invaluable help and guidance along the way. It has been a pleasure working with them for the past four years and I wish them all well for the future. Extra thanks goes to Inge Van Bogaert of Ghent University, whose advice has been extremely useful in helping me to understand the biochemical aspects of this project, particularly with respect to the cellobiose lipid chapter.

On a more personal note, I thank my family and friends for their encouragement, advice, and company throughout this four year journey. Nothing makes hard work feel easier than having people to hang out with and talk about something else!

Finally I would also like to acknowledge and give thanks for the enormous and enduring support of my long-suffering fiancée, Tharsini Sivapalan, without whom I may not have maintained the drive to complete my PhD. Dash has been there for me through thick and thin, and always encouraged me to try my best and believe in myself. It's thanks to her that I am where I am today, and I hope she knows how much she means to me.

1 Introduction

This thesis describes work to find and categorise particular metabolic gene clusters in a large new dataset of sequenced yeast genomes, as well as groundwork for the development of new methods of gene cluster discovery in such data-sets. In this chapter I seek to introduce the concepts of metabolic gene clusters and their evolution, as well as current computational methods for their discovery and analysis.

1.1 Metabolic gene clusters

A metabolic gene cluster, for the purposes of this thesis, is defined as a collection of functionally related (i.e. part of the same biosynthetic pathway) genes located adjacent to one another in a relatively compact region of the genome. The term "gene cluster" is also applied in the wider literature to collections of orthologous genes that have arisen through gene duplications followed by divergence in function, for example the Hox gene cluster in vertebrates, and to simple syntenic blocks of genes that are neither functionally nor structurally related but are nevertheless conserved across species. These latter two types are not the subject of this thesis and the term "gene cluster" should always be taken to mean the genetic constituents of a collocated biosynthetic pathway.

Gene clusters are ubiquitous constructs in prokaryotes, as is the operon, a similar structure where the entire region is transcribed into a single polycistronic mRNA encoding multiple proteins. Polycistronic mRNAs and operons are unknown in most eukaryotes, with the exception of *Caenorhabditis elegans* (Lawrence 1999) and some early branching eukaryotes including trypanosomes (Clayton 2019). It may be that polycistronic mRNA is an ancestral characteristic of the wider eukaryotic lineage that has been lost in the branches leading to the more familiar members (plants, fungi, animals). Alternatively the phenomenon may just be very rare in these higher eukaryotes and thus as yes undiscovered. Metabolic gene clusters of the type discussed here have been reported in plants (for example gene clusters producing avenacin, thalianol, momilactones, and phytocassane in oat, *Arabidopsis*, and rice, respectively (Shimura et al. 2007; Osbourn 2010)), animals, and fungi. In the latter, most

work has focused on filamentous fungi rather than yeast or yeast-like fungi.

Gene clusters are of interest due to the fact that they encode a wide range of secondary metabolites (biosynthesised compounds that are not essential to life but confer an adaptive advantage under certain conditions, for example an antibiotic compound that neutralises competitors in nutrient-poor or crowded conditions) that may be useful to a wide range of industrial, medical, and environmental applications. In addition the evolutionary mechanisms for their formation, particularly in eukaryotes, are poorly understood. Secondary metabolites are frequently discovered through screening analyses aiming to categorise and quantify the biosynthesis complement of bacterial or fungal strains. However since these compounds are, by their nature, only expressed under certain conditions, the screens cannot pick up all metabolic potential without very substantial and imaginative (and expensive) culture condition variation. For this reason it would be advantageous to be able to predict potential secondary metabolite gene clusters purely from the sequenced genome. These predictions could then act as markers to signal biosynthetic pathways that warrant further exploration. Ribosome engineering and rare earth metals have been shown to induce or increase natural product biosynthesis, even from cryptic clusters (Ochi et al. 2013). This could be a useful next step once these cryptic clusters are identified bioinformatically.

1.2 Known fungal gene clusters

Mannosylerythritol lipids

Mannosylerythritol lipids (MELs) are a class of biosurfactants (biologically synthesised surface-active compounds) produced by various members of the Ustilaginales (smut fungi in the Basidiomycota). The group (of MEL producers) was formerly known as the genus *Pseudozyma* but has since been reclassified as representing polyphyletic branches within the larger taxonomic group (Wang et al. 2015). MELs have wide-ranging applications including in cosmetics, environmental bioremediation, drug delivery, treatment of tumours and dopamine disorders, as well as displaying antibiotic, antifungal, and self-assembling properties relevant to foaming, emulsification and vesicle formation. The advantages of MELs versus conventional synthetic surfactants is that they are more environmentally friendly, being both biodegradeable and less toxic, and require comparatively mild and inexpensive production conditions (Yu et al. 2015). On the other hand, the conventional alternatives are often highly toxic to soil and aquatic ecosystems (Emmanuel et al. 2005), dermatologically irritating (Rodrigues et al. 2006), and produce harmful by-products during manufacture (Makkar et al. 2002) and when broken down (Scott et al. 2000). Additionally, biosurfactants can be effective

at wider temperature, salinity, and pH ranges, and at lower concentrations (Kitamoto et al. 2009; Fan et al. 2014).

MELs are composed of a mannosylerythritol disaccharide which is acylated and acetylated at the mannose moiety (Hewald et al. 2006). There are four main types of MEL, distinguished by the level of completion in terms of acetylation (see Figure 1.1). Three rarer configurations have also been described. These are a tri-acetylated form (Fukuoka et al. 2007c) with an additional acyl group on the erythritol moiety, a mono-acylated/tri-acetylated form (Morita et al. 2011) with the acyl group at C2 replaced by a third acetyl group, and a mono-acetylated form (Fukuoka et al. 2007a) where the C2 acyl is absent.

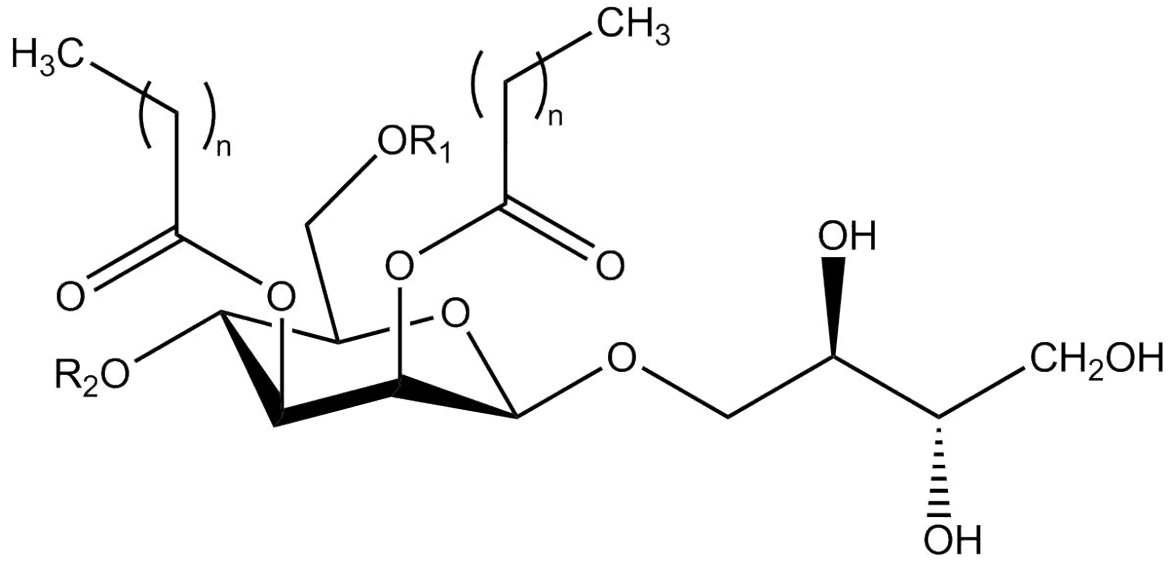

Figure 1.1: Mannosylerythritol lipid. Positions R_1 and R_2 may be acetylated depending on the type of MEL. MEL-A: R_1 & R_2 = acetyl, MEL-B: R_1 = acetyl & R_2 = H, MEL-C: R_1 = H & R_2 = acetyl, MEL-D: R_1 & R_2 = H

MELs are synthesised in all known cases by the action of a gene cluster similar to that characterised in *Ustilago maydis* (Hewald et al. 2006), see Figure 1.2. This gene cluster is made up of five genes spanning approximately 18Kb of chromosome 7. The five genes are *mat1*, encoding an acetyltransferase, *mmf1*, a member of the major facilitator superfamily, *mac1*, an acyltransferase, *emt1*, a glycosyltransferase, and *mac2*, another acyltransferase. Differences between the main types of MEL are likely due to differences in the activity of either the acetyltransferase *mat1p* or the transporter *mmf1p* as these are the final steps in the pathway.

Cellobiose lipids

Figure 1.2: MEL gene cluster as seen in *U. maydis*. Each coloured arrow represents a gene, with the relative orientation indicated. Gaps between arrows are not meant as a representation of distance.

The cellobiose lipids (CBLs) are another class of biosurfactants with a central cellobiose sugar moiety around which the fatty acid chains are arranged. There are two primary variants known in fungi at this time, ustilagic acid (UA - found in *U. maydis*) and flocculosin (*Pseudozyma flocculosa*). As is to be expected, the basic skeleton of both compounds is a cellobiose disaccharide linked to various fatty acid side chains. The structures of UA and flocculosin are shown in Figures 1.3 and 1.4. Both have similar antifungal properties and other potential uses in line with those of MELs, above. Flocculosin in particular is used to control powdery mildew pathogens (Teichmann et al. 2011).

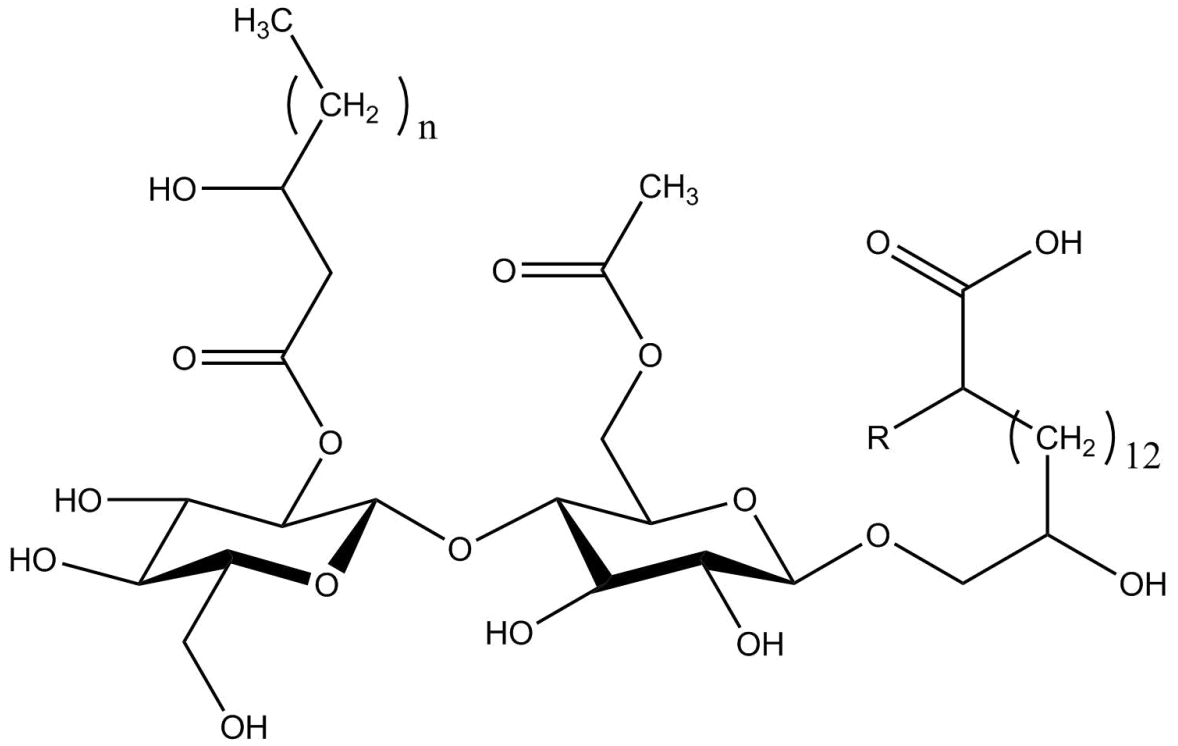

Figure 1.3: CBL ustilagic acid, produced by *Ustilago maydis*. The R group may represent differing length fatty acid chains.

The gene clusters responsible for CBL production in *U. maydis* and *P. flocculosa* are shown in Figure 1.5. That encoding UA is made up of twelve genes spanning ~45Kb of

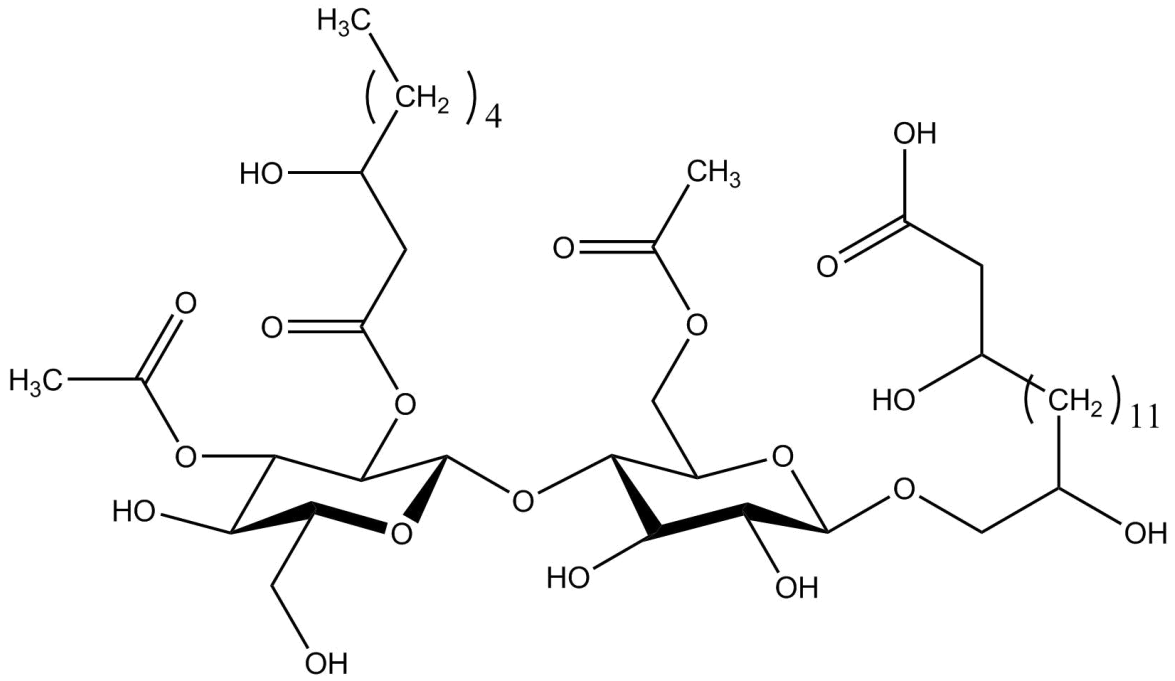

Figure 1.4: CBL flocculosin, produced by *Pseudozyma flocculosa*.

chromosome 23 (Teichmann et al. 2007, 2011) while that of flocculosin is comprised of eleven genes over ~60Kb. The genes involved are largely the same, albeit in a different configuration, with the exception of *orf2* and *ahd1* being unique to *U. maydis*, and *fat3* being unique to *P. flocculosa*. The ORF in the flocculosin gene cluster seems to be homologous to *orf1* of the UA gene cluster. Although other CBL producers have previously been reported, the existence or configuration of any underlying gene cluster in other species is unknown. In any case, the variation displayed between the two known CBL gene clusters invites investigation.

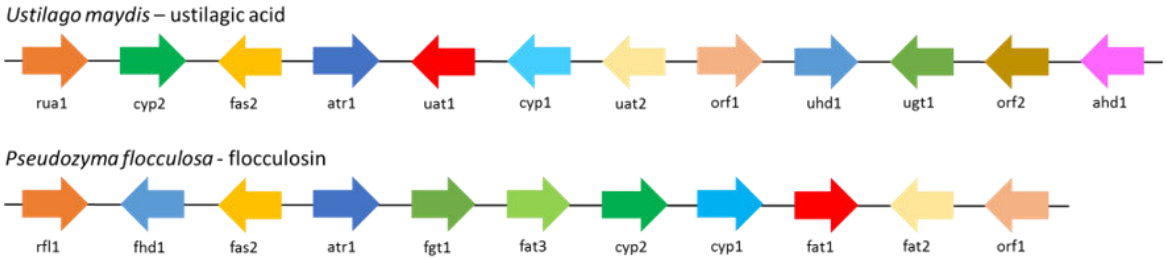

Figure 1.5: CBL cluster as seen in *Ustilago maydis* and *Pseudozyma flocculosa*. The ustilagic acid pathway has 12 genes, while the flocculosin pathway has 11. Arrows representing genes are colour coded for comparison between clusters and indicate orientation. Gaps between arrows are not meant as a representation of distance.

Sophorolipids

A third class of biosurfactants produced by fungi are the sophorolipids. These are secreted by a number of non-pathogenic yeast species including *Starmerella bombicola*, *Candida apicola* and relatives, and others such as *Wickerhamiella domercqiae*, *Pichia anomala*, and *Rhodotorula bogoriensis* (Van Bogaert et al. 2011). As with the other biosurfactants described previously, these have a number of potential applications in cleaning, food, cosmetics and pharmacy due to their emulsification and dispersal activities. These compounds are composed of a hydrophilic sophorose disaccharide and a hydrophobic fatty acid chain. Structural variations can take the form of acetylation on the sophorose part or lactonisation between the sophorose and the terminal end of the fatty acid, see Figure 1.6. As with MELs these structural modifications bring with them useful changes in the physiochemical properties of the molecule and the producing organisms each produce different types (Van Bogaert et al. 2011).

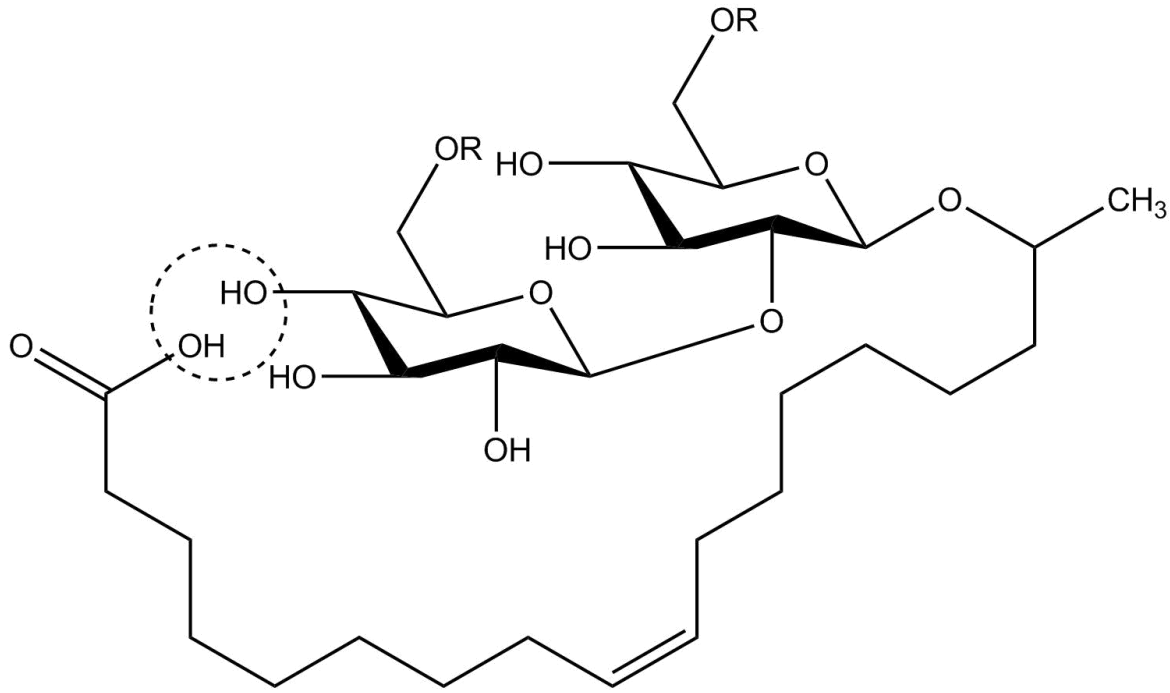

Figure 1.6: Sophorolipid. Positions marked by R are possible acetylation sites (otherwise H). Lactonisation may occur between the positions marked by the dashed circle.

Sophorolipids are synthesised by a gene cluster composed of at least five genes, fully characterised in *S. bombicola* by Van Bogaert et al. (2013), spanning roughly 12Kb. The five known genes are, in order of genomic appearance, *ugtB1*, the second glucosyltransferase in the pathway, a transporter *mdr*, an acetyltransferase *at*, *ugtA1*, the first glucosyltransferase, and a cytochrome P450 monooxygenase *cyp52m1*. There is also one modifying enzyme (a lactonesterase - *le*) outside of the cluster, further along the same chromosome (Roelants et al. 2014; Jezierska et al. 2018). See Figure 1.7 for a diagram of the gene cluster configuration.

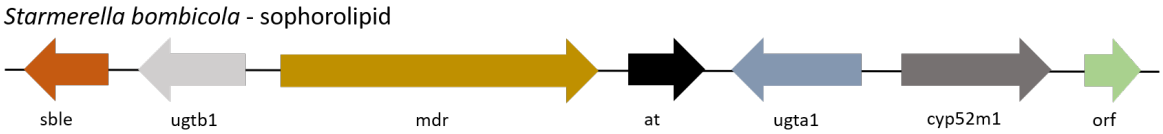

Figure 1.7: Sophorolipid gene cluster. Coloured arrows represent genes and their orientation. Gaps between arrows are not meant as a representation of distance. The lactonesterase *sble* is approximately 2.5Mb away (not clustered, denoted by red mark) but part of the biosynthetic pathway. The ORF is genomically associated with the gene cluster but it is not known whether it is related to the biosynthetic pathway.

Others

Other gene clusters have been described in fungi, some of which I will briefly mention here. The DAL gene cluster allows *Saccharomyces cerevisiae* to acquire nitrogen from allantoin and is one of very few examples of known gene cluster organisation in the small and simplified genomes of *Saccharomyces* (Wong et al. 2005). The GAL gene cluster is a relatively widespread Ascomycete gene cluster (apparently the pathway is independently assembled in *Cryptococcus* (Slot et al. 2010)) enabling use of galactose. Both these metabolic pathways are found in clustered (see Figure 1.8) and unclustered forms (Slot et al. 2010).

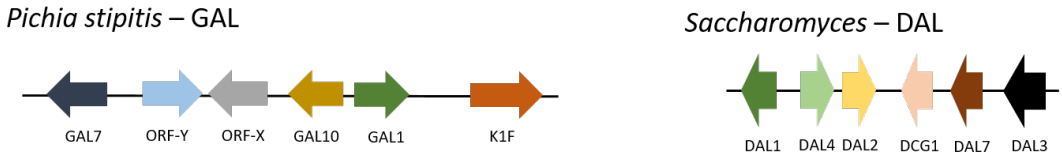

Figure 1.8: The clustered forms of the DAL and GAL metabolic pathways. In the case of the GAL gene cluster, other forms exist, for example a partially clustered pathway in *S. cerevisiae* and an independently assembled gene cluster in *Cryptococcus* (Slot et al. 2010).

Returning to secondary metabolism, filamentous Basidiomycetes such as *Aspergillus fumigatus*, *Aspergillus flavus*, and their relatives produce mycotoxins such as gliotoxin (Gardiner et al. 2005) and aflatoxin (Cary et al. 2006; Roze et al. 2015) from metabolic gene clusters, see Figure 1.9. Similarly the *Rhodotorula* produce carotenoids that give them their distinctive red colouration using, at least in part, the CAR gene cluster (Landolfo et al. 2018), see Figure 1.10. Other fungal gene clusters include those synthesising pulcherrimin in *Kluyveromyces lactis* (and possibly in other *Saccharomycotina* species) (Krause et al. 2018), tricothecenes in *Fusarium* & *Myrothecium*, sterigmatocystins in *Aspergillus*, melanin in *Alternaria* (Keller et al. 1997), kojic acid in *Aspergillus oryzae* (Takeda et al. 2014), penicillin in

Aspergillus nidulans (Shaaban et al. 2010) and *Penicillium chrysogenum* (Keller et al. 1997), and cephalosporins in *Acremonium chrysogenum* (Keller et al. 1997).

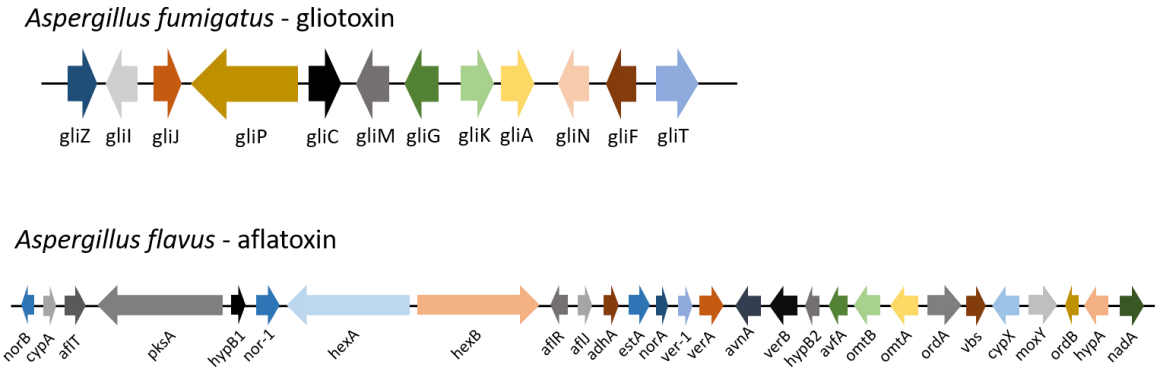

Figure 1.9: Gene clusters for gliotoxin and aflatoxin as found in *Aspergillus fumigatus* and *Aspergillus flavus* respectively. Gaps between arrows are not meant as a representation of distance.

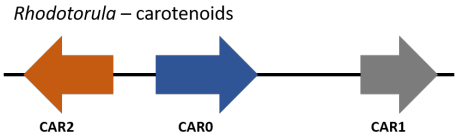

Figure 1.10: CAR gene cluster producing pigmentation carotenoids such as β-carotene. Gaps between arrows are not meant as a representation of distance.

1.3 Evolution and maintenance of gene clusters

Clustered biosynthetic pathways are ubiquitous in prokaryotes in the form of operons. In this form, each pathway is co-transcribed as one long polycistronic mRNA that is then used to make each gene product in turn. The reason for clustering in prokaryotes is therefore fairly obvious; it is required for co-expression of the genes and the biosynthesis of a complex metabolite (clustering also helps with rapid acquisition of new capabilities via horizontal gene transfer). Eukaryotic gene clusters work in the same way as normal eukaryotic biosynthetic pathways, that is each gene is transcribed and translated independently, just with the additional feature of being genomically close to one another. The reason for such clustering is less clear than in prokaryotes because trans-acting regulators are the norm in eukaryotes and therefore spatial clustering is not necessary for co-expression. Several models concerning the evolution and subsequent maintenance of metabolic gene clusters in eukaryotes have been proposed over the years and the main ones will be discussed here.

Selection for proximity of functionally related genes

The Fisher Model suggests that clustering of genes making up a biosynthetic pathway, or that are otherwise functionally linked, may be selected for as this reduces disruption by recombination (Lawrence 1999). The story of the DAL cluster, described above, illustrates this nicely in that it has been assembled recently via a series of genomic rearrangements (Wong et al. 2005). The cluster was seen in various degrees of completeness in related yeasts, moving through the taxonomy to the point of full clustering in *S. cerevisiae*. Genes were either moved or copied from their original locations to form the gene cluster, and this process of gene cluster development may still be in progress given that several pathway-related genes and transcription factors are still located distally within the *S. cerevisiae* genome.

Elsewhere, Slot et al. (2010) used ancestral state reconstructions to determine that the GAL gene clusters, also described above, in the *Saccharomyces* and *Cryptococcus* lineages were separately and independently assembled through gradual accumulation of pathway genes at one locus. This model of gene cluster assembly is also supported by analyses conducted outside of yeasts, for example the filamentous *Fusarium*. Genes encoding the tricothecene pathway are clustered to various degrees in the different clades of the genus. Proctor et al. (2009) showed that two of the pathway genes had been moved into the gene cluster from distant genomic loci.

Gene proximity allows co-regulation by chromatin modifications. In plants, clustered genes are frequently delineated by signature chromatin marks also seen at the human MHC cluster (Yu et al. 2016). Modifications are also evident at the site of the DAL gene cluster in yeast (Wong et al. 2005) and around secondary metabolite pathways in filamentous fungi (Shaaban et al. 2010). This marking may turn out to be another way to identify gene clusters, although it requires more information than is normally available from pure genome sequencing. Proximity of functionally related genes, i.e. gene clusters, may also facilitate horizontal transmission of entire biosynthetic pathways. In organisms where horizontal transfer is common, this may represent an additional selection pressure towards clustering. See also the "Selfish Operon Model" of Lawrence et al. (1996). This same selection pressure would also push towards a more condensed gene cluster (with less intergenic sequence) as the longer the chunk of DNA in question, the higher its chances of being split up.

Horizontal transfer from prokaryotes or other organisms

Horizontal gene transfer (HGT) between prokaryotes and eukaryotes has been documented in the past but tends to involve the transfer of just a single gene. The number of HGT events between eukaryotes is lower, especially for multi-gene events (Slot et al. 2007). Examples of HGT from prokaryotes include the apparent prokaryotic origins of the penicillin and cephalosporin pathway clusters mentioned earlier (Keller et al. 1997). The Keller paper also points out however that few fungal metabolic pathways have such clear prokaryotic counterparts. A potential example of HGT from another eukaryote is the purported transfer of the GAL gene cluster from *Candida* to *Schizosaccharomyces* (Slot et al. 2010). Of course this does not explain how the gene cluster originally came into being, merely how it came to occur in the latter genus.

Continuous horizontal transfer

Once a gene cluster has been formed, it is subject to forces serving to break it apart, i.e. recombination, or otherwise to deactivate it. Therefore for a gene cluster to persist it must counteract these forces. In particular, secondary metabolite gene clusters are at higher risk of being lost since they are not essential under all circumstances yet present a potential cost to the organism. One of the ways this could happen is if there is a significant amount of HGT between sub-populations or co-occurring species. This would maintain the gene cluster in the population despite pockets of local loss.

HGT is thought to be rare in eukaryotes but is occasionally observed. Even whole gene clusters have been seen to move in this way (Patron et al. 2007; Khaldi et al. 2008). Other evidence of HGT between eukaryotes was seen by Slot et al. (2007) regarding the sterigmatocystin gene cluster in relatively distinct fungal genomes. In that case there was notably greater synteny within the gene cluster than in the genomes as a whole, as well as phylogenetic evidence suggesting an HGT event. If such a large gene cluster can be successfully translocated between species, then HGT must be a viable method of gene cluster proliferation. Whether it actually occurs frequently, however, is unclear.

Toxic pathway intermediates

Frequently, a metabolic gene cluster may occur on several non-adjacent branches of a clade due to multiple total gene cluster losses in intermediate species. Secondary metabolite pathways often traverse intermediate biosynthetic steps resulting in "toxic intermediate" compounds that, if allowed to accumulate, would harm the producing organism. There may therefore be strong selection pressure to either keep the entire pathway intact or do away

with it entirely (Khaldi et al. 2008). As an example of the latter, Slot et al. (2010) reported that the GAL metabolism pathway is more likely to be lost when the genes are clustered, since this results in the loss of the entire pathway and does not leave the potential for toxic intermediates. In a similar vein, Hittinger et al. (2004) showed that the GAL pathway in *S. kudriavsevii* was inactivated very rapidly (i.e. all seven genes acquired nucleotide changes that rendered them nonfunctional), perhaps under selective pressure due to toxic intermediate compounds. Conversely, Wong et al. (2005) have suggested that the clustering of the DAL pathway has been maintained due to the DAL3 product being toxic to yeast and removed later in the pathway by DAL7. Keeping DAL3 and DAL7 linked is therefore beneficial.

Evidently this balance can tip either way, either acting to prevent gene cluster dispersal (i.e. retaining the pathway but losing the cluster dynamic) or acting to remove the gene cluster entirely if selected against. It may be that both situations occur in different circumstances, depending on environmental pressures.

Co-regulation of gene cluster components

Having all the genes in a biosynthetic pathway located in close proximity to one another has been proposed to be advantageous from a co-regulation efficiency point of view (Walton 2000). This is particularly relevant to *cis*-acting transcription factors. Shaaban et al. (2010) and Osbourn (2010) have also suggested that chromatin-mediated co-regulation is improved by gene cluster organisation in sub-telomeric regions.

The co-regulation theory seems to be more acceptable when talking about bacterial gene clusters, since regulation there is *cis*-acting while eukaryotic regulation is capable of acting in *trans*. There are many examples of eukaryotic pathways being co-regulated without the need for physical proximity (Walton 2000; Lee et al. 2003). If co-regulation is achieved via local transcription factors or chromatin modification then this may add weight to the co-regulation/co-localisation theory. With this in mind, this particular selection pressure may not act to encourage gene cluster formation, yet may still act later as a mechanism to maintain gene cluster contiguity.

1.4 Gene cluster discovery

There are two fundamental strategies for finding gene clusters in genomic data. Firstly one can look for homology with known gene clusters in new species/ datasets in order to find variants with potentially useful different properties. For example searching for the MEL gene

cluster, described above, in a variety of species may lead one to species producing some of the rarer MEL types which have unique physiochemical properties and therefore potentially novel applications. The second strategy to employ is to attempt to find entirely new gene clusters by trawling genome sequence data for sequence, gene, or domain patterns that may indicate the presence of a clustered biosynthetic pathway.

Homology with known gene clusters

This is the simplest way to find variants of known gene clusters. The technique relies on there being a reasonably well annotated reference gene cluster for which to search. Of course it cannot find totally new gene cluster types as there is nothing on which to base the search. Sequence matching tools such as BLAST (Altschul et al. 1990) and HMMER (Eddy 1998, 2015) are often put to use for this purpose, finding variants of known genes in public or custom databases (or genome assemblies).

As an example, Khaldi et al. (2008) used BLASTp (protein-protein matches) to search for the fifteen genes of the ACE1 gene cluster (as found in *Magnaporthe grisea*) in 26 fungal genomes, finding 9 previously unknown partial gene clusters. Similarly, the flocculosin gene cluster described previously was discovered by searching the *P. flocculosa* genome for homologues of the *U. maydis* ustilagic acid gene cluster genes (Teichmann et al. 2011). Later in this thesis, I will describe a computational pipeline built around the HMMER toolset that can execute such a homology search for a whole gene cluster in large multi-genome datasets (e.g. the NCYC genome sequencing project), without having to search for each gene individually and then manually piece together the cluster composition for each species.

de novo gene cluster prediction

In a similar vein to *de novo* gene prediction, where patterns of codon usage and splice site signatures are used to identify possible coding sequences, patterns of gene placement and functional types may be used to find novel gene clusters. For example, genes encoding metabolite-synthases occurring in close proximity to those for structure-modifying enzymes, regulatory proteins and transporters (Proctor et al. 2009) could constitute a potential gene cluster. Additionally there may be duplicated copies of primary metabolism genes, a common origin of secondary metabolite genes, or simply syntenic blocks shared across genomes (Medema et al. 2015). In the latter case this may point to regions experiencing selection to prevent chromosomal rearrangement, possibly due to the issues discussed in the previous section.

Common components of natural product pathways in bacteria are non-ribosomal peptide synthases (NRPS), polyketide synthases (PKS), and terpenoid genes. The compounds built around these backbone proteins are used by micro-organisms to inhibit competitors (Li et al. 2009) and therefore are often useful as antibiotics. This is probably true of a lot of secondary metabolites. NRPS & PKS products in particular are often important in clinically used pharmaceuticals (Ziemert et al. 2012). It has also been suggested that transposable and transposon-like elements may act as flags at the borders of gene clusters (Shaaban et al. 2010), which may prove useful in searching for novel gene clusters.

Many computational tools exist already for gene cluster discovery, leveraging both strategies mentioned here. Most were originally conceived with prokaryotes in mind and have been coded or trained with bacterial data, however they may work just as well in yeasts. Recently more progress has been made in adapting some of the existing tools to work with eukaryotic genomes, and there are also methods in the pipeline that aim to take advantage of modern computational techniques such as machine learning as well as relying on the original pattern/rule based search strategies. Some of these tools are described below to give an overview of the current landscape. The following is based on the review of Medema et al. (2015), with a few more recent approaches added.

High confidence/ low novelty

The tools described below use methods that rely mostly on recognition of known patterns, thus providing high-confidence predictions of gene cluster locations and configurations. They suffer from limitations imposed by these methods, primarily that unknown patterns are missed by definition.

antiSMASH Medema et al. (2011), Weber et al. (2015), and Blin et al. (2017). Uses a library of Hidden Markov Model (HMM) profiles assembled from alignments of known signature cluster genes to identify regions of a genome containing multiple matches in close proximity. Also captures accessory genes by extending each putative cluster by several thousand base-pairs from the terminal signature genes. Included in the output is a functional annotation of discovered clusters including predicted substrates. Capable of predicting clusters in both bacterial and eukaryotic genomes. This is the state of the art 'go-to' software for gene cluster prediction, with an increasing number of developers involved and contributions from many research groups aiming to refine the search rules. Other tools are often integrated into the antiSMASH package as optional extras.

SMURF Khaldi et al. (2010). Uses HMMs to identify backbone genes such as non-ribosomal

peptide synthases (NRPS) or poly-ketide synthases (PKS) and common "decorating genes" such as transporters and regulators. It then scans the genomic region surrounding the backbone genes for domains commonly found in gene clusters. Works with fungal genomes.

np.searcher Li et al. (2009). Identifies potential clusters and offers predictions of the chemical structure of the products. Focusses on NRPS and PKS clusters, searching for acyltransferase and adenylation domains within translated genomic sequences and then adding any adjacent modifying domains. Additionally detects trans-acyltransferase PKS and terpenoid clusters but cannot predict their products.

ClustScan Starcevic et al. (2008). Uses HMMER (Eddy 1998, 2015), GeneMark (Besemer et al. 2005), and Glimmer (Delcher et al. 2007) to predict genes based on profiles of known proteins (the default being a set of NRPS/PKS profiles, with the addition of user-generated profiles if desired) and then allows the user to manually define a cluster. Suggests possible biosynthetic orders based on the chosen cluster constituents. Designed for bacterial genomes but is also useful in lower eukaryotes such as slime moulds.

CLUSEAN Weber et al. (2009). Uses BLAST (Altschul et al. 1990) and HMMER (Eddy 1998, 2015) searches against general databases to identify and annotate biosynthetic clusters in bacterial genomes.

MultiGeneBlast Medema et al. (2013). Performs multiple BLAST (Altschul et al. 1990) searches of all gene sequences in the GenBank database (formatted database supplied with programme and updated regularly) to find clusters homologous to the input cluster (specified either as a genomic locus or a list of genes) and contained within a defined distance of each other. Other features include searching a user-generated gene database (in the correct format), searching un-annotated genome sequences via user-generated nucleotide databases, and using as input a user-generated series of protein sequences (not already known as a cluster) that imply a pathway containing specified biosynthetic steps.

NaPDoS Ziemert et al. (2012). Uses BLAST (Altschul et al. 1990) and HMMER (Eddy 1998, 2015) to search genomes for conserved domains associated with NRPS and PKS products. Created to search for individual genes and classify products based on phylogeny. Has potential for use within a larger pipeline as conserved domains may act as flags marking the location of biosynthetic gene clusters. Low stringency results in a high number of false positives (often in the form of fatty acid synthases as opposed to secondary metabolite enzymes). Optimised for bacterial genomes although the reference

database is being expanded to include more eukaryotic sequences.

eSNaPD Reddy et al. (2014). Searches raw metagenomic sequence reads for matches to a reference database of known metabolic gene clusters. Combines sequence analysis with geographic data in an effort to guide further metagenomic environmental sampling to discover novel natural products.

Low-confidence/high-novelty

The tools in this section are the opposite of those above. They use methods that are more open to new unknown patterns and therefore have a better chance of discovering novel types of gene cluster. They are however limited by generally higher false-positive rates due to their more speculative nature and require more post-process validation.

ClusterFinder Cimermancic et al. (2014). Uses an HMM to hunt for genomic regions containing high frequencies of Pfam domains found in known gene clusters. Secondary metabolite pathways often contain regulators and transporters belonging to larger domain families so a high frequency of these genes in a region may indicate the presence of a biosynthetic gene cluster. Something of a middle-ground between the two gene cluster-finding methodologies as it can find clusters whose internal biosynthetic genes are not similar to anything already known, but still requires some degree of similarity in terms of transporter and regulator genes. This algorithm was integrated into version 3.0 of antiSMASH (Weber et al. 2015). Apparently has a high false positive rate due to probability based algorithm.

EvoMining Cruz-Morales et al. (2015) and Sélem-Mojica et al. (2018). Searches for diverged paralogues of primary metabolism genes under the premise that this is a common source of secondary metabolism genes. Developed for bacterial genomes. Version 2.0 released in late 2018 with some extra features and user interface.

Motif-Independent Prediction Takeda et al. (2014). Identifies syntenic blocks of genes within different genomes that retain similar gene composition despite divergence across the genome as a whole. May find clusters that do not share "signature genes" with any other known pathways. Works in filamentous fungi.

A notable limitation in current methods that has yet to be solved is the difficulty in accurately predicting gene cluster boundaries, i.e. which genes are or are not included in the pathway either side of the principal biosynthetic backbone. I am aware of two more methods, both in their infancy, attempting to improve this. One is a machine learning approach for

classifying biosynthetic gene clusters by protein domains (Hayda Almeida - Personal Communication). This has, so far, been applied only to a test dataset of Aspergillus gene clusters. The other is a deep-learning approach to identifying gene clusters using protein family domain content (DeepBGC) which purports to outperform ClusterFinder (Christopher Woelk - Personal Communication). This method has only been tested on bacterial genomes to date. Finally there is also a project at the Institute of Pharmacological Science, Germany, aiming to predict substrates and product structures of metabolic gene clusters (Stefan Gunther - Personal Communication).

1.5 Project Aims

The aim of this project is to use a combination of published tools and homegrown computational methods to investigate the metabolic gene cluster content of the National Collection of Yeast Cultures (described in the following chapter). This is facilitated by an ongoing effort to sequence the genomes of the entire collection. At present almost 1,000 genomes (979) have been sequenced and assembled, representing roughly 200 (of 530 in the collection) yeast species. Exploration of this large and diverse dataset has the potential to uncover a wealth of interesting compounds, or at least lay the groundwork for such discoveries. Yeasts have the additional advantage that they are fairly simple to grow in large batches, often on waste materials as energy sources, making them amenable to industrial use in natural product production.

An additional aim of the project is to describe the evolutionary history of gene clusters in yeast, through phylogenetic analysis of known biosynthetic pathways in well sampled lineages (chosen due to the availability of relevant strains, and published descriptions of the gene clusters). Finally, the project aims to examine the utility of third-party methods in finding gene clusters in yeast genomes and, if necessary, develop reliable computational methods for the discovery and analysis of gene cluster variants, and novel gene clusters, in unexplored (yeast) genomic sequences. This will draw on existing knowledge of common backbone genes frequently found in gene clusters.

1.6 Summary of Thesis

This thesis will first describe the National Collection of Yeast Cultures (NCYC) and the genome sequencing project currently underway, as well as some exploratory phylogentic analysis of the collection. Chapters 3 and 4 discuss the search within the sequenced NCYC genomes for the mannosylerythritol lipid and cellobiose lipid gene clusters, respectively, as

well as the results of evolutionary analysis of those gene clusters. Chapter 5 regards the search for novel cytochrome P450 genes within the sampled yeast genomes, and their utility within the field of natural product research, while Chapter 6 describes the production and evaluation of novel software intended to complement the growing field of bioinformatic gene cluster discovery.

2 Background work - sequencing and phylogenetics

2.1 The National Collection of Yeast Cultures

The National Collection of Yeast Cultures (NCYC, see `www.ncyc.co.uk`) was set up in 1951 with the primary aim of storing and maintaining cultures of diverse yeast strains. Given the widespread use of baker's and brewer's yeast in a broad range of industries and in academia, a large part of the collection is given over to strains of *Saccharomyces cerevisiae*. Today, as then, some of the main activities undertaken by the NCYC include identification, storage, and supply of specific yeast strains to industrial and academic customers, all on a commercial basis. Since 1980 the collection has been housed within the Institute of Food Research, and now in its successor the Quadram Institute, on the Norwich Research Park in Norwich, UK. In addition to the *Saccharomyces* cultures, the NCYC hosts several hundred (~530 in total) species of generally non-pathogenic yeasts, isolated from a wide array of environments around the world. Today the total number of strains held exceeds 4,000.

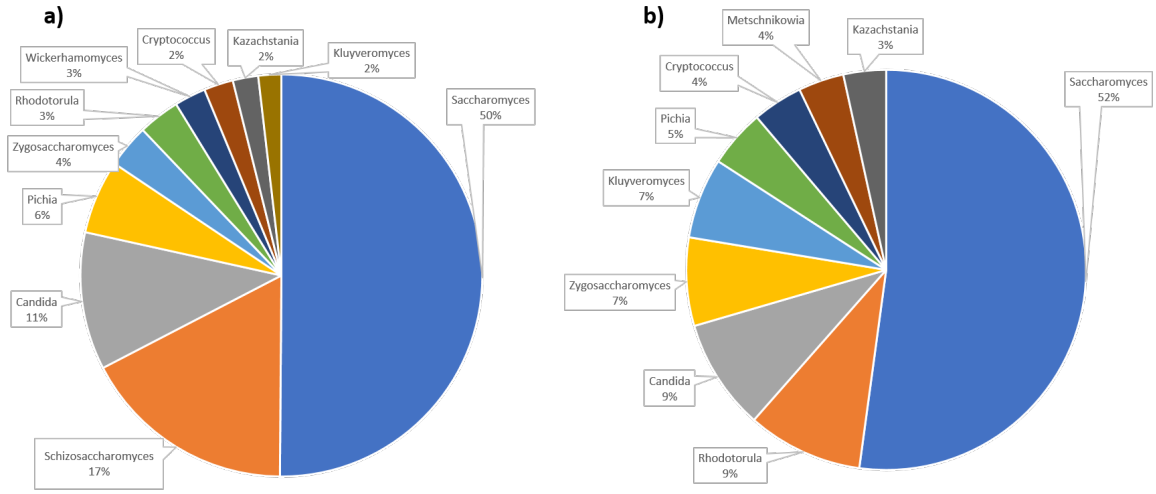

Figure 2.1: Top ten genera represented in (a) the whole NCYC collection, and (b) the sequenced genomes. Figure generated by Dr Jo Dicks (NCYC).

2.2 Sequencing the NCYC collection

Since 2015, the NCYC has sequenced the genomes of almost 1,000 strains from the collection, with representatives from 200 species. The genomes of 967 strains (944 from the NCYC and 23 from other collections) were sequenced over eleven sequencing plates by the Earlham Institute (plates 1-3, 7-9), Eurofins (plates 4-6, 10), and the Wellcome Trust Sanger Institute (plate 11). The sequencing took place concurrently with this project (2015-2019), and was overseen by Dr Jo Dicks (Quadram Institute). Table A.1 contains details of the technologies used, while Tables A.2 and A.3 show more detail on the strains themselves.

2.3 Phylogenetics of the NCYC collection

There are estimated to be several thousand species of yeast across the globe and these are divided into two major groups: the monophyletic group containing the brewing and baking yeasts, within the Ascomycota phylum, and the paraphyletic assortment found within the Basidiomycota phylum. The latter are often referred to as 'yeast-like' fungi due to the fact that they resemble brewing yeasts in lifestyle (single celled, non-hyphal growth with small simplified genomes and metabolism) despite not being part of the same taxon. The Basidiomycete yeasts are spread across the phylum rather than grouped together, having presumably evolved the 'yeast-like' life cycle independently. Some species can switch between the two growth strategies. The NCYC collection, and in particular the set of strains selected for sequencing, includes strains from across this great diversity, with an Ascomycete to Basidiomycete ratio of approximately 9:1.

The raw sequencing reads produced by the NCYC genome sequencing project were preprocessed by Dr Jo Dicks. Firstly, regions of low quality and any remaining adapter sequences were removed using Trimmomatic v0.32 (Bolger et al. (2014) - default parameters and adapter sequence files relevant to the sequencing library used for each plate). Draft sequence assemblies were subsequently estimated from the paired trimmed reads using ABySS v1.9.0 (Simpson et al. 2009), with the -k=80 option.

The assembled genomes of the 967 strains were used to construct a phylogenetic tree. This phylogeny will be used in later projects in order to apply taxonomic context to phenotypic datasets. The tree seen in Figure 2.5 was constructed from the D1/D2 region of the large subunit (LSU) of the 26S ribosomal DNA. D1/D2 sequences were collected from all assembled NCYC genomes using HMMER v3.1b2, with the *Ustilago maydis* D1/D2 sequence as the base query, and retrieved using the mapCoordinates.py script described later in chapter 6. The

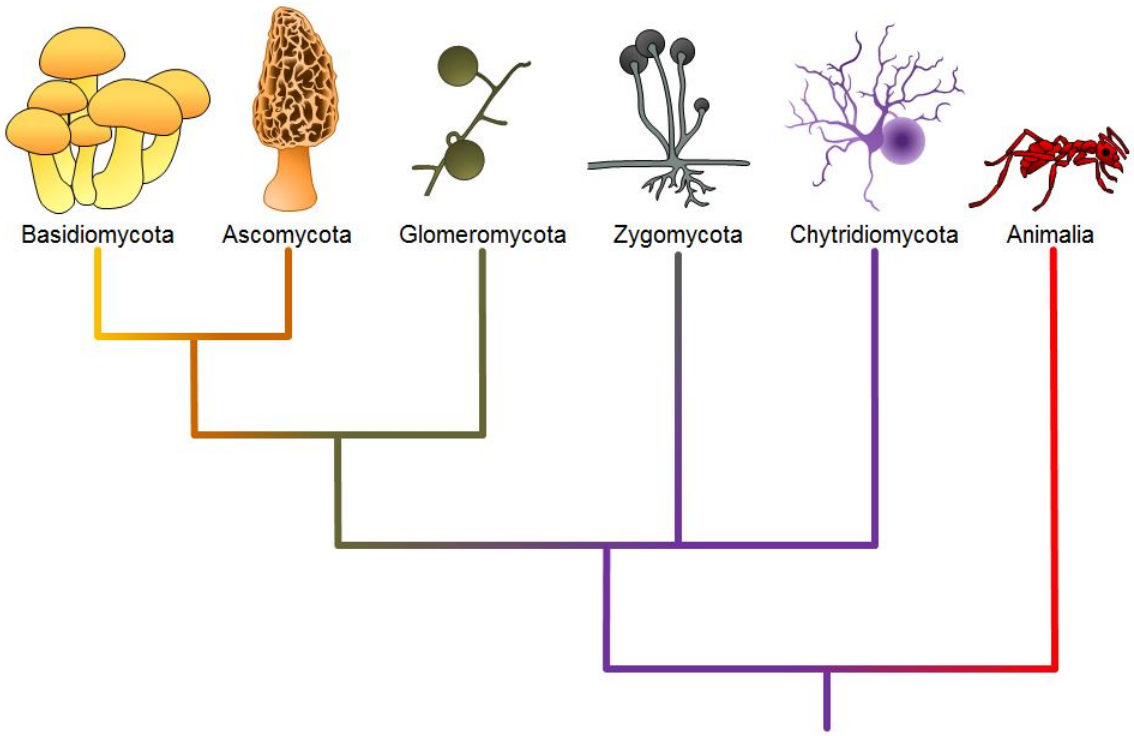

Figure 2.2: Overview of fungal phyla. Yeasts are found in the Ascomycota and Basidiomycota. Taken from `https://courses.lumenlearning.com/suny-osbiology2e/chapter/classifications-of-fungi/`.

tree was estimated using RAxML (Stamatakis 2014) with 100 bootstraps and model set to GTR + GAMMA (other options: -f a, -x 12345, -p 12345). Model selection was performed with PartitionFinder2 (options: model_selection=BIC, models=all, search=greedy, Lanfear et al. (2017)). Sequence alignments were constructed and trimmed using MUSCLE v3.8.31 (Edgar 2004), with default options, and trimAl v1.2rev59 (Capella-Gutierrez et al. 2009), with the -strict option, respectively. Conversion to phylip format prior to tree estimation was done using Geneious (Kearse et al. 2012). A summary tree is reproduced in Figure 2.4 to aid the reader's appreciation of the larger phylogeny.

While the majority of strains in Figure 2.5 exhibit the expected phylogenetic relationships, some do not. For example there is a clear distinction between the Basidiomycetes (marked with a purple bar) and the Ascomycetes, and within the Basidiomycete clade there is a clear distinction between the three major groups represented in the study: the Ustilaginales, the *Rhodotorula*, and the Tremellomycetes. Within the Ascomycetes, the Dipodascaceae, *Metschnikowia*, and, to a lesser extent, *Pichia* clades have all resolved correctly, with the rest of the tree being made up of *Saccharomyces* complex strains at varying levels of resolution. On the other hand, several groups of strains are jarringly misplaced. For example the *Torulaspora*, part of the *Saccharomyces* complex, appear to be interspersed within the Basidiomycete

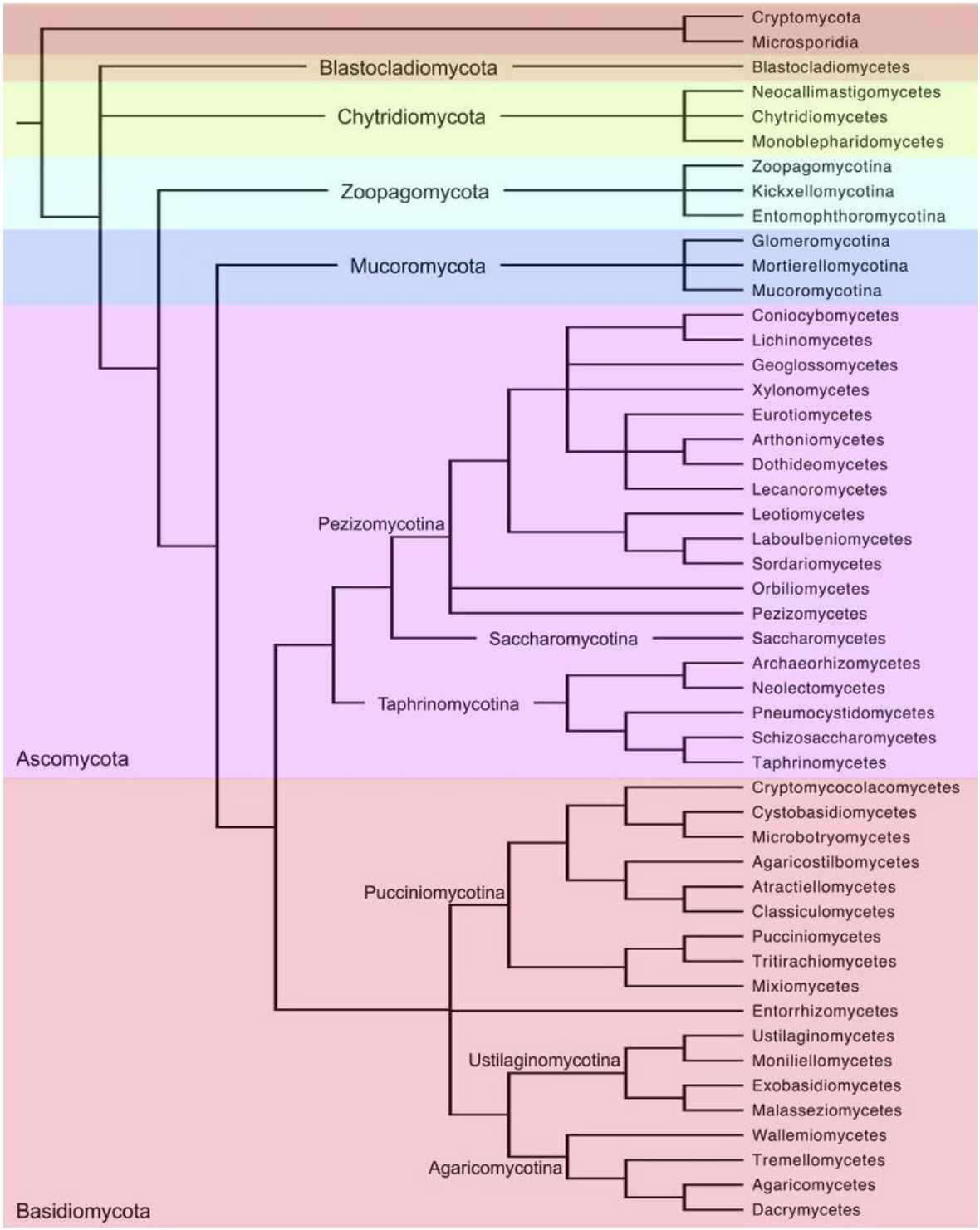

Figure 2.3: Phylogenetic relationships of the fungal kingdom, concentrating on the Ascomycota and Basidiomycota, the phyla of interest in this thesis. Taken from `https://www.researchgate.net/publication/319869622_The_Fungal_Tree_of_Life_from_Molecular_Systematics_to_Genome-Scale_Phylogenies`. A phylogeny specifically for yeast species has never been achieved but this offers a good overview. Most of the Ascomycete NCYC strains are from the Saccharomycotina and the Taphrinomycotina, with a small number (3) from the Pezizomycotina, and the Basidiomycete NCYC strains are from the Agaricomycotina, Pucciniomycotina and Ustilaginomycotina.

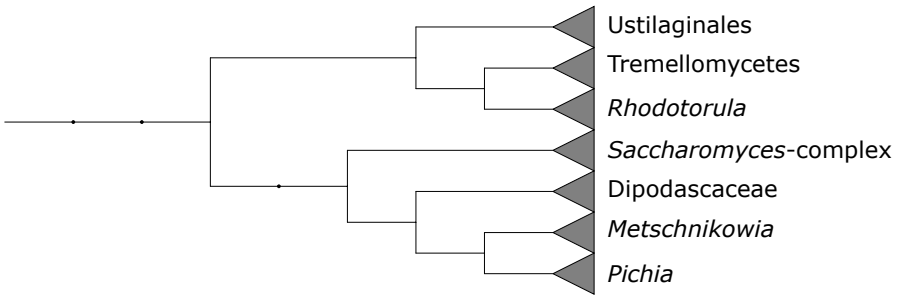

Figure 2.4: This is a summary of the phylogeny shown in full in Figure 2.5. The major clades have been collapsed to show the global arrangement. See the full figure and explanation in text for discussion of anomalous strains not seen in the summary.

Rhodotorula, and vice versa. Additionally, there are a few apparent *S. cerevisiae* strains within the Ustilaginales clade, and several Tremellomycete strains (Basidiomycetes) nestled within the *Saccharomycete* complex. The reasons for this will be complex, ranging from strain mix-ups to sequencing errors and contamination of samples. As established further in later chapters, the Ustilaginales interlopers are clearly the result of contamination, and this is also likely the case for the aberrant *Rhodotorula* and *Torulaspora* strains given that they were sequenced on the same plate (P2). Furthermore, the mapping of sequencing Plate 11 samples (which include most of the abberrant Tremellomycete strains) to NCYC accessions has very recently been called into doubt (supported by this figure). It is known that the sequencing manifest was changed between the submission and sequencing of this plate and it is hoped that the mapping error will prove to be systematic and therefore recoverable. This issue is currently under investigation. Bootstrap values are also generally low in many parts of the tree, perhaps due to the same issue or potentially a function of the very large taxonomic distances between strains included in the tree. This may be why no detailed phylogeny exists that includes both yeast phyla.

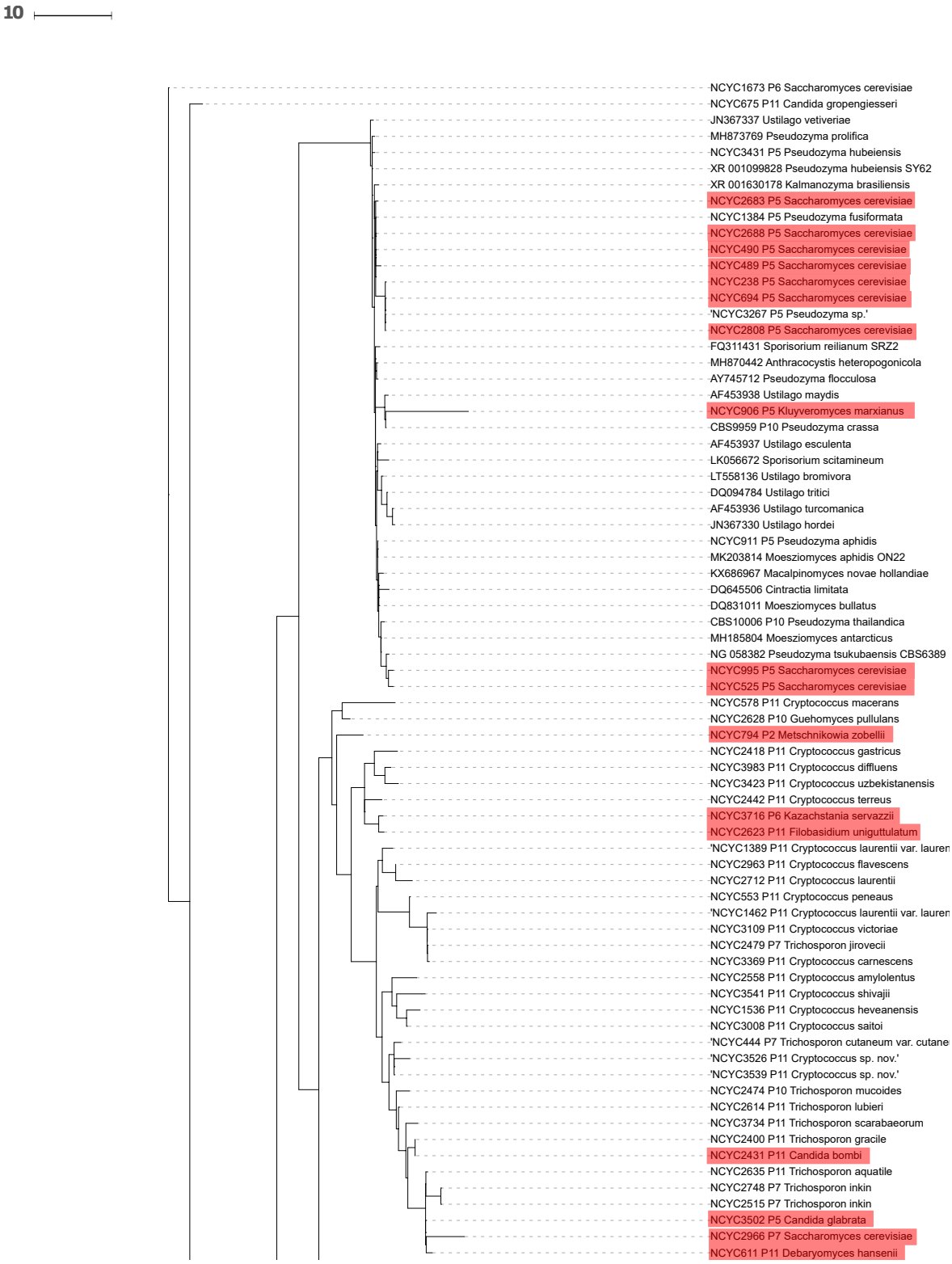

Figure 2.5: Continued on next page.

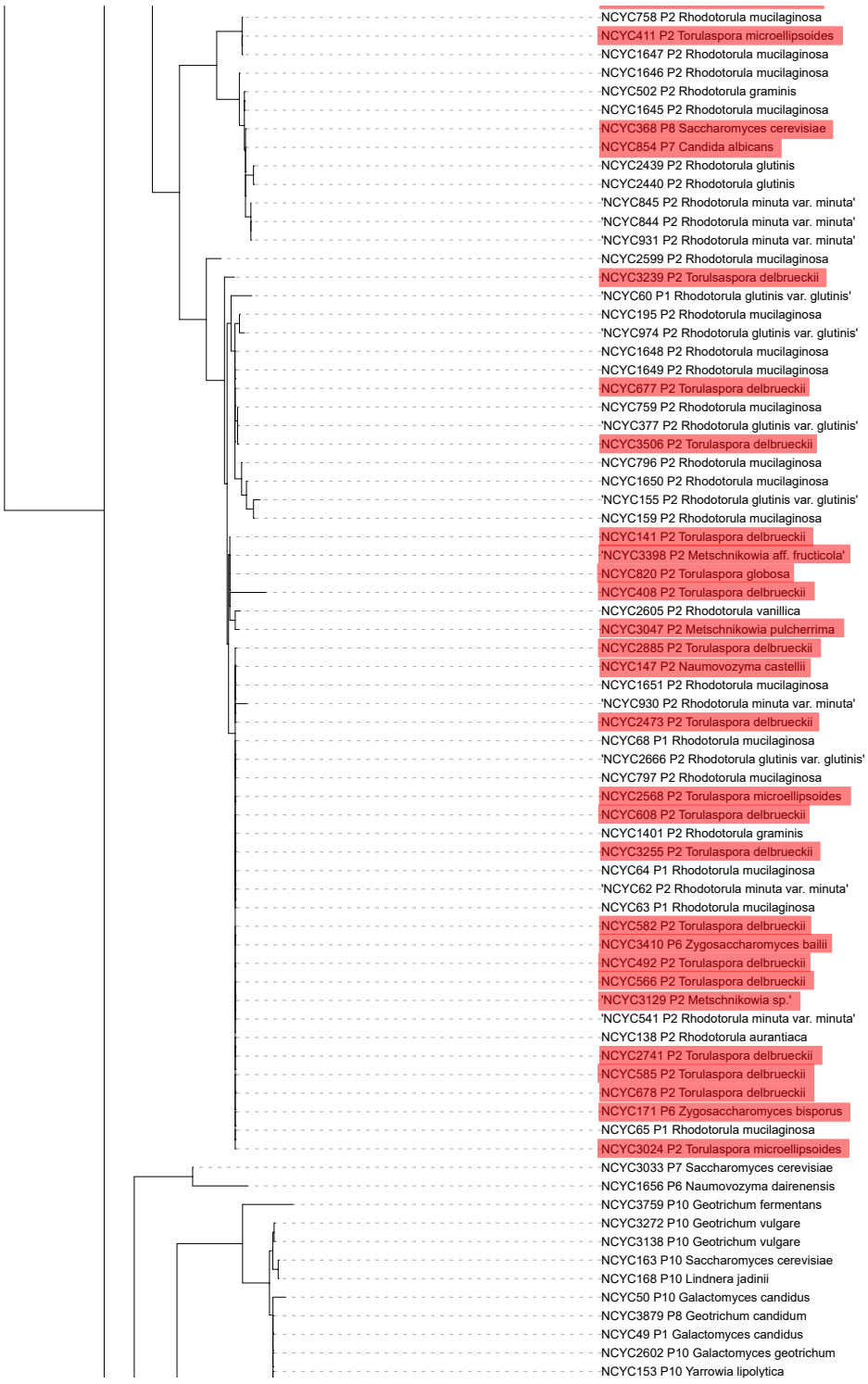

Figure 2.5: Continued on next page.

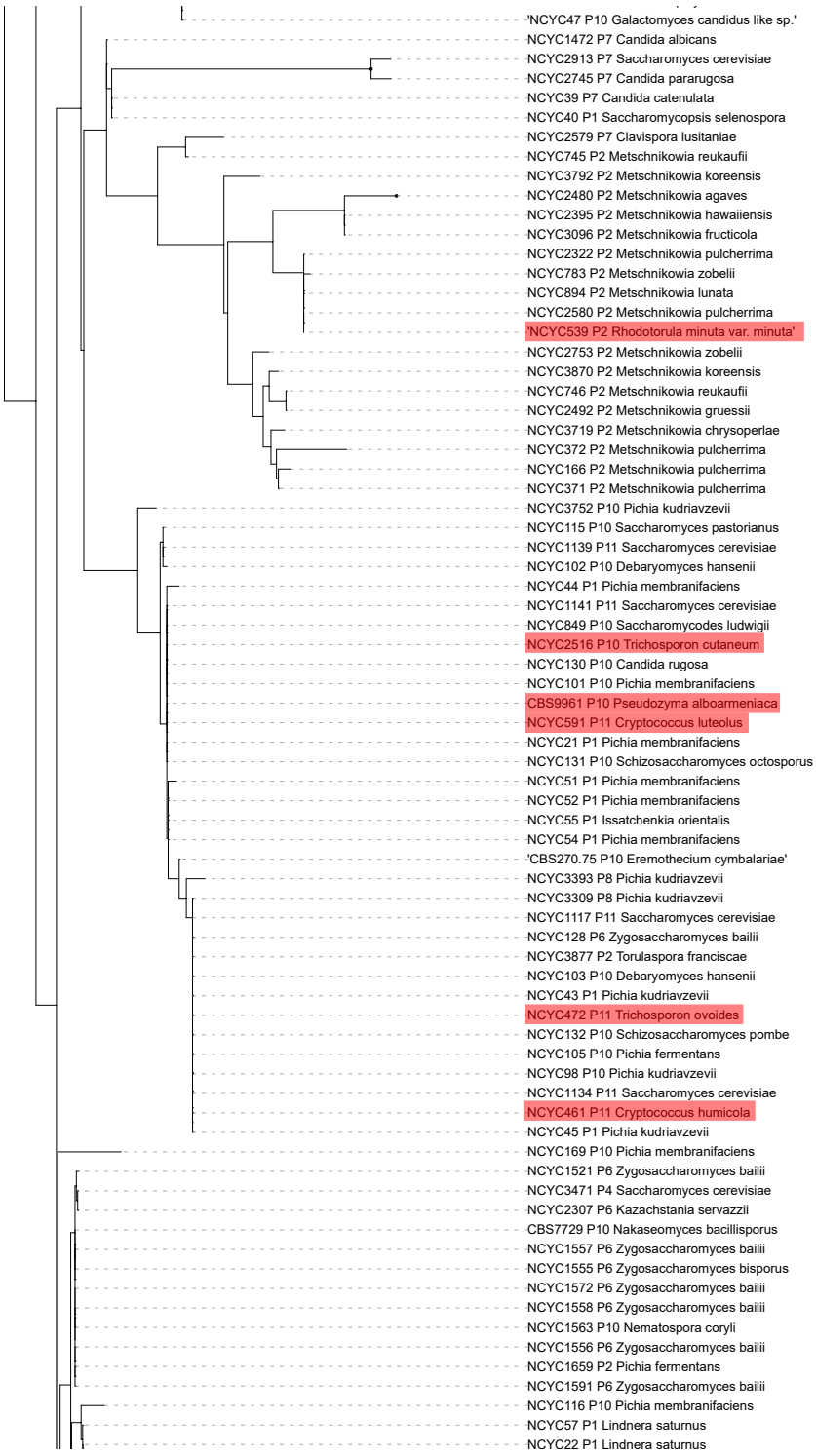

Figure 2.5: Continued on next page.

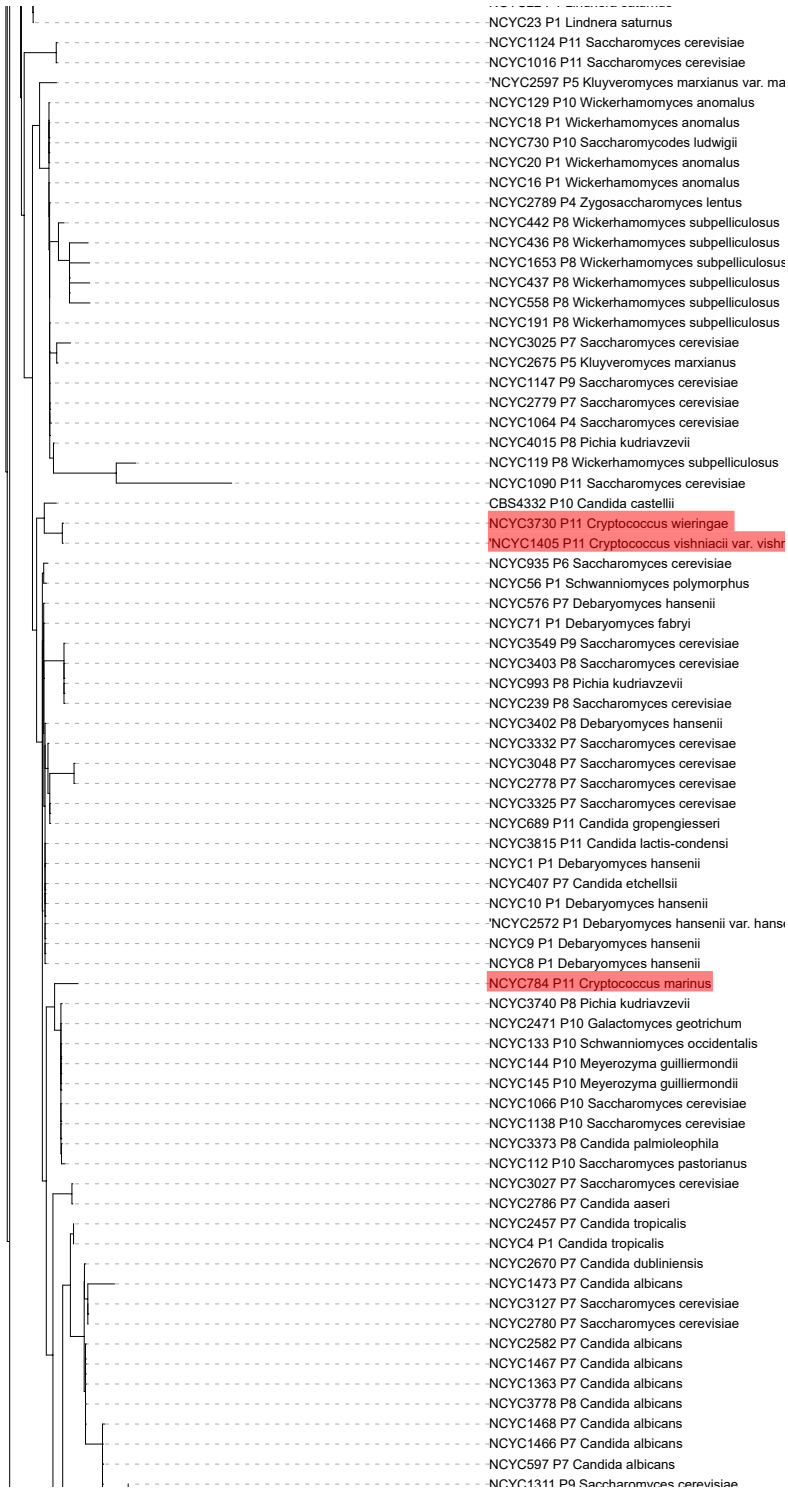

Figure 2.5: Continued on next page.

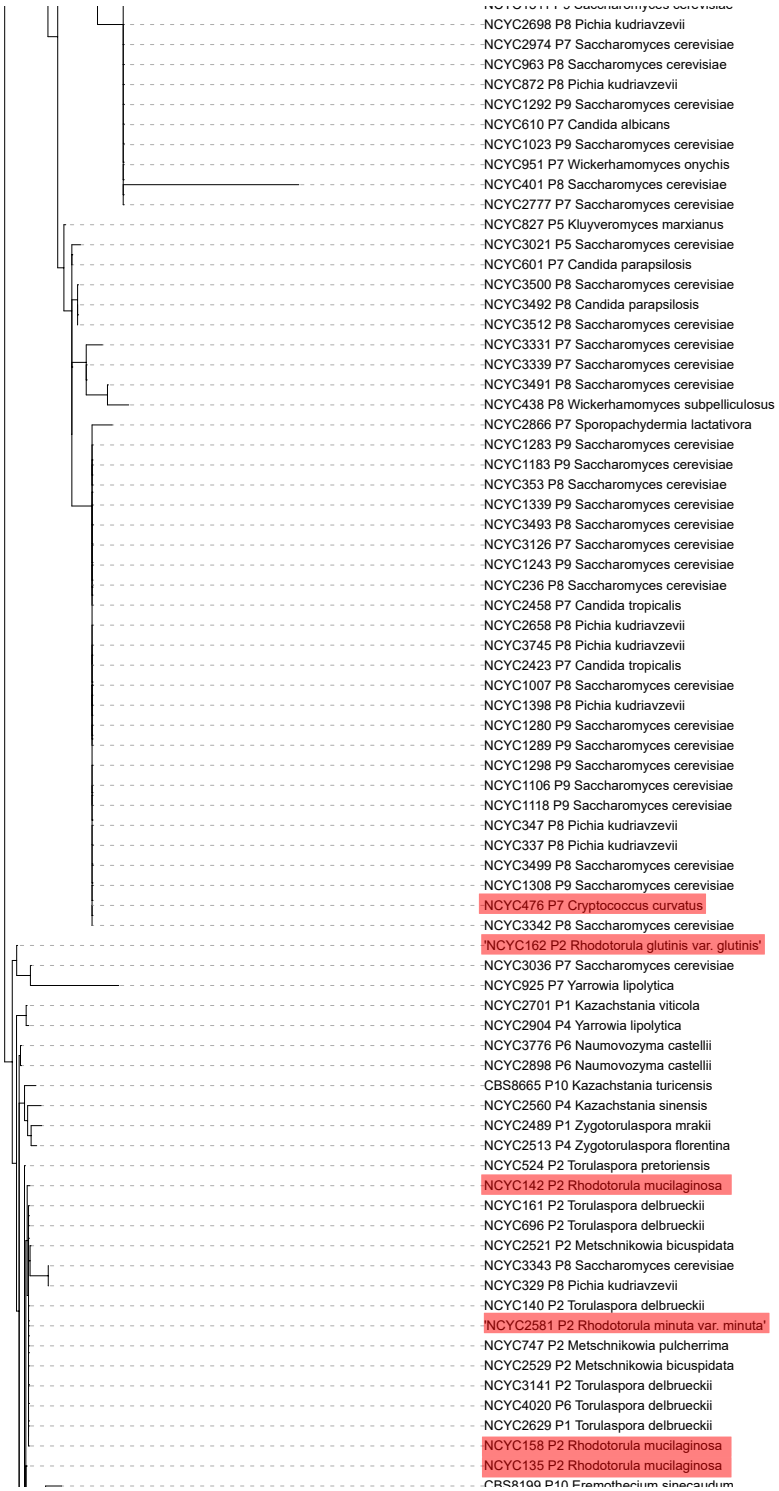

Figure 2.5: Continued on next page.

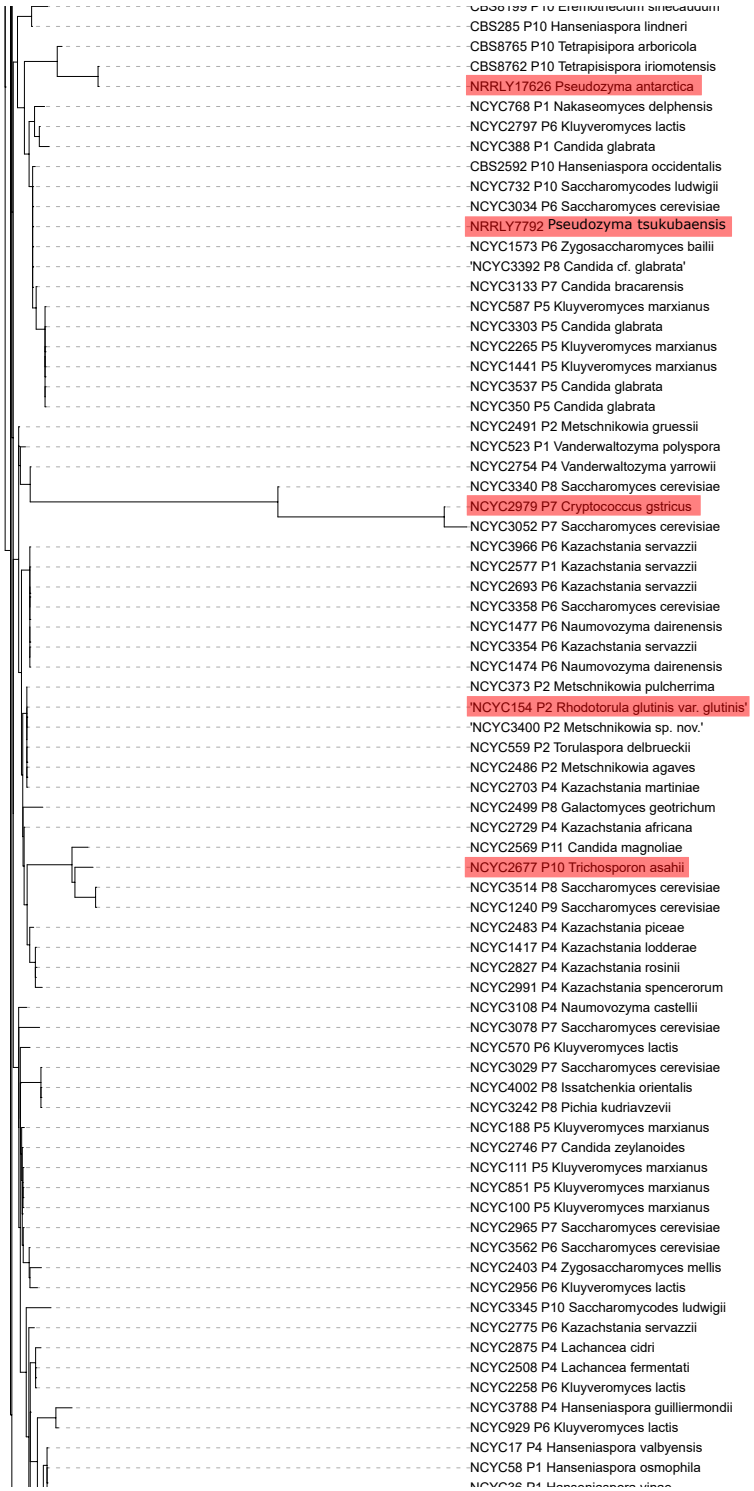

Figure 2.5: Continued on next page.

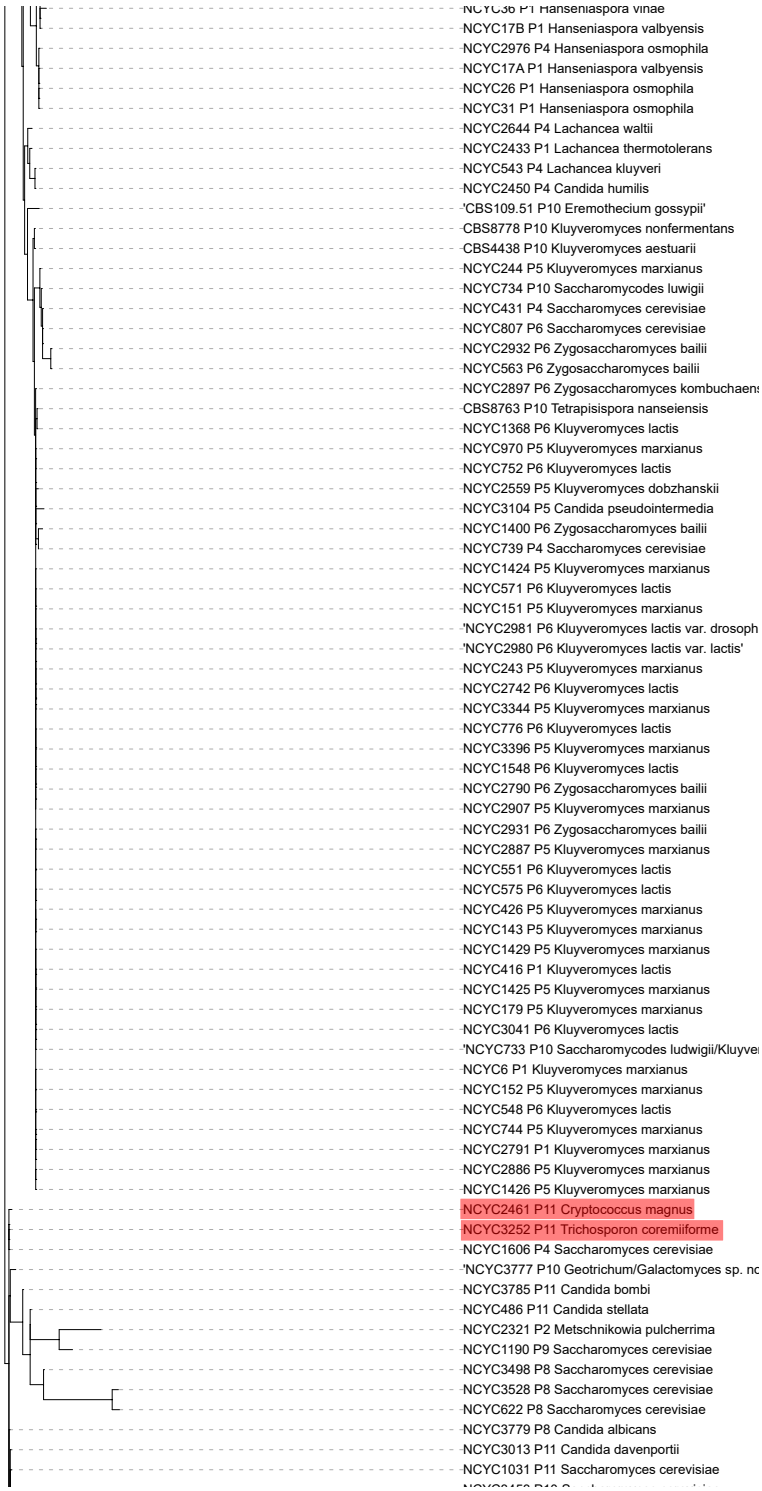

Figure 2.5: Continued on next page.

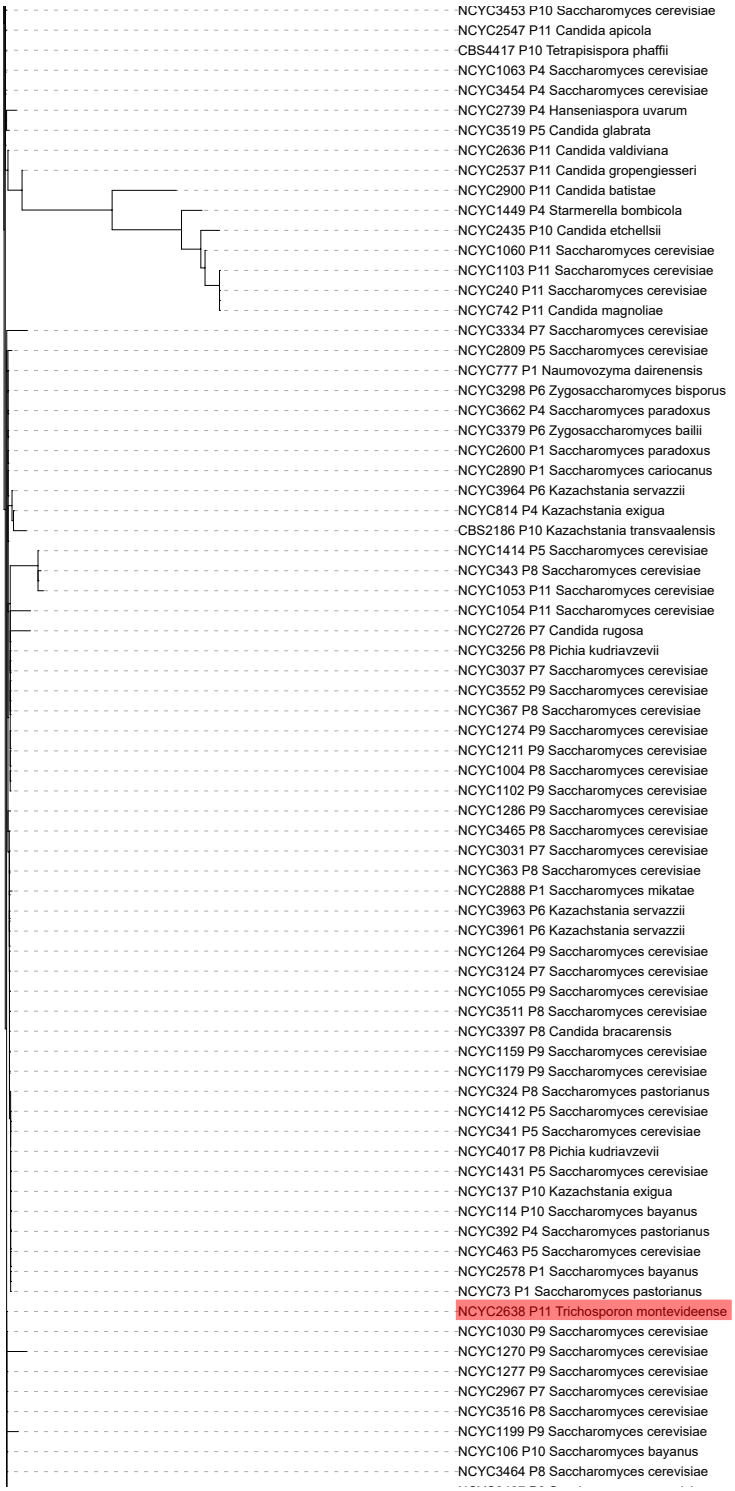

Figure 2.5: Continued on next page.

Figure 2.5: Continued on next page.

Figure 2.5: Continued on next page.

Figure 2.5: Continued on next page.

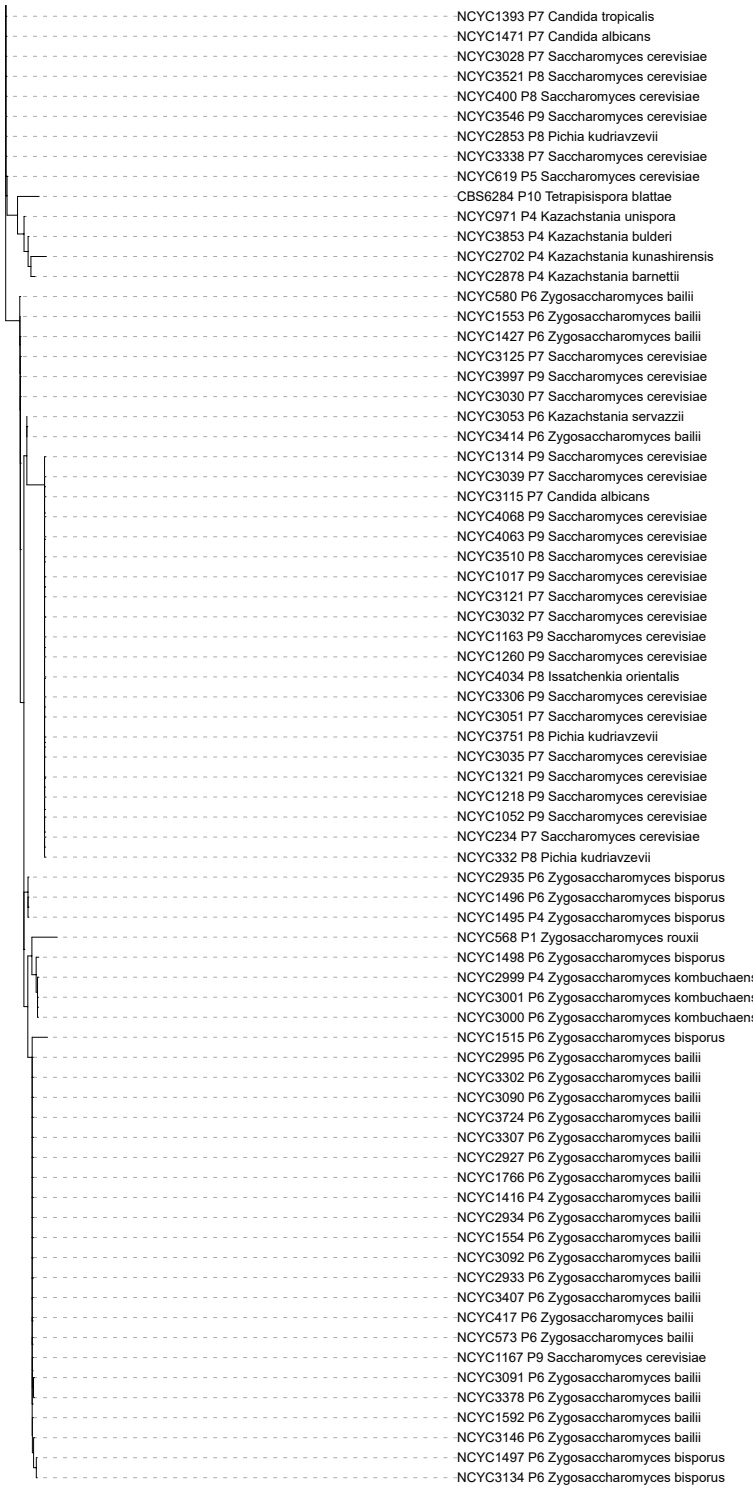

Figure 2.5: Phylogenetic tree constructed from the D1/D2 subunit of the ribosomal DNA of all sequenced NCYC strains and a number of publicly available Ustilaginomycete genomes. Model = GTR + Gamma. Final ML Optimization Likelihood: -18697.674333. The tree is split into two major clades, the Basidiomycetes (indicated by purple bar) and the Ascomycetes. Anomalous placements are highlighted in red.

3 The Mannosylerythritol Lipid gene cluster in the Basidiomycetes

3.1 Summary

- The NCYC genome collection, along with publicly available genomes, is searched for the gene cluster producing Mannosylerythritol lipids.
- The evolutionary history and taxonomic extent of the gene cluster are investigated through phylogenetic analysis.
- The species group containing the gene cluster is explored in more detail to deduce how it was formed.

3.2 Introduction

Mannosyl-erythritol lipids (MELs) are a class of biosurfactants (compounds affecting surface tension, often used as detergents) believed to be solely manufactured by a small group of Basidiomycetous yeasts (or yeast-like fungi), closely related to the smut fungus *Ustilago maydis*. Most of the species previously implicated in MEL production were formerly classified within the now-defunct genus *Pseudozyma* but are now considered paraphyletic and renamed as such (Wang et al. 2015). MELs are glycolipids made up of a mannosyl-erythritol disaccharide with a number of fatty acid chains attached. The number and length of these chains is variable and confers corresponding variation onto the compounds' biochemical activities and applications (Hewald et al. 2006; Fukuoka et al. 2007c,b; Morita et al. 2011; Fukuoka et al. 2007a). See Figure 3.1 for the generic structre of MELs and common variants. Other rare variants have also been observed, see Table 3.1. Applications of MELs include use in cosmetics, bioremediation (specifically oil spill clean-up), and drug delivery via vesicle formation (Cameotra et al. 2004). They also have antibiotic, anti-fungal, and anti-tumour effects (Rodrigues et al. 2006). The benefit of biosurfactants such as MELs over their synthetic counterparts is that they are both biodegradable and less toxic (Yu et al. 2015). They also

work at wider temperature, salinity, and pH ranges (Fan et al. 2014). This is a considerable advantage over conventional synthetic surfactants that have been reported to be highly toxic to both soil and aquatic environments (Emmanuel et al. 2005), irritating to skin (Rodrigues et al. 2006), and prone to producing harmful by-products during manufacture and use (Makkar et al. 2002; Scott et al. 2000).

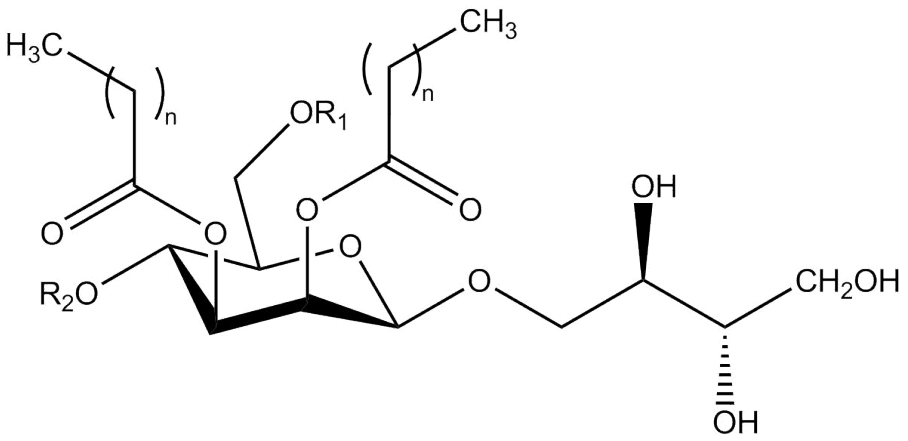

Figure 3.1: Mannosylerythritol lipid. Positions R_1 and R_2 may be acetylated depending on the type of MEL. MEL-A: R_1 & R_2 = acetyl, MEL-B: R_1 = acetyl & R_2 = H, MEL-C: R_1 = H & R_2 = acetyl, MEL-D: R_1 & R_2 = H

Table 3.1 details the various types of MEL manufactured by Ustilaginales. This list represents experimentally validated lipid production, as opposed to that predicted from gene cluster presence as will be shown later. In most cases, MELs are produced as a complement of compounds (i.e. all or most types of MEL are produced, but in different ratios) and the principal component, if one type is dominant, is specified in the table. With regards to the different types of MEL mentioned, MEL-A seems to have the greatest efficacy in mediating liposome fusion (Inoh et al. 2004).

The gene cluster underlying MEL biosynthesis comprises five genes and was first described in *U. maydis* (Hewald et al. 2006). In that species, the cluster inhabits an ~18Kb region of chromosome seven. The five genes are an acetyltransferase (*mat1*), major facilitator (membrane-bound transporter) (*mmf1*), acyltransferase (*mac1*), glycosyltransferase (*emt1*), and acyltransferase (*mac2*), see Figure 3.2. This gene cluster has since been found in a number of *Pseudozyma* strains (Morita et al. 2011; Konishi et al. 2013).

This chapter details a comprehensive search of publicly available genomes for the presence of the MEL gene cluster, as well as a search of *de novo* sequenced and assembled genomes from the NCYC collection. The aim is to answer questions regarding the taxonomic spread

Species	**MEL type(s)**	**Whole genome sequence available**	**Reference**
Ustilago maydis	MEL-A,B,C,D	Public	Hewald et al. (2006)
Ustilago cynodontis	MEL-C	Public	Morita et al. (2008b)
Ustilago scitaminea (*Sporisorium scitamineum*)	MEL-B	Public	Yu et al. (2015)
Pseudozyma antarctica (*Moesziomyces antarctica*)	MEL-A,B,C,D mono-acylated MEL tri-acylated MEL	Public	Kitamoto et al. (1990) Arutchelvi et al. (2008) Arutchelvi et al. (2008)
Pseudozyma aphidis (*Moesziomyces aphidis*)	MEL-A,B,C,D	Public, This study	Rau et al. (2005)
Pseudozyma hubeiensis	MEL-A,B,C	Public, This study	Konishi et al. (2011)
Pseudozyma tsukubaensis (*Macalpinomyces spermophorus*)	modified MEL-B	This study	Morita et al. (2010) Yamamoto et al. (2013)
Pseudozyma fusiformata (*Kalmanozyma fusiformata*)	MEL-A	This study	Morita et al. (2007a)
Pseudozyma rugulosa (*Moesziomyces bullatus*)	MEL-A,B,C tri-acylated MEL	This study	Morita et al. (2006)
Pseudozyma churashimaensis (*Dirkmeia churashimaensis*)	Mono-acylated tri-acetylated MEL	This study	Morita et al. (2011)
Pseudozyma crassa (*Triodiomyces crassus*)	MEL-A,B,C	This study	Fukuoka et al. (2008)
Pseudozyma graminicola (*Sporisorium graminicola*)	MEL-A,B,C	Unknown	Morita et al. (2008a)
Pseudozyma parantarctica (*Moesziomyces parantarcticus*)	MEL-A,B,C mono-acylated MEL tri-acylated MEL	Unknown	Morita et al. (2008c) Morita et al. (2013c)
Pseudozyma shanxiensis (*Ustilago shanxiensis*)	MEL-C	Unknown	Fukuoka et al. (2007b)
Pseudozyma siamensis	MEL-B,C	Unknown	Rodrigues et al. (2006)

Table 3.1: MEL production in yeast-like Basidiomycetes. If the species name used in the paper referenced differs from the strain's current name (according to Mycobank.org), the current name is noted in parentheses.

Figure 3.2: MEL gene cluster as seen in *U. maydis*. Each coloured arrow represents a gene, with the relative orientations indicated.

and evolutionary history of the MEL gene cluster. Is the gene cluster really restricted to this small taxonomic group? In addition, in those strains where gene cluster genes are found, the chapter concerns more detailed investigations into the evolutionary history of the gene cluster and variation found amongst different cluster-containing strains. By examining the individual genes, can we gain insight into how the gene cluster may have formed?

3.3 Methods

MEL gene sequences from *U. maydis* were extracted from GenBank (Sayers et al. 2019) and used to collect matching sequences from other publicly available MEL-producing *Pseudozyma* strains (*P. antarctica* T34, *P. antarctica* JCM10317, *P. aphidis* DSM70725, *P. hubeiensis* SY62), via tBLASTn (Altschul et al. 1990). *P. brasiliensis* GHG001 and *P. flocculosa* PF1 were also searched as they are closely related to *U. maydis*. The gene sequences were annotated with regards to intron and boundary positions using the Geneseqer USD webserver (Brendel et al. 2004) with the species specific splice site model set to "Yeast". Sequences were aligned using MUSCLE v.3.8.31 (Edgar 2004), with default options, and checked manually.

HMMER 3.1b2 (Eddy 1998, 2015) was used to create Hidden Markov Models (HMMs) for each gene (run *hmmbuild* on stockholm alignment of known sequences, default options), which were then used to search the genomes of the NCYC collection for homologues of the MEL genes (run *nhmmer* on each genome - assembled contigs - with the HMMs as queries). This sequence later formed the basis of the FindClusters pipeline described in Chapter 6.

GenBank was later searched for additional MEL genes (in February 2019) to pick up newly deposited sequences. The *U. maydis* sequences were used again as the query for blastn and tblastn searches against both the nr/nt and wgs databases. Further alignments (one per gene) were generated by combining the top hit from each strain searched via HMMER with the sequences obtained from GenBank. Gene trees were estimated using RAxML (trimmed FASTA alignments were generated by MUSCLE, as above, and trimAl v1.2rev59 (Capella-Gutierrez et al. 2009) with the "-strict" option, and converted to Phylip format using Geneious (Kearse

et al. 2012)) with 100 bootstrap iterations and the model set to GTR + GAMMA (other options: -f a, -x 12345, -p 12345). To check which *mmf1* sequences were true *mmf1* genes rather than similar members of the Major Facilitator Superfamily, a separate gene tree was estimated using the top two hits for each strain. This tree was computed using MrBayes v3.2.1 (Ronquist et al. 2012) over 1,000,000 MCMC generations (other options: printfreq=1000, samplefreq=500, nchains=4, starttree=random, nrun=2, burninfrac=0.25). Model selection for all trees was performed using PartitionFinder2 (options: model_selection=BIC, models=all, search=greedy, Lanfear et al. (2017)).

3.4 Results & Discussion

Gene cluster identification

Searches of publicly available Ustilaginales genomes and draft genome assemblies of approximately one thousand NCYC strains revealed that the MEL gene cluster does indeed appear to be confined to a fairly narrow taxonomic group. See Table 3.2 for a list of publicly available and NCYC genomes containing the MEL cluster. Of all the NCYC strains, gene clusters were found only in NCYC911 (*P. aphidis*), NCYC1384 (*P. fusiformata*), NCYC1510 (*P. tsukubaensis*), NCYC3267 (*sp. nov.*), and NCYC3431 (*P. hubeiensis*). Work in this project confirmed the presence of the MEL gene cluster in most known producers for which there is genome sequence information (confirming, and adding to, the results of Morita et al. (2013c,b), Lorenz et al. (2014), and Saika et al. (2014) and Saika et al. (2016)). One additional potential producer was identified, NCYC3267, a novel species apparently related to *P. tsukubaensis*. Sequences matching *mmf1* were found in many of the strains searched (330), presumably because it is a member of the Major Facilitator gene superfamily and does not vary much from other members (common in all genomes). Most of those matching sequences are less than 500bp in length. The results from the FindClusters Pipeline are shown in Figure 3.3. Although the gene clusters found in NCYC1384 and NCYC3267 are split across contig ends, we assume that they are nevertheless complete clusters. The NCYC1510 genome assembly is fragmented but given the presence of all 5 genes, we assume the same for this strain. This strain was sequenced twice and in the other version the genes are clustered in a manner identical to those of NCYC3267. It should be noted that the gene cluster is not found in NNRL_Y-7792 (another *P. tsukubaensis* strain), NRRL_Y-17626 (*P. rugulosa*), CBS9959 (*P. crassa*), or JCM16988 (*P. churashimaensis*), despite these being reported producers. It is possible that these specific strains do not contain the gene cluster. Alternatively, this could just be due to missing sequence (i.e. that section of the genome is poorly covered).

Species identification was confirmed for species purporting to contain the gene cluster using a combination of reciprocal BLAST searches for the 26S ribosomal DNA region, and k-mer tests against representative reference genomes by Ann-Marie Keane (AM - NCYC), using a custom yeast version of Kraken v1.0 (Wood et al. 2014). This process confirmed the identity of the *Pseudozyma* strains and discounted several supposed *S. cerevisiae* MEL cluster carriers (NCYC238, NCYC489, NCYC525, NCYC694, and NCYC2808), with these spurious non-Basidiomycete hits likely the result of contamination during the genome sequencing phase. The latter strains have been omitted from the gene cluster arrangement diagram in Fig. 3.3.

Apart from those sequences described above, partial sequence matches were found for *mat1*, *emt1*, and *mac2* in 38, 3, and 2 NCYC strains, respectively. These strains are mainly *Kluyveromyces marxianus*, plus four *S. cerevisiae*, two *Rhodotorula glutinis*, and one each of *K. lactis*, *Kazachstania servazzi*, *R. minuta*, *Cryptococcus diffluens*, and *C. flavescens*. NCYC2683 has matches for all three genes, and is the only strain to have more than one, but can be discounted as above due to evidence of contamination. The remainder of the partial matches are short, under 80-100bp in length. Kraken analysis, by AM as above, did not indicate contamination of the other genome sequences and these are presumably spurious matches, given this short length and the taxonomic distances between the species concerned. BLAST searches suggest the *K. marxianus* and *K. servazzi* sequences are part of a lactoylglutathione lyase gene (GLO1) which seems to share a very short portion of sequence with *mat1*, while the other sequences (from *Cryptococcus* and *Rhodotorula*) do not match anything in GenBank.

Figure 3.4 shows the locations of introns in the MEL genes of the initial subset of MEL producers. Intron information seems to correlate with taxonomy according to previously mentioned phylogenies (Oliveira et al. 2014; Wang et al. 2015). *P. hubeiensis* closely resembles the *U. maydis* arrangement, while the *Moesziomyces* species (*P. antarctica* and *P. aphidis*) are mostly in accordance with each other (exact placement aside). Notably, *P. fusiformata* appears to have lost all introns. This is an unusual phenomenon and while this has been observed in a gene cluster context before (Sad7 in the oat avenacin gene cluster - Jo Dicks, personal communication), this case was only in a single gene. One of the more popular theories concerning intron loss drivers is that of reverse transcriptase (RT) mediated intron loss (i.e. a partially processed mRNA is reverse transcribed and the sequence reintegrated into the genome, minus the introns) (Cohen et al. 2011).

Gene tree estimation

Species	Strain	Location	Coordinates
Ustilago maydis	521	chr7 AACP02000117	37080-55624
Ustilago cynodontis	NBRC 9727	node435 LZZZ01000435	756-11290
		node711 LZZZ01000710	1100-2641
Ustilago bromivora	UB2112	chr7 LT558123	2600-23099
Ustilago trichophora	RK089	sc42 LVYE01000042	45407-61393
Ustilago xerochlorae	UMa702	node702 MAIN01000049	73619-90697
Ustilago esculenta	MMT	sc2 JTLW01000012	630566-644981
Ustilago hordei	Uh364	sc6 NPMZ01000006	1094741-1111153
Sporisorium scitamineum	SSC39	chr7 CP010919	7244-25482
Sporisorium iseilematis-ciliati	BRIP 60887	node100 MJEU01000100	59410-63125
		node176 MJEU01000176	90-13733
Sporisorium reilianum	SRS1-H2-8	chr7 LT795060	865916-883744
Pseudozyma antarctica	JCM10317	contig0110 BBIZ01000110	13338-27631
	T34	contig00998 BAFG01000551	1839-10128
		contig00997 BAFG01000550	2662-4672
Pseudozyma aphidis	DSM70725	Seq20 AWNI01000012	2361-16973
Pseudozyma aphidis	NCYC911	contig1257	35271-49561
Pseudozyma hubeiensis	SY62	K001P39 BAOW01000069	6934-22533
Pseudozyma hubeiensis	NCYC3431	contig154	59749-75469
Melanopsichium pennsylvanicum	4	sc75 HG529597	95379-114578
Pseudozyma fusiformata	NCYC1384	contig538	358314-369127
		contig412	53337-54984
Pseudozyma tsukubaensis	NCYC1510	contig76892	87-1206
		contig108322	3087-4922
		contig64698	1-1372
		contig109495	1843-3313
		contig95440	1516-3059
Novel species	NCYC3267	contig5214	4604-6101
		contig5978	11172-27532

Table 3.2: MEL gene clusters found in publicly available and NCYC Ustilaginales genomes, with locations. The strains listed are confirmed to contain the cluster, it may also be present in other strains of the same species but this information is not available. Where multiple locations are specified, it is because the cluster spans the ends of two contigs and is assumed to be intact. One extra case is included, that of NCYC1510, which has all the genes of the gene cluster but they are found on separate, very short, contigs due to the fragmented nature of that genome assembly.

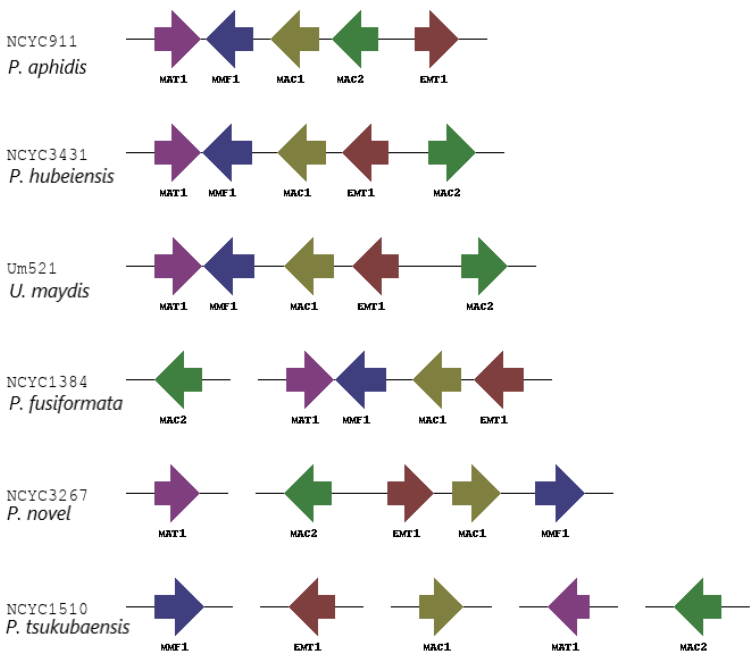

Figure 3.3: Results from the FindClusters pipeline described in Section 6 of this thesis. The analysis has been run on all sequenced NCYC genomes (plus the 23 from other collections), with *U. maydis* included for reference. Colours indicate homology, arrow direction indicates gene orientation, genes connected by a black line are found on the same contig.

Gene trees (Figures 3.6-3.10) were estimated using the matching sequences gleaned from the searches of the NCYC genome dataset and those in GenBank for the five MEL genes (spurious matches, as described in the previous section, were omitted). Due to the apparently highly unique sequences seen in four of the MEL genes, it was not possible to root the trees with a suitable outgroup (i.e. there is no non-cluster homologue of the genes). A search of the Pfam database at [http://pfam.xfam.org/search] (El-Gebali et al. 2019) did not reveal any useful domains with which to find an outgroup, so these gene trees are left unrooted. They are rotated where necessary to aid visual comparisons, with the *Moesziomyces* clade basal as known from Wang et al. (2015). These trees are in broad agreement with each other with regards to clade integrity, with the *Moesziomyces* (*P. aphidis* & *P.antarctica*), *Sporisorium*, and *Ustilago* clades resolving intact in all trees, although low support for the global arrangement is evident in some cases (notably the mid-tier bootstrap support between 30-60% in the MAC2 & MMF1 trees). This may be due to the lack of rooting. The clade arrangement, where supported, is also in agreement with the species tree constructed from the

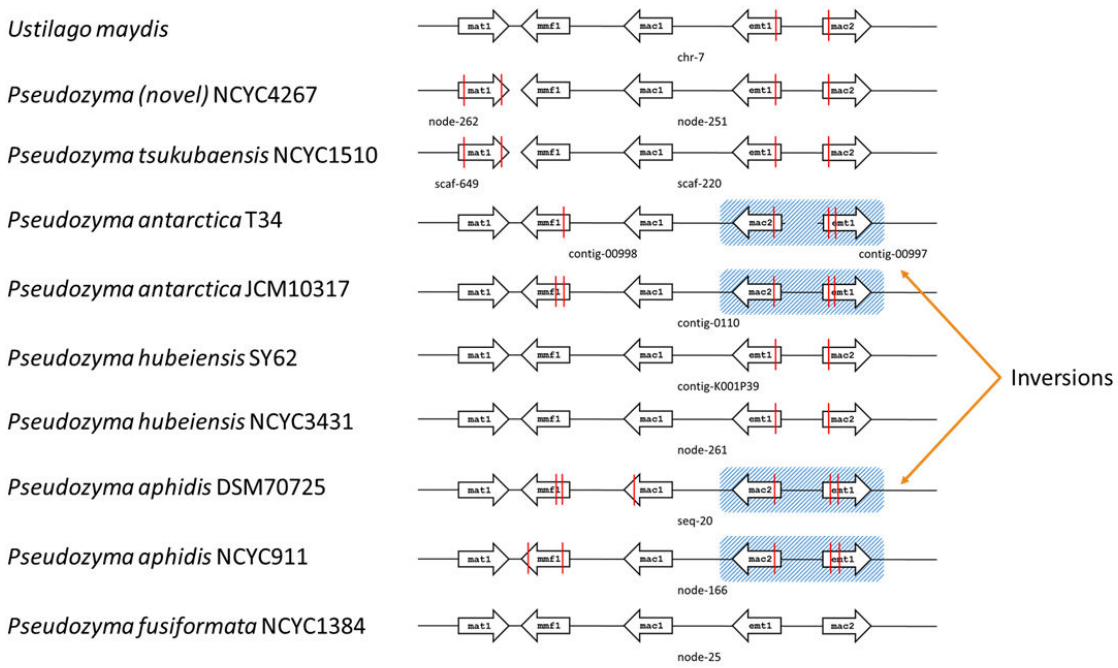

Figure 3.4: Approximate distribution of introns in the MEL genes found in the initial search of the known NCYC and public MEL producers. Arrows indicate genes and direction. Black lines represent connecting intergenic sequence. Labels underneath each gene cluster detail the contig, scaffold or chromosome on which the genes were found. Introns are marked with a red vertical line showing the approximate location of the intron. The inversion of *emt1* and *mac2* in the *Moesziomyces* producers is highlighted in blue hatching.

D1D2 region of the ribosomal DNA (26S), Figure 2.5, and with the Ustilaginales phylogeny of Wang et al. (2015), reproduced in Figure 3.5. This suggests that, on the whole, the MEL genes have originated in a strain ancestral to this taxonomic group. One thing that is left unclear however, is where the genes originally arose. Since no homologues exist in strains that do not contain the gene cluster, their evolutionary progress cannot be tracked through the phylogenetic tree, as seen in Wong et al. (2005).

The exception to the unavailability of an outgroup is the *mmf1* gene tree, which is due to the ubiquitous nature of the Major Facilitator superfamily to which *mmf1* belongs. Evidently, *mmf1* is quite similar to other members of the superfamily, resulting in sequence hits from all searched genomes. A separate gene tree (Figure 3.11) was constructed using the top two hits for each gene cluster-containing strain (as of January 2016) to determine whether the *mmf1* sequence hits found in *P. flocculosa* and *P. brasiliensis*, which lack the gene cluster, were truly *mmf1* genes or just similar members of the Major Facilitator superfamily. It is clear that these two sequences sit with the second hit found in the known gene cluster strains,

meaning that the sequences found in these two species are not true *mmf1* genes. For this reason, Fig. 3.7 has been cropped to show only the region within the clade bounded by *P. flocculosa* and *P. brasiliensis* as these can be taken to be a reasonable outgroup based on the findings of Fig. 3.11.

Gene cluster arrangement

The MEL gene cluster had previously been described in *U. maydis* (Hewald et al. 2006) and a small number of close taxonomic relatives (Konishi et al. 2011; Lorenz et al. 2014; Saika et al. 2014, 2016). The only notable variant in cluster arrangement is found in the *Moesziomyces*, formerly *Pseudozyma aphidis and P. antarctica*, where *emt1* and *mac2* are inverted relative to all other known arrangements (see Figure 3.12, Morita et al. (2007b)). *M. parantarcticus* and *M. rugulosa* are the other reported producers in this clade and presumably share this alternate arrangement seen in the strains investigated here. However, this cannot yet be confirmed in the former as there are no genome sequences available for this species, and the latter shows no evidence of the gene cluster (NRRL Y-17626 analysed earlier in this chapter). There is one other isolated variant in *U. cynodontis* where the gene cluster matches that of *U. maydis* except that *mac1* is apparently inverted relative to that of *U. maydis*. This is not seen in any other strain, even those close to *U. cynodontis*, e.g. *U. xerochloae*. It is not currently known whether this variation in gene cluster arrangement has any tangible effect on the composition of the MEL biosurfactant output, or the structure of the molecules themselves.

3.5 Conclusions

The first question posed in this chapter was whether the MEL gene cluster is restricted to the tight taxonomic group surrounding *Ustilago maydis* and the few other known MEL producers. Searches of both public databases and NCYC genome sequences did not expand the known taxonomic breadth of this gene cluster, suggesting that it is indeed confined to the Ustilaginomycetes. In addition, no partial or modified gene clusters were discovered, apart from two minor inversions. Long read sequencing may be useful to confirm intact gene clusters in NCYC1384, NCYC1510, and NCYC3267.

The second question was whether, by examining the MEL genes themselves, it is possible to infer their evolutionary history and reconstruct the route by which the MEL gene cluster was formed. These results suggest that the MEL gene cluster is made up of evolutionarily unique sequences, at least in the context of what is currently available in GenBank and the

NCYC genome dataset. Apart from sequences sharing the primary domain of the *mmf1* gene (part of the Major Facilitator Superfamily), there are no convincing sequence matches for any MEL gene in any genome searched, other than those actually in the gene cluster. This suggests that the genes making up the MEL gene cluster have diverged substantially and possibly quite quickly from whatever sequences they are descended from, or perhaps that they have evolved from previously non-genic sequences.

The next chapter will deal with the search for a similar gene cluster, that producing cellobiose lipids, which exhibits contrasting patterns of evolution.

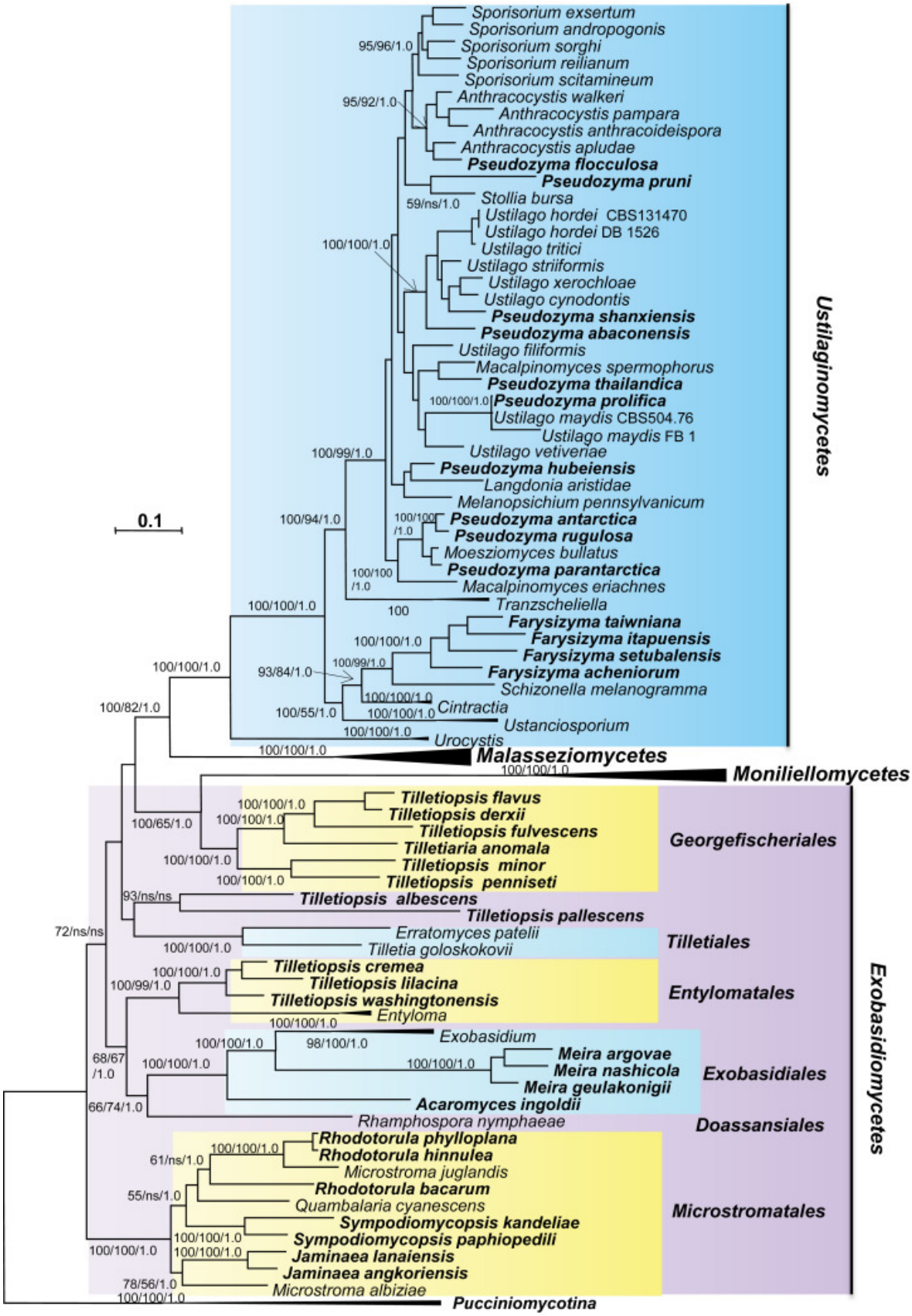

Figure 3.5: Part of the Basidiomycete phylogeny reported in Wang et al. (2015), showing the group containing the known MEL producers and their relatives.

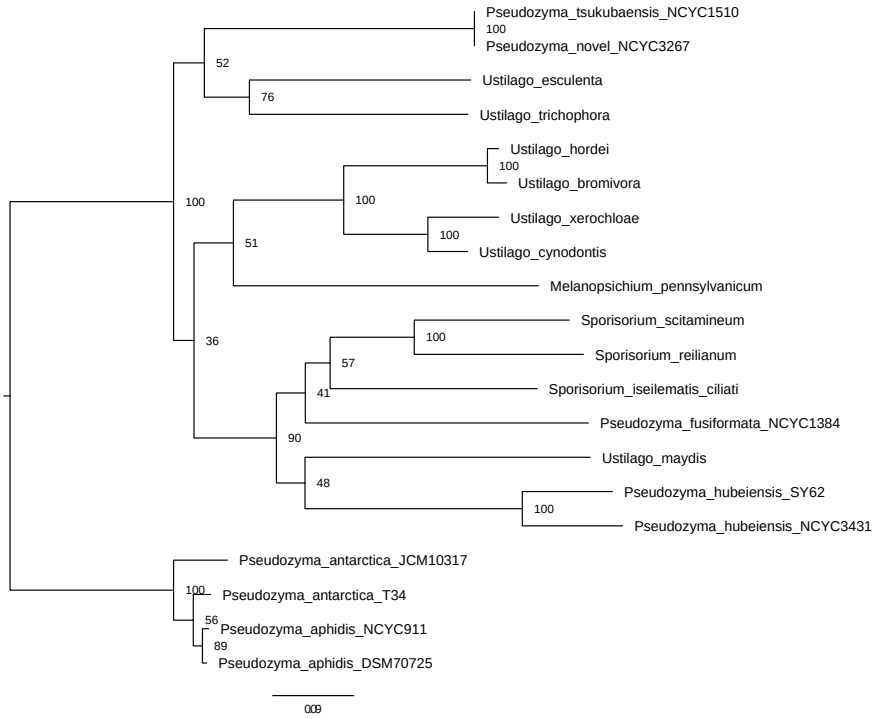

Figure 3.6: MAT1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *mat1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis mat1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -14217.442780.

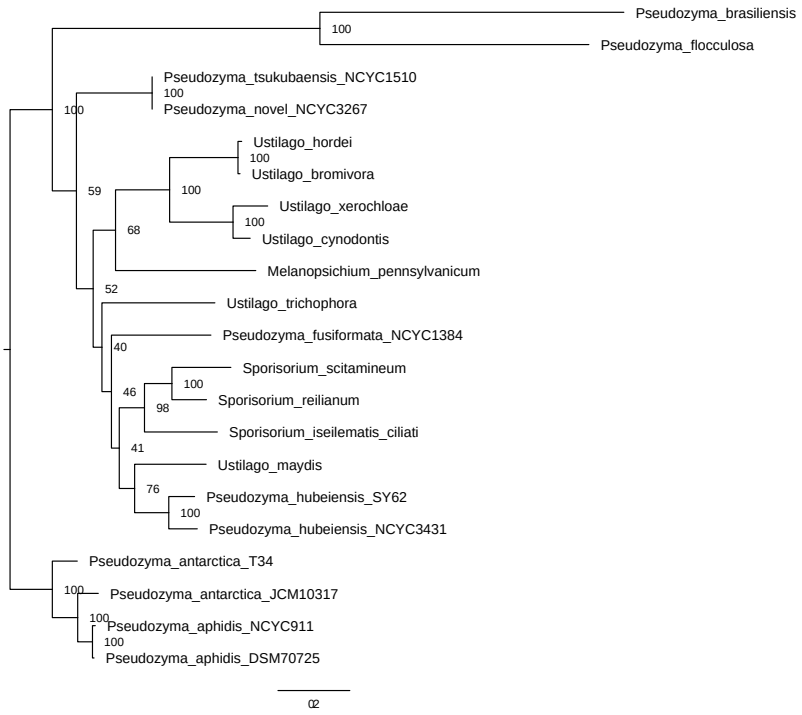

Figure 3.7: MMF1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *mmf1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis mmf1* gene as query. Rooted with *P. brasiliensis* and *P flocculosa*, which do not contain the cluster. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -18768.386626.

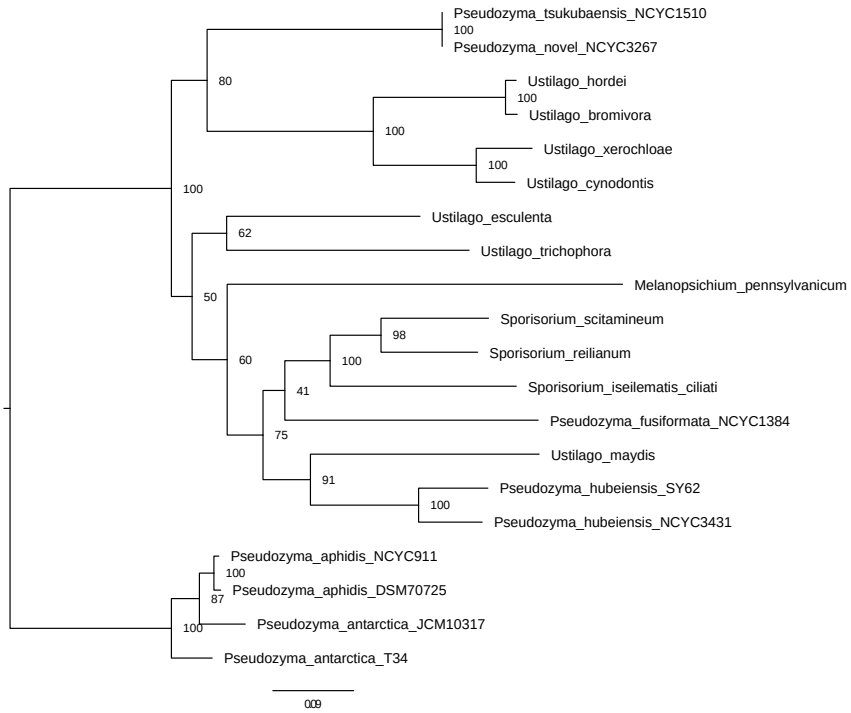

Figure 3.8: MAC1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *mac1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis mac1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -15834.722394.

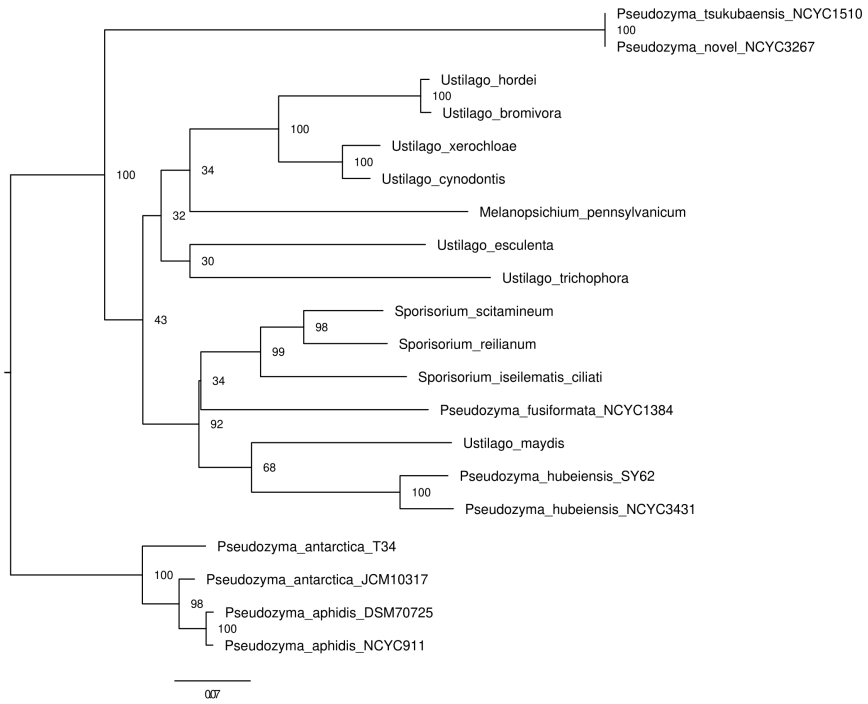

Figure 3.9: EMT1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *emt1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis emt1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -13265.420592.

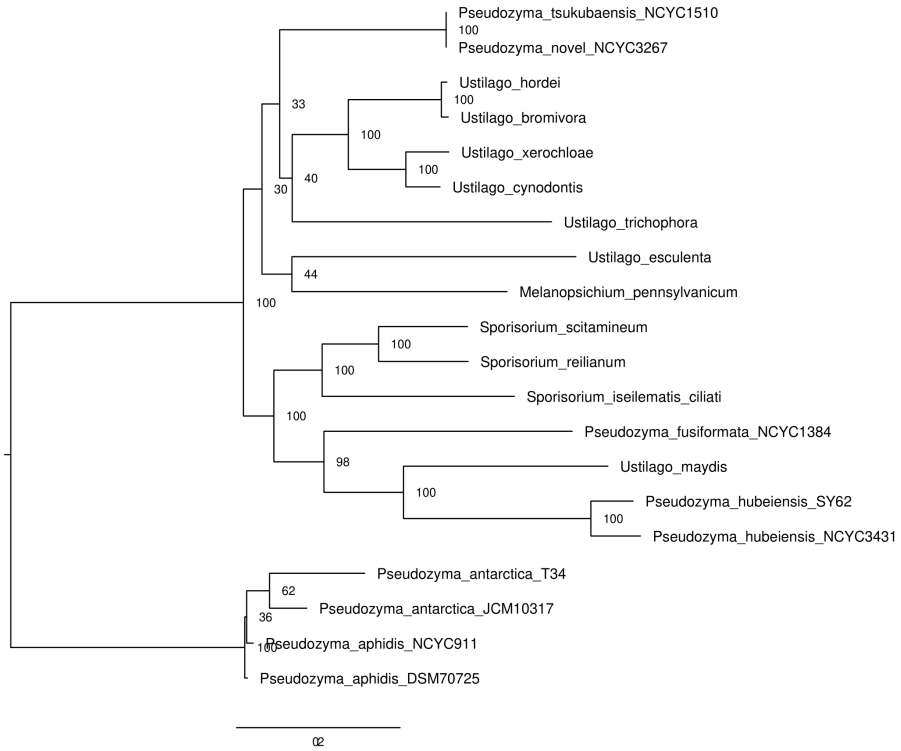

Figure 3.10: MAC2 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *mac2* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis mac2* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -15898.989072.

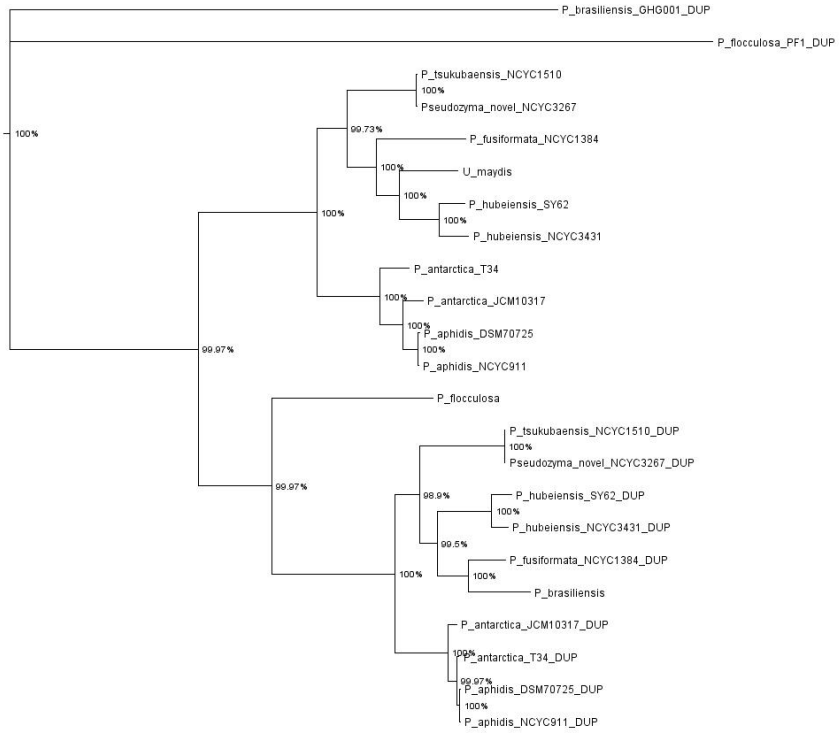

Figure 3.11: Gene tree of top two *mmf1* hits for cluster species (as of 2016). Node labels show percentage support based on Bayesian phylogeny reconstruction. Sequences labelled with a "DUP" suffix are the second best hit for that strain. The top hits for *P. flocculosa* and *P. brasiliensis* clearly group with the second best hits for the other strains, showing that there is no true *mmf1* gene in these two strains.

Figure 3.12: MEL cluster as seen in *Moesziomyces aphidis*, where *emt1* and *mac2* are switched. It is not known what effect this modification has on MEL production. Arrow colour is coded to gene identity, the same as in Fig. 3.2, to aid comparison.

4 The Cellobiose Lipid gene cluster in the Basidiomycetes

4.1 Summary

- The NCYC genome collection, along with publicly available genomes, is searched for the gene cluster producing Cellobiose lipids.
- The evolutionary history and taxonomic extent of the gene cluster are investigated through phylogenetic analysis.
- The species group containing the gene cluster is explored in more detail to deduce how it was formed.
- The reported producers, including *P. aphidis* and *T. porosum* were investigated in detail to determine how such diverse species could be producing the same product (i.e. the similarity of gene clusters).

4.2 Introduction

Cellobiose lipids (CBLs) are another class of biosurfactants, similar to MELs, and produced by a metabolic gene cluster previously identified by Teichmann et al. (2007). Like MELs, CBLs have surface tension modifying effects and are therefore of use in foods, cosmetics, and pharmaceuticals, etc. (Mimee et al. 2009). One well known example of this utility is flocculosin, described below, which exhibits strong antifungal effects that contribute to the use of *P. flocculosa* as a biocontrol agent (Marchand et al. 2009). CBLs are known to be produced by a number of Ustilaginomycete species, roughly corresponding to the MEL producers mentioned in the previous chapter (Roelants et al. 2014). However, CBL production has also been reported in species that do not produce MELs (namely *P. flocculosa* (Mimee et al. 2009)), including species that are taxonomically distant to the Ustilaginomycetes, for example *Vanrija humicola*, aka *Cryptococcus humicola*, and *Trichosporon porosum* (Golubev et al. 2008; Kulakovskaya et al. 2007). See Table 4.1 for a list of species in which CBL pro-

duction has been reported, principally from liquid chromatography assays. Figure 4.1 shows a species tree of CBL producers and close relatives, a subset of the D1/D2 phylogeny seen in Figure 2.5.

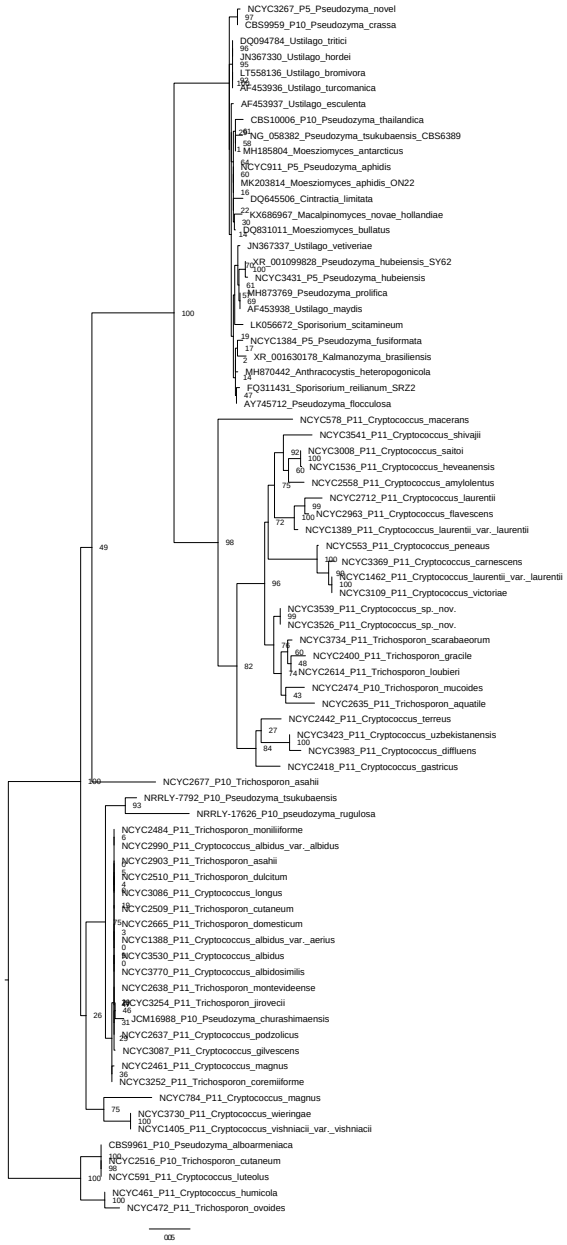

Figure 4.1: Species tree of CBL producers and relatives. Produced from D1/D2 sequences pulled from assembled genomes of all NCYC strains labelled as either *Trichosporon* or *Cryptococcus*, plus all *Pseudozyma* and some publicly available Ustilaginales strains.

CBLs are comprised of a cellobiose sugar with fatty acid chains attached, see Figures 4.2 and 4.3 for the chemical structures of the two principle variants known. The gene clusters for the cellobiose lipid pathway have been previously characterised in *Ustilago maydis* (ustilagic acid - UA, Teichmann et al. (2007)) and *Pseudozyma flocculosa* (flocculosin, Teichmann et al. (2011)), two species within the Ustilaginales. In the former, the gene cluster has 12 genes

while in the latter there are 11, see Figure 4.4 for the cluster arrangements, which vary substantially. Notice also that the UA cluster contains two genes at its rightmost end that are not found in the *P. flocculosa* cluster, *orf2* and *ahd1*, while the *P. flocculosa* cluster has a third acyltransferase, *fat3*, not found on the other cluster (this gene may be a duplication of *fat1* as they are very similar in sequence). The structural differences between the products of the two gene clusters are believed (by Roelants et al. (2014)) to be a result of the actions, or lack thereof, of *ahd1* and *fat3* (there is a very good figure describing the biosynthetic pathway for both products in Jezierska et al. (2018)). There is also evidence from publicly available genome sequences that CBL production in *C. humicola* and *T. porosum* is underpinned by a gene cluster that shares some, but not all, component genes with those of *U. maydis* and *P. flocculosa* (Personal Communication, I. van Bogaert, 2016). There is also a suggestion that the gene cluster is not present in *P. aphidis*, despite it being a reported producer of CBLs (Personal Communication, I. van Bogaert, 2016).

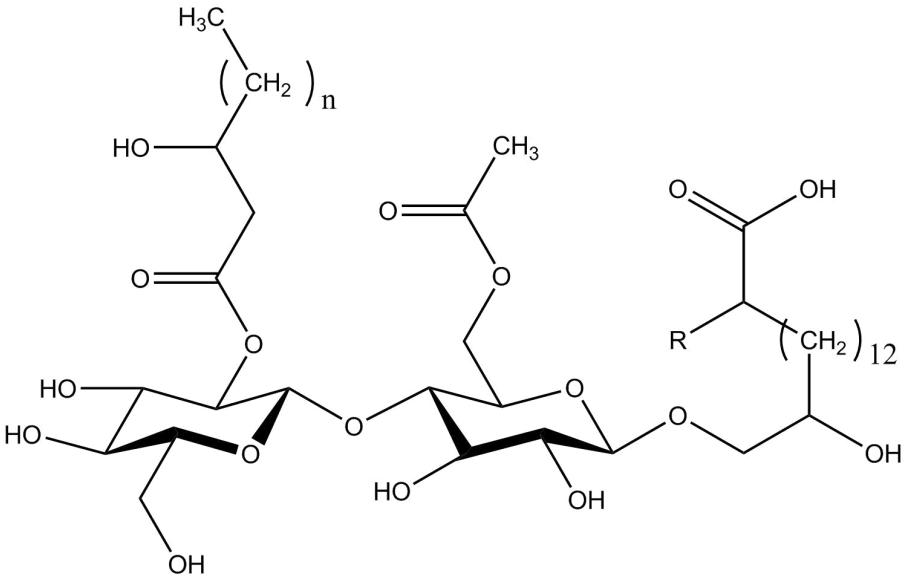

Figure 4.2: CBL ustilagic acid, produced by *Ustilago maydis*. The R group may represent differing length fatty acid chains.

In this chapter, the NCYC genome sequences have been used to gain further insight into the CBL gene cluster, including the different configurations found in the various producers, and the taxonomic spread of the cluster. How widely spread is the gene cluster found in *U. maydis* and *P. flocculosa*, and what variation in cluster construction exists among producers? Is this gene cluster responsible for CBL production in all species? This second question is of particular relevance given the taxonomic distance between the Ustilaginomycetes and the Tremellales (*Cryptococcus* and *Trichosporon*). Lastly, where is the biosynthetic pathway responsible for CBL production in *P. aphidis*, and what form does it take?

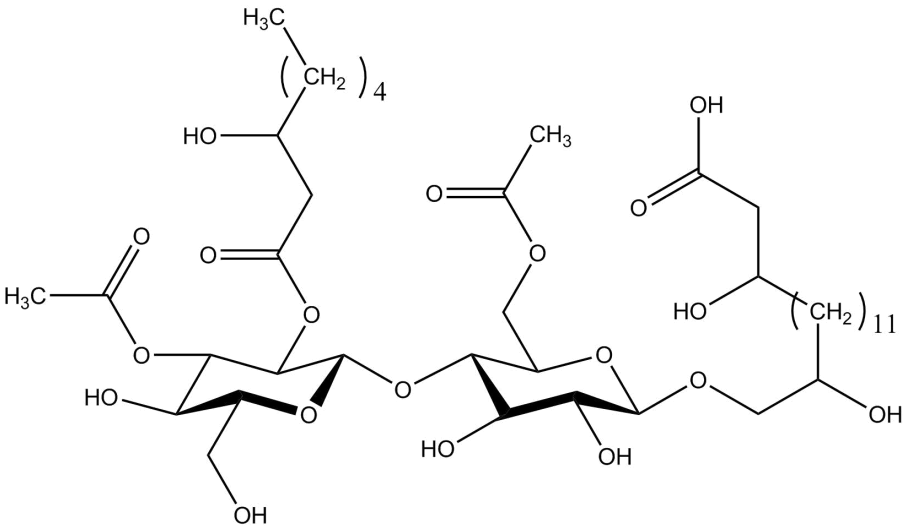

Figure 4.3: CBL flocculosin, produced by *Pseudozyma flocculosa*.

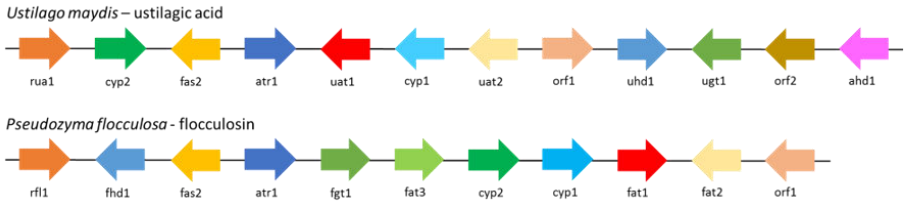

Figure 4.4: CBL cluster as seen in *Ustilago maydis* and *Pseudozyma flocculosa*. The ustilagic acid pathway has 12 genes, while the flocculosin pathway has 11. Arrows representing genes are colour coded for comparison between clusters and indicate realtive orientation.

4.3 Methods

CBL gene sequences from *U. maydis* and *P. flocculosa* were extracted from GenBank and used as queries for BLAST searches of the nr/nt and wgs databases. This search was initially focused on publicly available genomes of reported CBL producers and their close relatives (*M. aphidis* DSM70725, *M. antarctica* JCM10317 and T34, *U. esculenta*, *Vanrija humicola*/*Cryptococcus humicola*, *Trichosporon porosum*, *P. fusiformata*, *P. brasiliensis*, *S. scitamineum*, *P. hubeiensis* SY62, and *P. tsukubaensis*). The resultant sequences (if found as a gene cluster) were then aligned using MUSCLE v3.8.31 (Edgar 2004), with default options, and used to create Hidden Markov Models (HMMs) for each gene (*hmmbuild*). Those HMMs were then used to run a HMMER (v3.1b2, Eddy (1998, 2015)) search (*nhmmer*) within the NCYC genome dataset, looking for homologues of each gene. Top hits (1 per gene per strain) were extracted for further analysis using a custom script (mapCoordinates.py, part of the FindClusters Pipeline described in Chapter 6).

Species	CBL type(s)	Whole genome sequence available	Reference
Ustilago maydis	Ustilagic acid	Public	Haskins (1950)
Pseudozyma flocculosa (*Anthracocystis flocculosa*)	Flocculosin	Public	Mimee et al. (2005)
Pseudozyma aphidis (*Moesziomyces aphidis*)	Unknown	Public, This study	Morita et al. (2013a)
Pseudozyma hubeiensis	Unknown	Public, This study	Morita et al. (2013a)
Pseudozyma graminicola (*Sporisorium graminicola*)	Unknown	Public	Golubev et al. (2008)
Pseudozyma fusiformata (*Kalmanozyma fusiformata*)	Unknown	Public, This study	Kulakovskaya et al. (2005)
Sympodiomycopsis paphiopedili	Unknown	Public	Kulakovskaya et al. (2004)
Cryptococcus humicola	Unknown	Public	Puchkov et al. (2002)
Trichosporon porosum	Unknown	Public	Kulakovskaya et al. (2010)

Table 4.1: CBL production in yeast-like Basidiomycetes. If the species name used in the paper referenced differs from the strain's current name (according to Mycobank.org), the current name is noted in parentheses.

Gene trees were estimated from the aligned sequences found for each gene, using RAxML (Stamatakis 2014) with 100 bootstraps and model set to GTR + GAMMA (other options: -f a, -x 12345, -p 12345). Model selection was performed with PartitionFinder2 (options: model_selection=BIC, models=all, search=greedy, Lanfear et al. (2017)). Alignments were constructed and trimmed using MUSCLE v3.8.31 (Edgar 2004), with default options, and trimAl v1.2rev59 (Capella-Gutierrez et al. 2009), with the -strict option, respectively. Conversion to phylip format prior to tree estimation was done using Geneious (Kearse et al. 2012).

Unknown genes (predicted using AUGUSTUS v3.2.2 with *U. maydis* as model species, Stanke et al. (2008)) associated with the CBL gene cluster in some strains (*C. humicola* and *T. porosum*) were run against the nr/nt NCBI database using BLASTn, and against the Pfam database at [http://pfam.xfam.org/search], in an attempt to identify them.

4.4 Results & Discussion

Gene cluster identification

BLAST and HMMER searches of publicly available Ustilaginomycete genomes and the assembled genomes of 967 NCYC strains (and additional non-NCYC strains sequenced in this study) have shown that gene clusters resembling those found in *U. maydis* and *P. flocculosa* are present in the genomes of several Ustilaginomycete and Tremellales strains (see Table B.1). Outside these two widely separated taxa though, there is little evidence of any clustered CBL pathway, corroborating the experimental evidence mentioned previously that CBL production has only been reported in these two groups. However, while clustered pathway genes were only observed in the Ustilaginomycetes and Tremellales, many of the genomes searched contained matching sequences to at least one of the CBL genes.

Homologous sequences were found in most genomes searched, and at several loci in each genome. This means that, unlike the MEL genes discussed in the previous chapter, the majority of the CBL gene cluster is made up of genes that are members of larger gene families. The exception to this is *fat3*, found in the *P. flocculosa* gene cluster, which is found in no other genome. Several matching sequences are returned from the homology searches, but all are matches to a sub-region of *fat1*, presumably due to high sequence similarity between the two genes (*fat3* is likely a simple duplication of *fat1*). This presumably means that the duplication of *fat1* to create *fat3* is unique to *P. flocculosa*. Similarly, *orf2*, which is associated with the *U. maydis* gene cluster but not confirmed as part of the pathway, matches almost nothing (partial matches present away from the gene cluster, only in NCYC3431, NCYC1384, and NCYC2581).

Results from the FindClusters pipeline are shown in Figure 4.5. This figure has been truncated to remove strains identified as contaminated (Kraken analysis described in previous chapter). Strains outside the CBL producer groups (Ustilaginomycetes and Tremellales) have also been removed, to provide clarity on the gene arrangements present in those groups. It is assumed that gene matches outside these groups are due to shared domains within the gene families to which the CBL genes belong. In any case, no genuine clustering was observed in any of the strains removed (they were included by the pipeline due to having two genes on the same contig, but manual checks show that the two genes are separated by vast distances, so this is coincidental and no evidence of clustering).

It appears that the seven genes *ahd1*, *fhd1*, *fas2*, *cyp1*, *cyp2*, *atr1* and *orf1* are common amongst all of the species shown. This suggests that these genes have been present in the Basidiomycetous yeasts for a considerable time. Furthermore, while there are syntenic links between many gene pairs (*cyp2*,*fas2*; *fas2*,*ahd1*; *fas2*,*orf1*; *ahd1*,*cyp1*; *ahd1*,*atr1*, *cyp1*,*orf1*), large genomic distances are also common between them. In a particularly striking example, a

genomic contig of NCYC 3833 harbours copies of *orf1*, *fhd1* and *cyp1* in an 5.5Mbp region. It must be said however, that there is no confirmation that these dispersed genes are CBL pathway genes (as opposed to just close homologues). This is something that needs to be tested before any concrete conclusions may be drawn. Moving up the figure from the bottom, the growing cluster appears to become more compact and the genes *fgt1*, *fat1*, and *fat2* are recruited to join it. These genes were originally present in an uninterrupted cluster (e.g. as in NCYC 3252), and only later became partially separated by the insertion of *cyp1* between *fat2* and *fat1*. The gene *rfl1* looks to have been most recently recruited to the cluster, taking terminal or sub-terminal positions within it.

Intriguingly, both T. inkin genomes show CYP1 and CYP2 in tandem, a feature only seen elsewhere in the P. flocculosa genome. This may suggest the two cytochrome P450 genes are the product of a duplication (and then moved apart). Alternatively, it may simply be that these strains lack a good match for either *cyp1* or *cyp2* and what we see in Figure 4.5 is the two genes being mapped to the same region (as we see with *fat3* mapping to *fat1* in species that lack *fat3*). Alignment of the two CYP genes from the *U. maydis* gene cluster is fairly poor, suggesting the latter explanation is correct.

Gene tree estimation

Phylogenetic trees were estimated for each gene in the CBL gene cluster, in order to determine whether the genes in the gene cluster have evolved in line with the rest of the genome (i.e. by comparing to a species phylogeny). Given the lack of matching sequences, gene trees were not estimated for *orf2* and *fat3*. The trees are presented in the order in which the genes are found in the *U. maydis* gene cluster, see Figures 4.6-4.16. They are also intended to show the relatedness of the Tremellales genes, thus determining whether the gene cluster found in that group is homologous with that of the Ustilaginomycete species (compounds of the same type are being produced, but are they convergent?) A finding that the two gene clusters are homologous (despite the vast evolutionary distance between the groups) could imply horizontal gene transfer, potentially at the gene cluster level. The alternative explanation would be a joint origin with massive losses in intervening clades.

The gene trees are in general agreement that the Ustilaginomycete and Tremellomycete genes group separately and in accordance with the species tree in Figure 4.1, for example *U. maydis* & *P. hubeiensis* consistently group together, as do *P. tsukubaensis* & the novel *Pseudozyma*, while in the Tremellomycete branches *T. montevideense* & *T. domesticum* are consistent sister species. This is evidence that they have evolved side by side rather than

originating from horizontal transfer (in which case I would expect one of the two groups to be clustered within the other). The trees do suffer from poor bootstrap support in places, perhaps due to marked differences in sequence causing alignment difficulties (in several cases, the trimmed alignment was just a small fraction of the overall gene length). In any case, the evidence suggests that the two gene clusters have evolved independently, from similar building blocks.

Gene cluster arrangement

The gene clusters responsible for CBL production had previously been described in *U. maydis* (Teichmann et al. 2007) and *P. flocculosa* (Teichmann et al. 2011). These two gene clusters shared most of their constituent genes, albeit in a different arrangement. This study aimed to determine the arrangement in other CBL producers. We can see in Figure 4.5 that all other Ustilaginomycete CBL producers follow the pattern seen in *U. maydis*. This suggests that the CBLs being produced by these species will be more similar to ustilagic acid than to flocculosin. It also suggests that the flocculosin gene cluster is somewhat unique, at least in this dataset. *Anthracocystis*, the genus into which *P. flocculosa* has now been placed (Wang et al. 2015), is perhaps a more derived clade that has acquired some novel characteristics versus the rest of the Ustilaginomycetes. *P. flocculosa* itself is known to have transitioned to a biocontrol/commensal role on the plant leaf (versus Botrytis pathogens, etc.) from the general Ustilaginomycete smut pathogen lifestyle (Lefebvre et al. 2013). This may explain the marked difference seen in this gene cluster. Further sampling of the *Sporisorium* genus (closely related to *Anthracocystis*) may be needed to rule out further spread of the flocculosin gene cluster structure, although evidence from *S. scitamineum* shown here suggests that the arrangement is indeed unique to *P. flocculosa*.

The gene cluster as found in the Tremellales species sampled here seems to differ from that of *U. maydis* and *P. flocculosa*, and is discussed in more detail next.

The Tremellales CBL gene cluster

Clearly from these results (see Figure 4.5) the gene cluster present in the Tremellales species is different to that of the Ustilaginomycetes. Despite reports of CBLs being produced, only a partial gene cluster is apparent in these species, and what is found varies substantially between species. The unknown genes found associated with the *T. porosum* and *C. humicola* gene clusters (checked via BLAST) did not appear to be obvious candidates for metabolic pathway constituents. See Figure 4.17 for a representation of the cluster as seen

in *Trichosporon porosum*. The results from the NCYC search are inconclusive, with genome assembly fragmentation preventing unambiguous gene cluster identification. In some cases, for example NCYC2903 (*T. asahi*) and NCYC3086 (*C. longus*), part of the Tremellomycete CBL gene cluster is evident, while in others, for example NCYC2638 (*T. montevideense*) and NCYC2665 (*T. domesticum*), the same genes are more spread out. This may be an indication that the genes in this biosynthetic pathway have been moved closer together across taxa, as seen in Wong et al. (2005), but this requires further investigation to confirm. In any case there is no clear resolution of the entire gene cluster in any of the new genomes searched. These results do however offer tentative confirmation that CBL production in the Tremellomycete species is underpinned by a gene cluster that is noticeably different to the previously described examples in *P. flocculosa* and *U. maydis*.

The CBLs produced by the Tremellomycetes differ somewhat compared to those of the Ustilaginales. There is no acyl chain, potentially explaining the lack of *fas2* and *uhd* homologues in the Tremellomycete gene cluster (note that lots of *fas2* homologues were found by FindClusters, Fig. 4.5, but none were clustered and may not be pathway related).

The CBL gene cluster in *P. aphidis*

P. aphidis was previously reported to be a producer of CBLs, as shown by Morita et al. (2013a). However, searches of several *Moesziomyces* (the genus to which *P. aphidis* now belongs according to Wang et al. (2015)) genome sequences (Public: DSM70725, T34, JCM10317. NCYC: NCYC911, NRRL-Y17626) have not revealed a CBL gene cluster, or even a dispersed biosynthetic pathway. Matching sequences are apparent for several of the CBL genes in these strains (*rfl1*, *cyp2*, *fas2*, *atr1*, *ahd1*, *cyp1*, *orf1*, and *fhd1*) but these are widely dispersed and do not group together on the aformentioned gene trees with the clustered homologues found in other species. This suggests that these matching sequences are other members of their respective gene families and are not part of any CBL biosynthesis pathway in the *Moesziomyces*.

4.5 Conclusion

This chapter aimed to answer four main questions: how widely spread is the ustilagic acid/flocculosin gene cluster? How does that gene cluster vary in terms of gene order and content? Is the same gene cluster present in distant CBL producers (Tremellales)? Does *P. aphidis* posess the CBL gene cluster?

The search for the CBL gene cluster across the NCYC collection's sequenced genomes revealed that it is found in the same approximate taxonomic group as the MEL gene cluster, among a slightly wider range of species. Generally the results seem to show minimal variation in gene order within the Ustilaginales, with the flocculosin gene cluster being a notable outlier in this respect.

It was already known that CBL production is not confined to the Ustilaginales, unlike with MEL production. In this case, similar products are produced by species in the Tremellomycetes. Gene trees show that the Tremellomycete genes are likely to be evolutionarily separate from the previously described CBL genes. In such widely separate lineages, it seems that two pathways have independently been assembled from common gene types, making a chemically convergent product. As with the MEL gene cluster of the previous chapter, long read sequencing would be useful in getting a better picture of the CBL gene cluster in those genomes that were too fragmented to be of use in this study.

The story for *P. aphidis* and its relatives is no clearer. Extensive searches of multiple genome sequences did not reveal any clues as to the whereabouts of the CBL biosynthetic pathway in this species. Given the taxonomic proximity of *P. aphidis* to the rest of the ustilaginomycete CBL producers, it seems strange that the pathway might be non-homologous, but this seems the best conclusion based on the evidence presented here. This non-homologous biosynthetic pathway may also be a dispersed one, which may explain why it does not appear in other gene cluster searches.

In contrast to the MEL gene cluster genes discussed in the previous chapter, the search for the genes making up the CBL gene cluster has revealed that many of them are clear members of widespread gene families. Homology searches find multiple hits in most genomes searched, suggesting that the genes making up this gene cluster are largely similar to their ancestral sequences and have retained key domains.

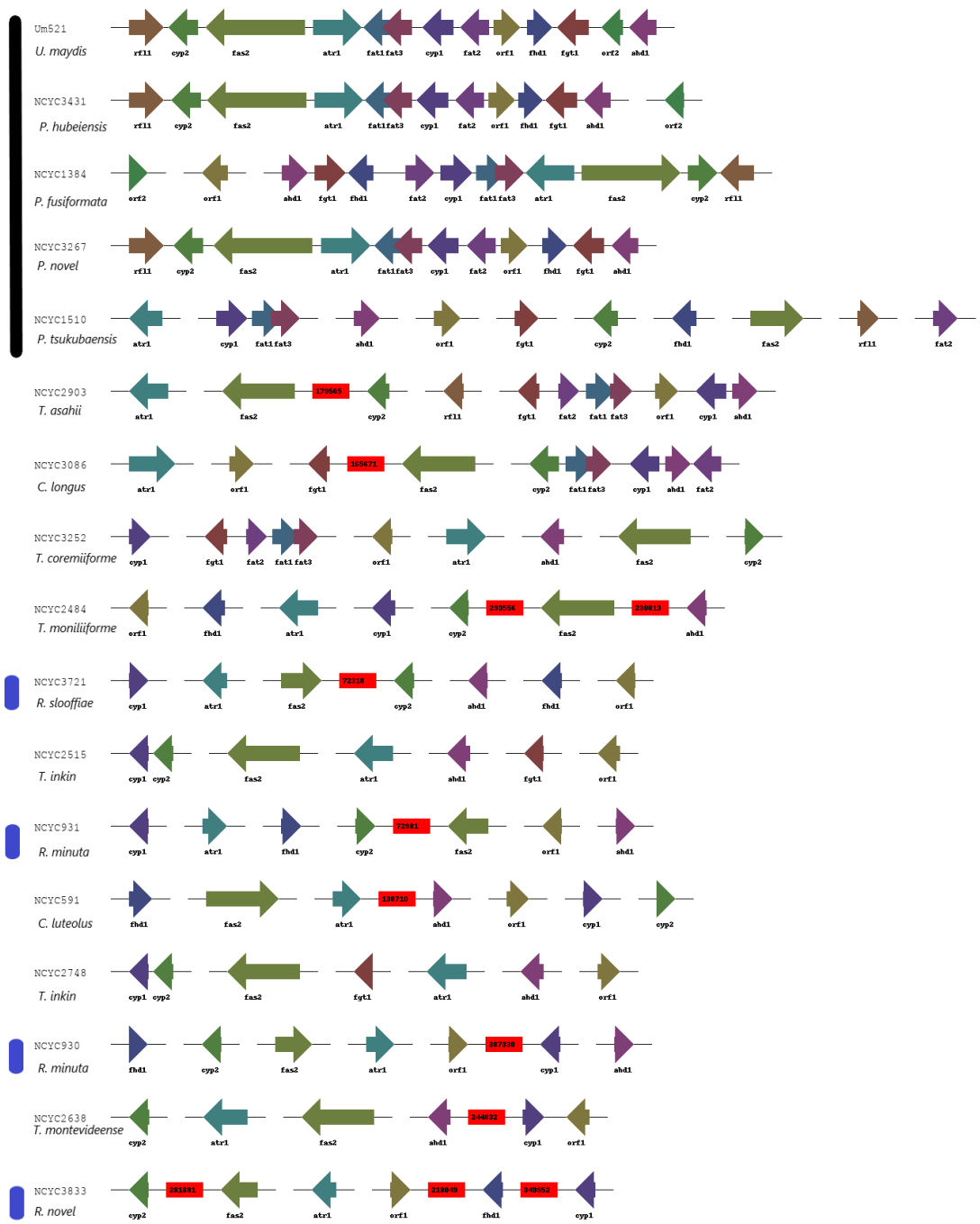

Figure 4.5: Continued on next page.

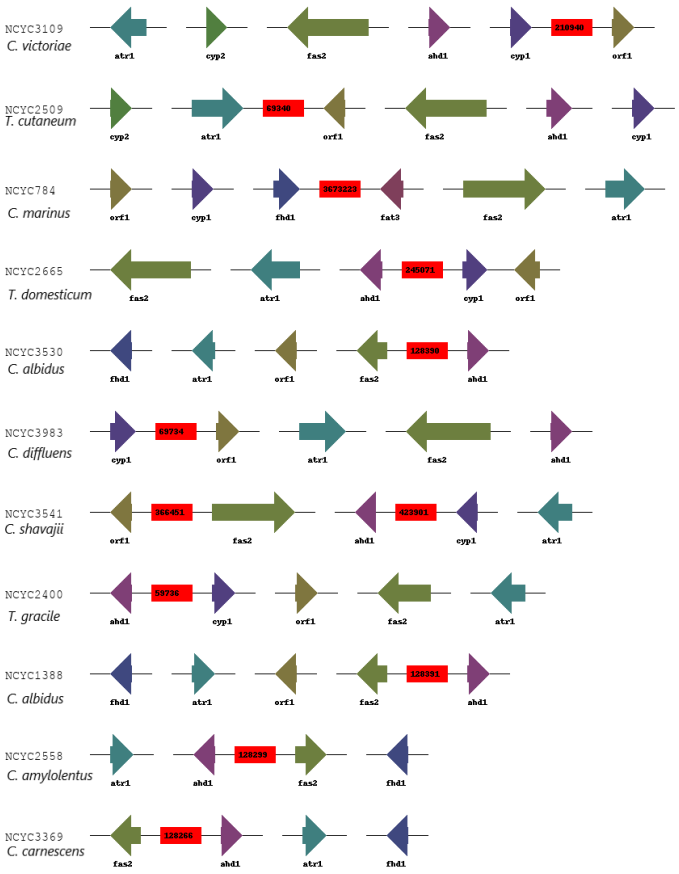

Figure 4.5: Results from the FindClusters pipeline described in Chapter 6 of this thesis. The analysis has been run on all sequenced NCYC genomes (plus the 23 from other collections), with *U. maydis* included for reference. Colours indicate homology, arrow direction indicates gene orientation, genes connected by a black line are found on the same contig. Red rectangles indicate distance between genes in cases where genes are found on the same contig but separated by more than 10,000 nucleotides. The first 5 strains are Ustilaginomycetes (black bar), NCYC930, NCYC931, NCYC3721, and NCYC3833 are Rhodotorula (blue bars), and the rest are Tremellomycetes.

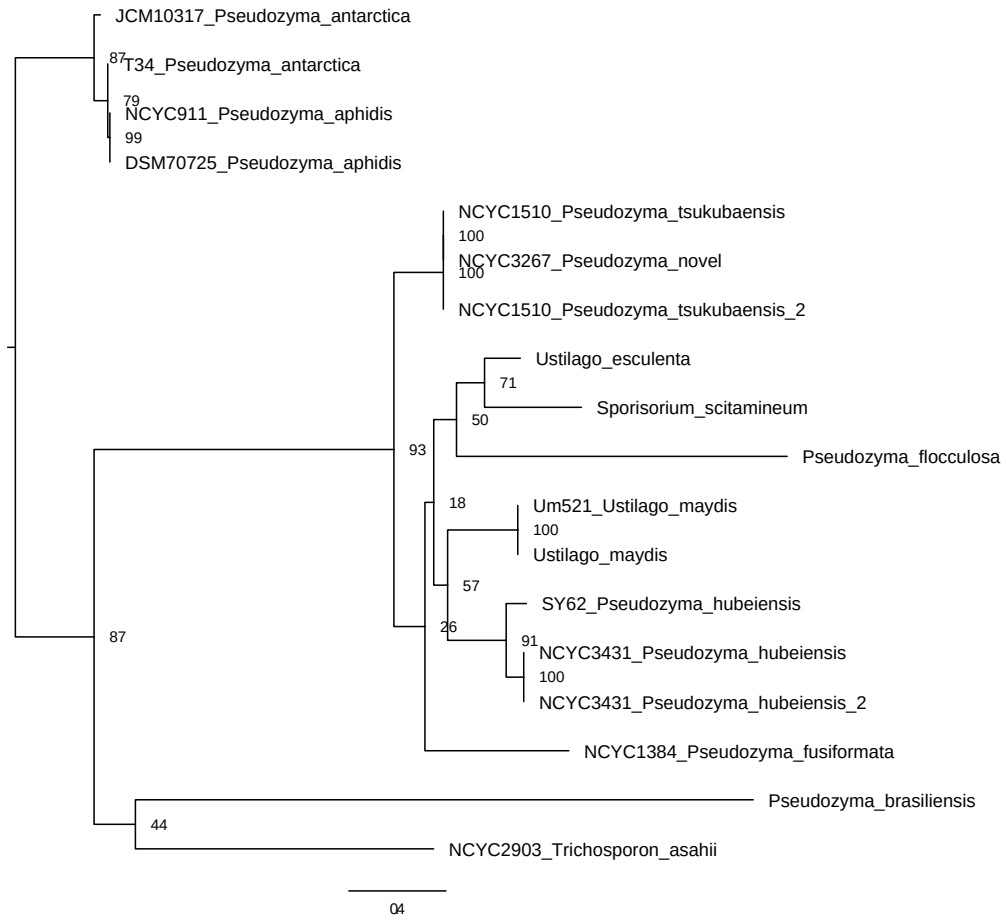

Figure 4.6: RFL1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *rfl1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis rfl1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -14051.011882.

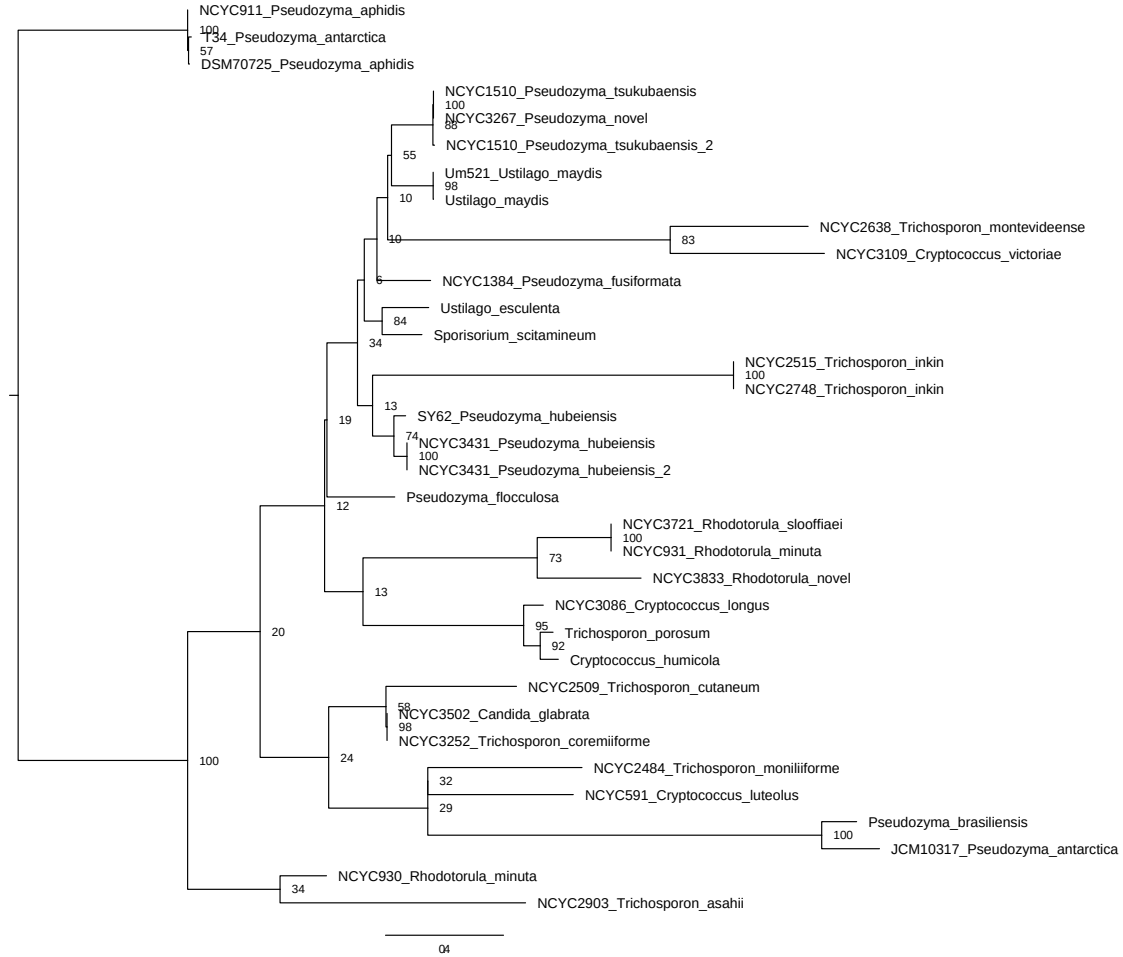

Figure 4.7: CYP2 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *cyp2* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis cyp2* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -11952.639400.

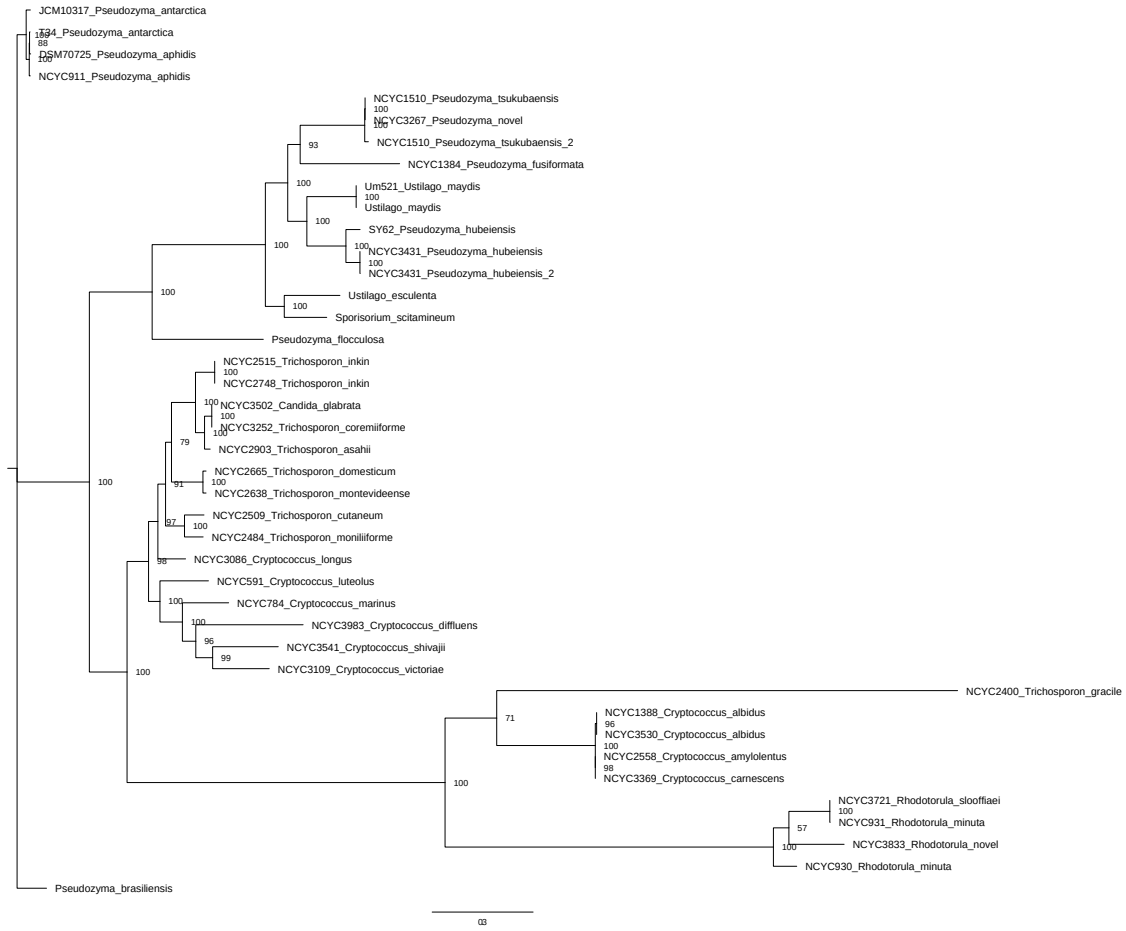

Figure 4.8: FAS2 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *fas2* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis fas2* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -83912.007835.

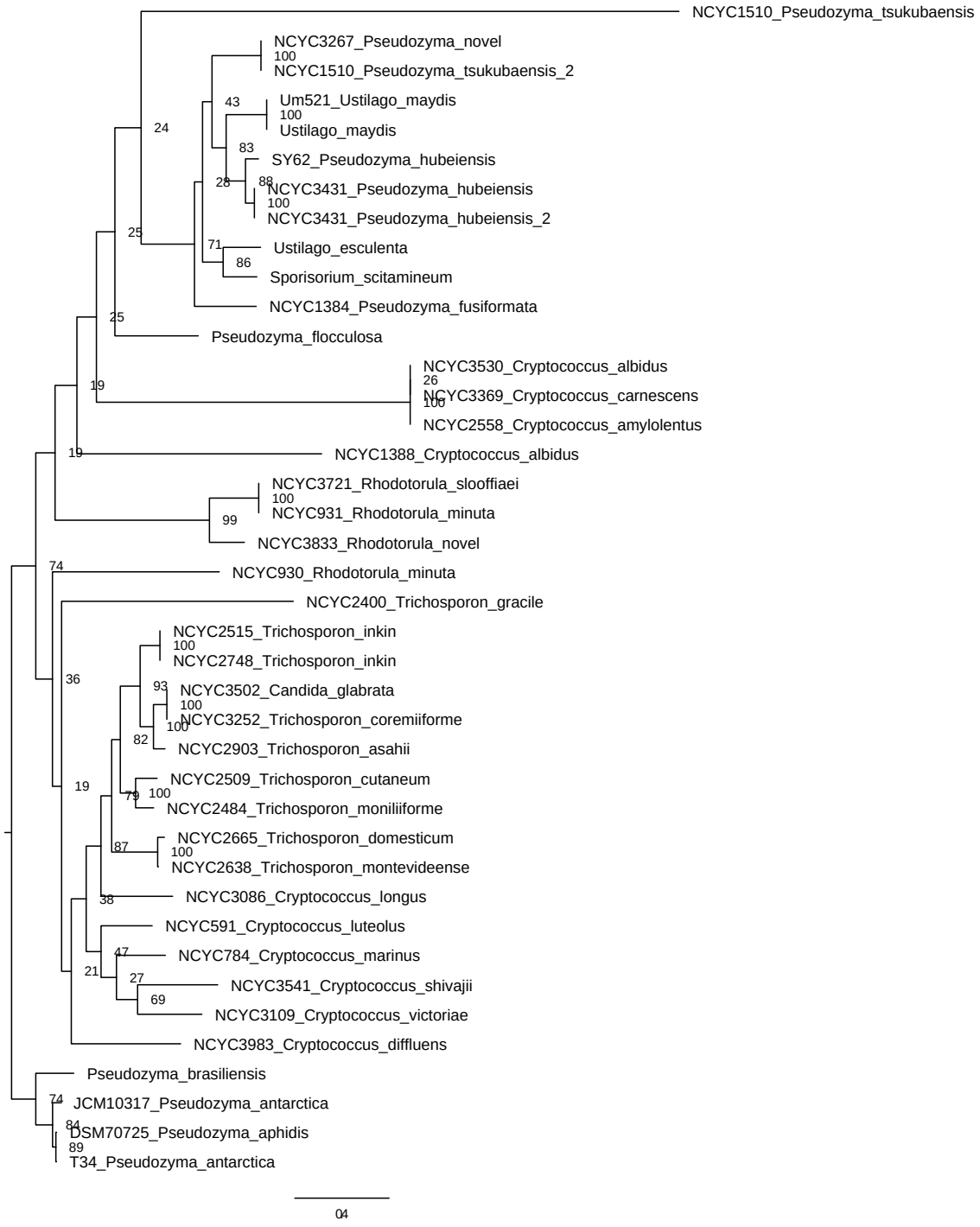

Figure 4.9: ATR1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *atr1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis atr1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -22788.974411.

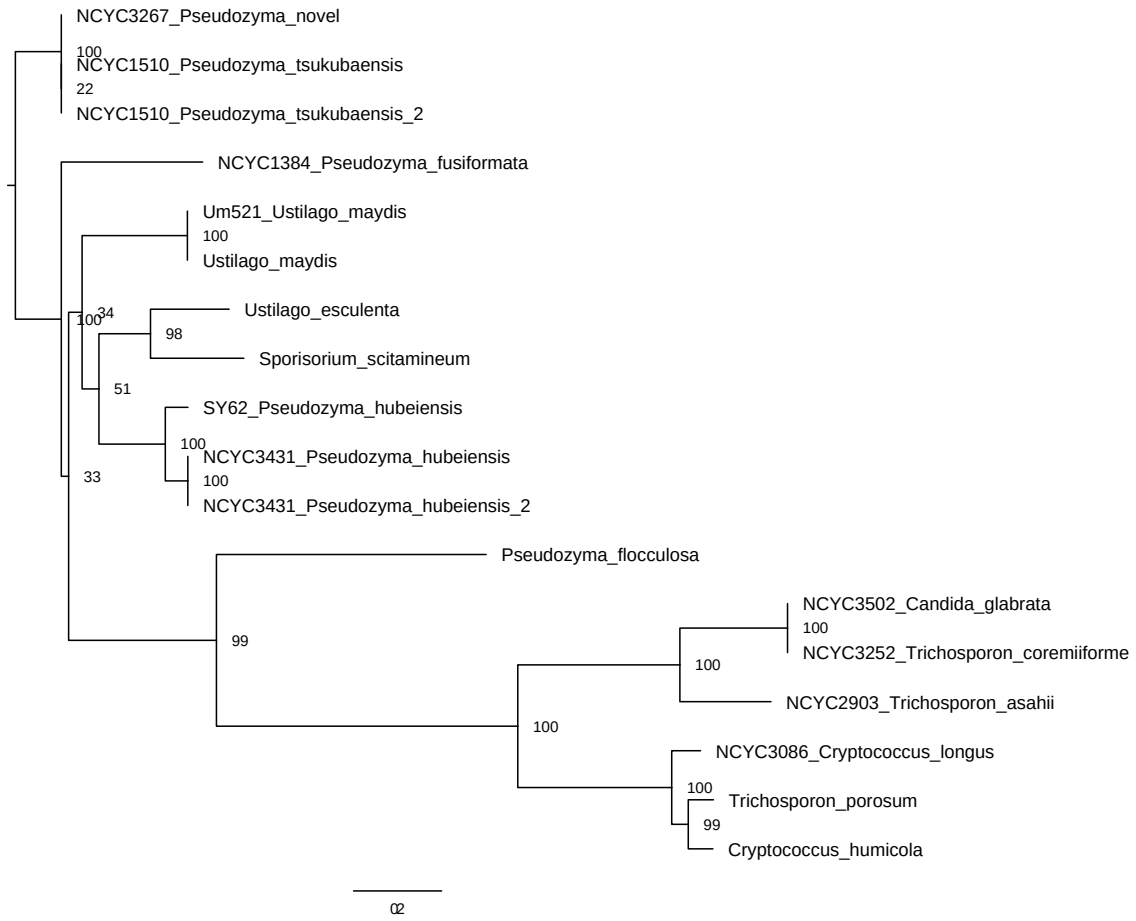

Figure 4.10: FAT1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *fat1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis fat1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -11826.209468.

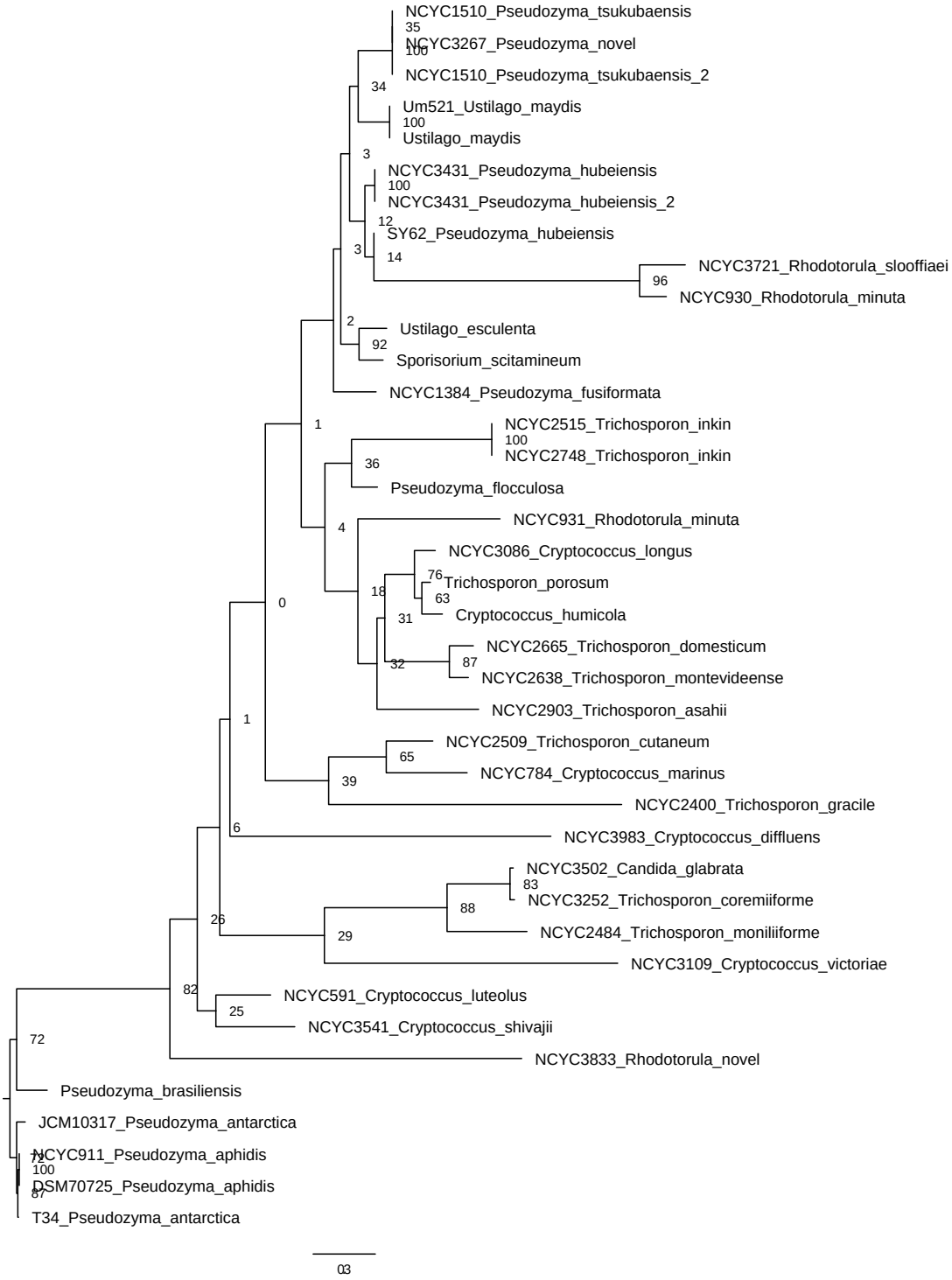

Figure 4.11: CYP1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *cyp1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis cyp1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -17067.493738.

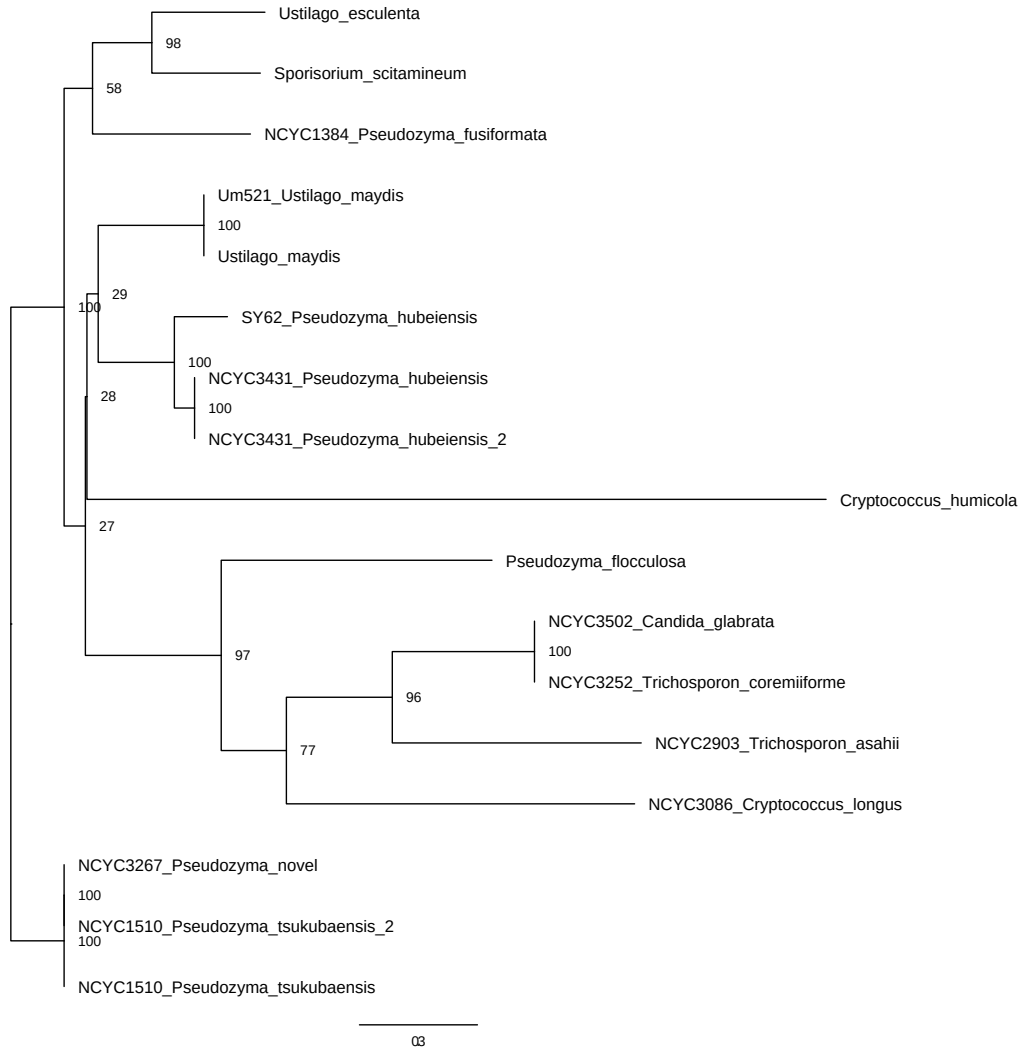

Figure 4.12: FAT2 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *fat2* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis fat2* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -10478.607431.

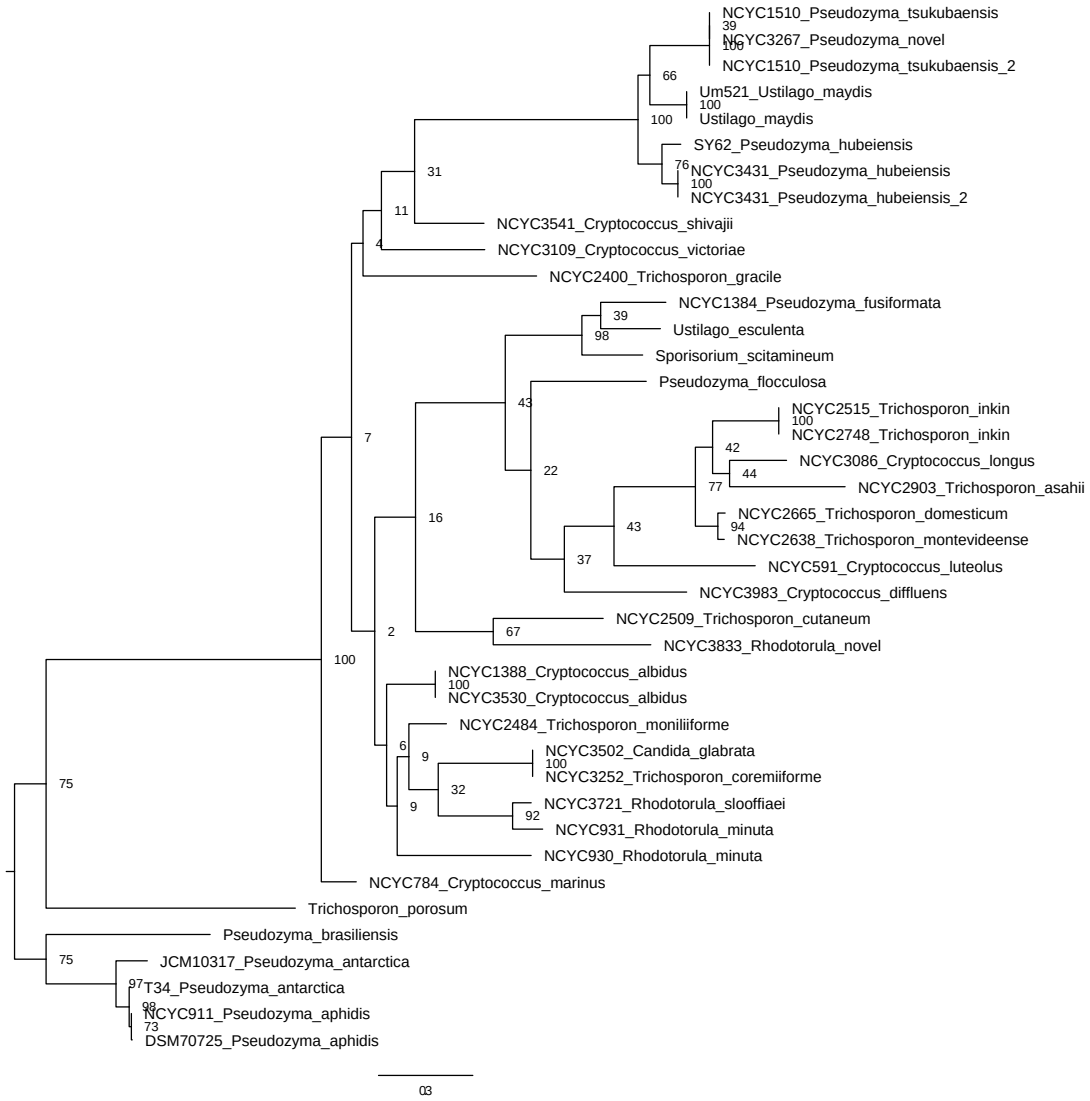

Figure 4.13: ORF1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *orf1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis orf1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -13094.364634.

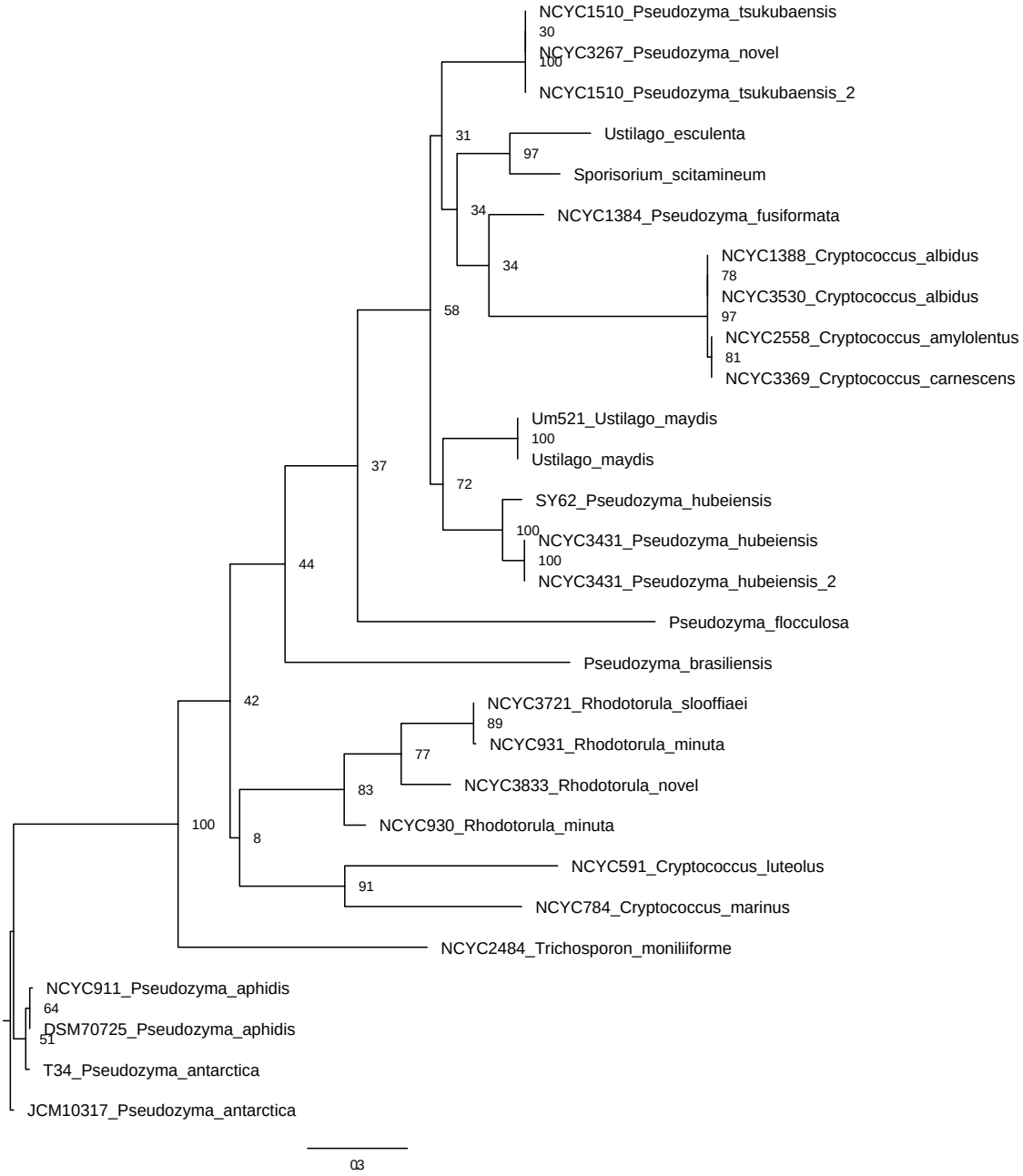

Figure 4.14: FHD1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *fhd1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis fhd1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -7596.477872.

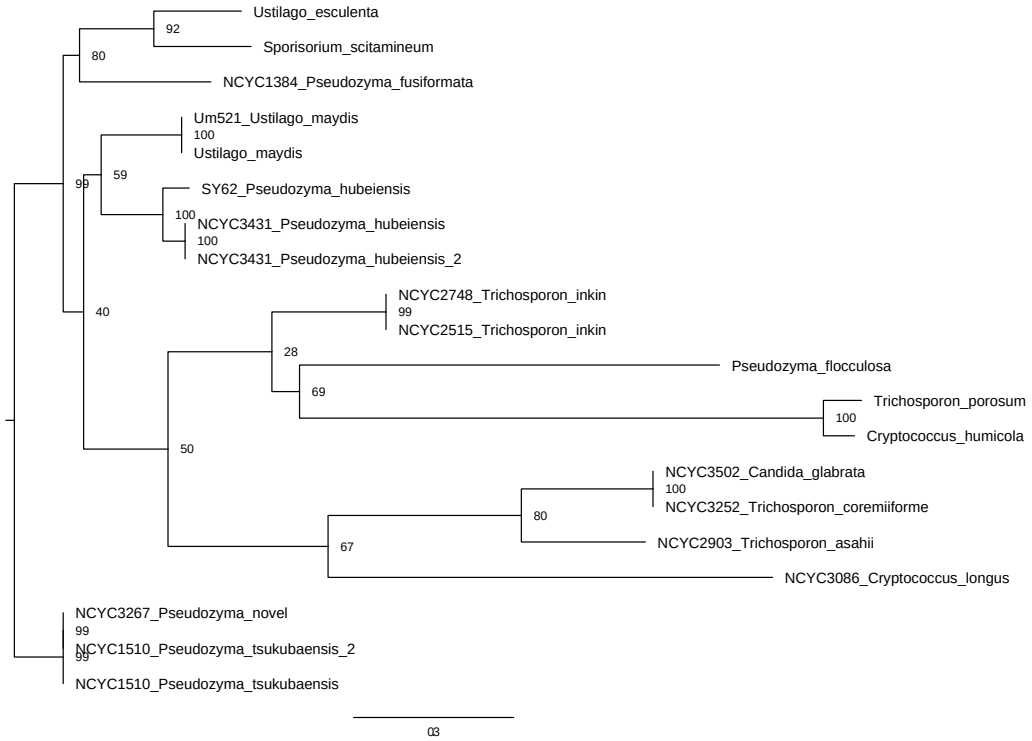

Figure 4.15: FGT1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *fgt1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis fgt1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -11685.236691.

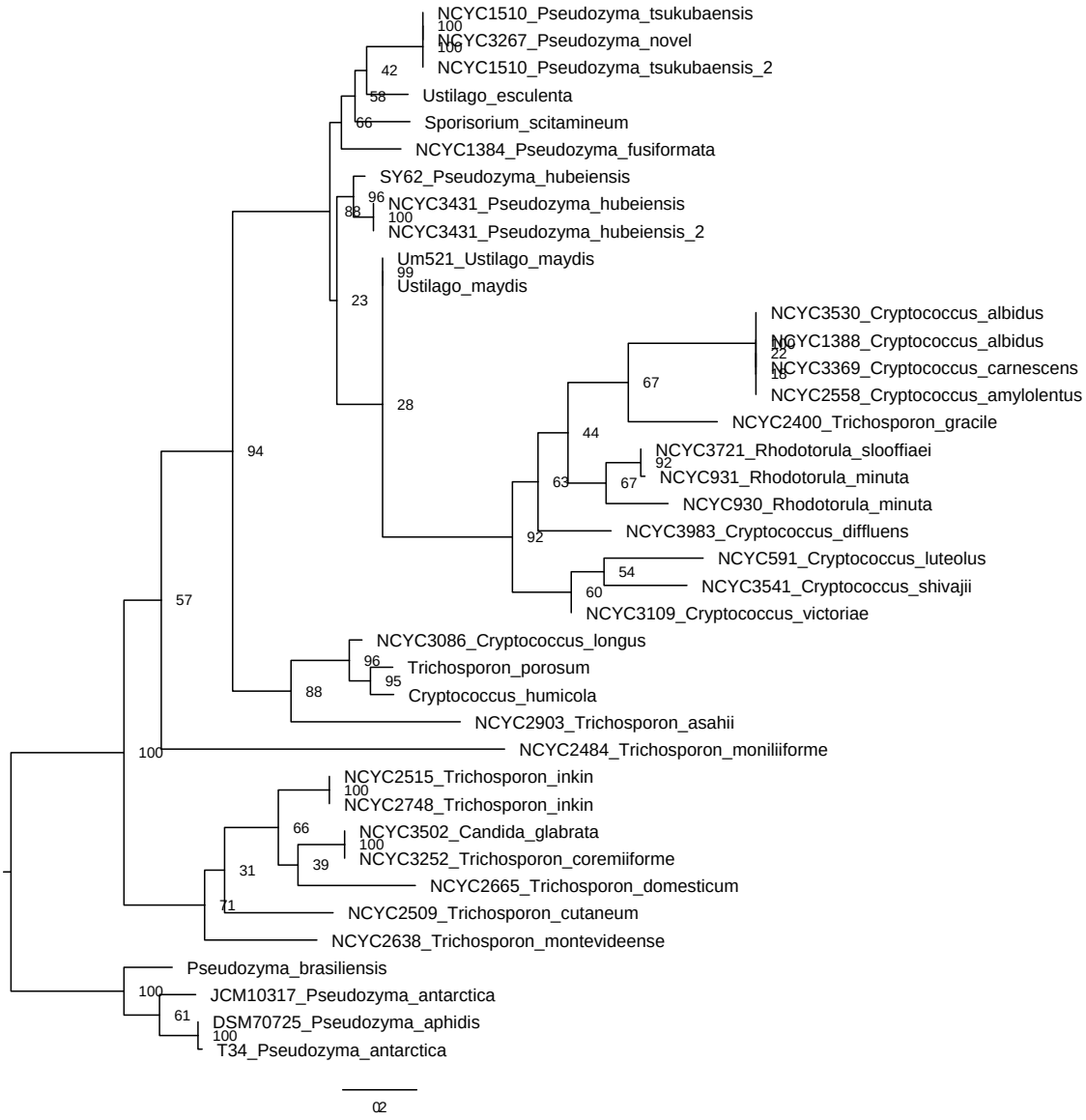

Figure 4.16: AHD1 gene tree: estimated from sequence hits from an nhmmer search of all sequenced NCYC genomes, using the *ahd1* HMM generated earlier, and from blastp and tblastn searches of the nr and wgs databases, using the *U. maydis ahd1* gene as query. Nodes are labelled with the bootstrap support values (out of 100). Final ML Optimization Likelihood: -7840.837001.

Figure 4.17: CBL cluster as seen in *Trichosporon porosum* and (approximately) *Cryptococcus humicola*. Only a partial cluster is apparent. Arrows representing genes are colour coded for comparison between clusters and indicate direction. Black diamonds represent predicted genes of unknown identity.

5 Cytochrome P450 content of the NCYC collection

5.1 Introduction

In this chapter, the sequenced genomes of the NCYC collection were searched for Cytochrome P450s (CYPs), with any found being classified according to the method of David Nelson (David Nelson, Personal Communication), a leading authority on CYP genes. The aim of this chapter is to catalogue the cytochrome P450 content of the NCYC collection (the ~1000 sequenced genomes described in earlier chapters serving as a representative sample of the whole collection). A key question then to be addressed is whether the number of cytochromes P450 correlate with taxonomy. Finally, the number of unknown CYPs in each strain has been assessed, again asking whether the number is associated with taxonomy, which would indicate undersampled clades.

Cytochrome P450s (CYPs) are a large superfamily of proteins that bind heme and generally function as part of electron transfer chains (as the final oxidase enzyme) or as modifying enzymes within biosynthetic pathways (Cook et al. 2016). They are also well known, particularly in fungi, as part of toxin production and degradation pathways (Shin et al. 2018). In humans, insects, plants, and other organisms, CYP enzymes themselves may function to metabolise xenobiotics, toxins, herbicides, pesticides, etc. (Shin et al. 2018). The 3D structure of CYP51 is shown as an example in Figure 5.1. As such, CYP-encoding genes are frequently found in metabolic gene clusters of the type being investigated in this thesis. Examples of metabolic gene clusters containing CYPs in fungi include the cellobiose lipid gene cluster described in the previous chapter (Teichmann et al. 2007, 2011) and the sophorolipid gene cluster (Van Bogaert et al. 2013). Examples in plants include the gene clusters for thalianol, avenacin, and momilactone (Nützmann et al. 2016), which contain CYP85, CYP51, and CYP99A2, respectively. See Figure 5.2 for diagrammatic representations of some of these examples, with the CYPs highlighted. CYPs are also known to be part of mycotoxin production pathways, for example aflatoxin in *Aspergillus* (Shin et al. 2018).

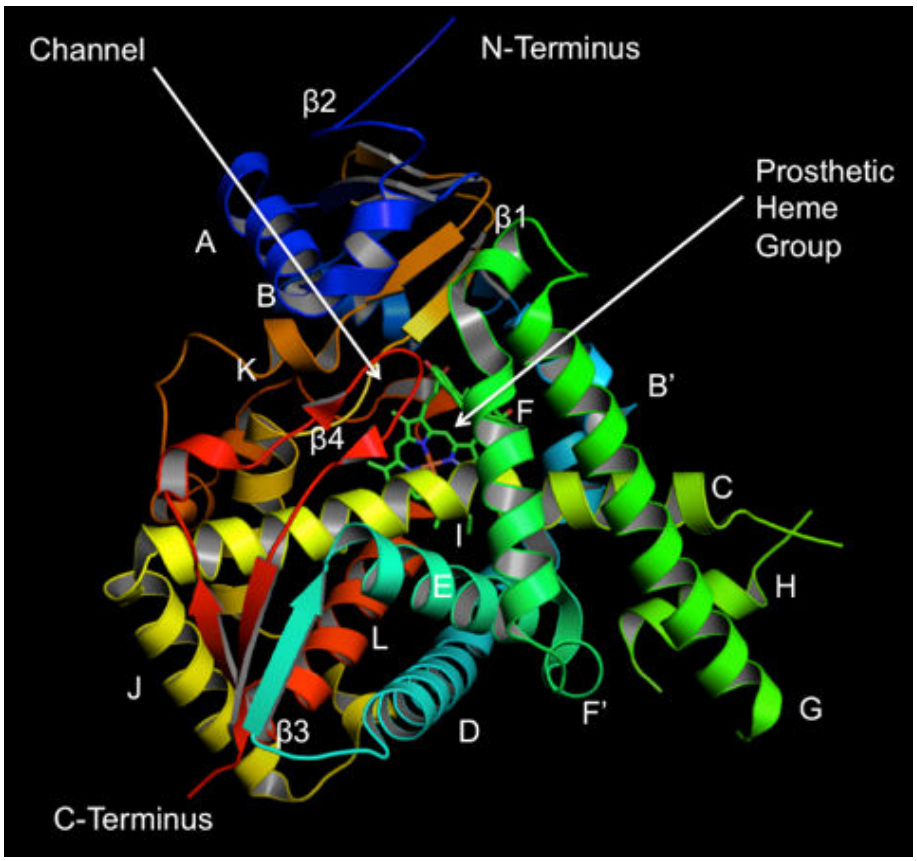

Figure 5.1: 3D structure of a representative cytochrome P450 protein. Specifically, this is lanosterol 14A-demethylase (CYP51), found in almost all fungi as a crucial cell wall development protein. It is thus the target of several antifungal drugs (Shin et al. 2018). Image taken from the Cytochrome P450 Wikipedia page at `https://en.wikipedia.org/wiki/Cytochrome_P450` and shared under the Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) licence.

CYPs are found in varying numbers in the vast majority of living organisms, ranging from several hundred in some animals and plants (*Oryza sativa* has 334) to just three in *S. cerevisiae* (Shin et al. 2018). Despite exhibiting quite extraordinary sequence diversity at the amino acid level, particularly in fungi (Shin et al. 2018), they are easily identified by the presence of protein motifs that control their 3D structure, AGXDTT, EXXR, PERW (the W being a fungi-specific variant), and FXXGXRXCXG, where X is any amino acid (Kelly et al. 2009), see Figure 5.3 for further information about the relative positions and makeup of the motifs. CYPs are classified into families and subfamilies by way of amino acid sequence similarity, and named in the style CYP52A1 (52 is the family, A is the subfamily, 1 is the gene itself) (Shin et al. 2018). As of 2018, more than 85,000 fungal CYPs had been identified (Nelson 2018).

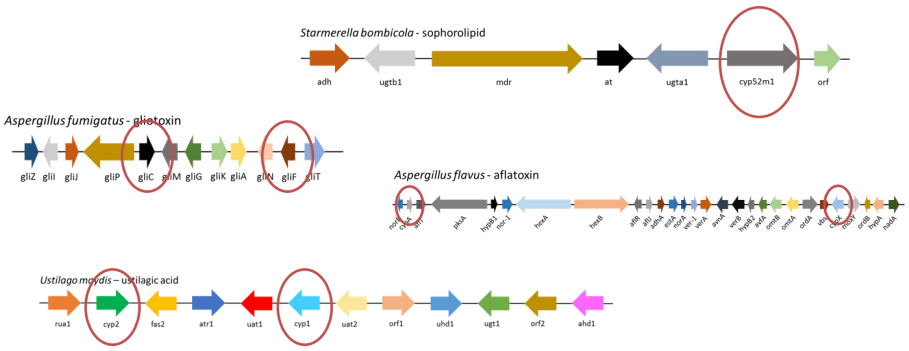

Figure 5.2: Cytochrome P450 genes are found in several fungal metabolic gene clusters. CYP genes in these gene clusters are noted with a red circle. No homology is implied by colour across gene clusters in this figure.

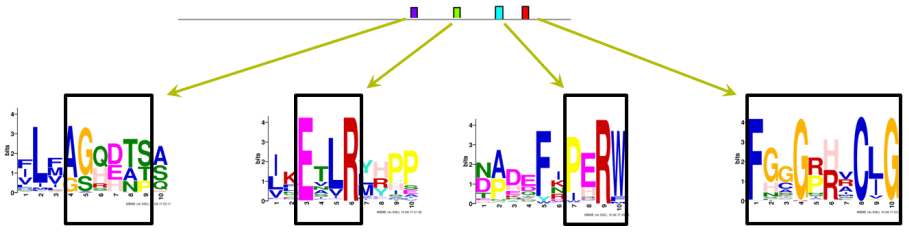

Figure 5.3: The relative positions of the four amino acid motifs found in cytochromes P450. These motifs, and their relative positions in the protein, are very strongly conserved due to their crucial role in the 3D structure of the protein. The rest of the amino acid sequence is, conversely, highly diverse.

Given that CYPs are a fairly common component of metabolic or biosynthetic gene clusters, it seems sensible to use them as potential nucleation points in the search for unknown gene clusters. Since they are easily identified, they can be easily found within a new genome and the region surrounding each one can be subjected to a more rigorous search for biosynthetic pathways.

5.2 Methods

Known fungal (yeast) CYP sequences were downloaded from the David Nelson database at `http://drnelson.uthsc.edu/CytochromeP450.html` and from GenBank and manually compiled to a FASTA file (169 sequences). These sequences were aligned using MUSCLE v3.8.31 (Edgar 2004) with default options enabled. HMMER v3.1b2 (Eddy 1998, 2015) was used to make an HMM profile (*hmmbuild* with default options - known CYP FASTA alignment converted to STOCKHOLM format for input), which was subsequently used to search the sequenced genomes of ~1000 NCYC strains for putative CYPs (*nhmmer* with default options). The resulting sequences were collected from the genome assemblies using

the mapCoordinates.py script from the FindClusters pipeline described in Chapter 6.

Protein sequences were predicted for each strain using AUGUSTUS v3.2.2 (Stanke et al. 2008) with an appropriate reference species selected according to each strain's taxonomic position, e.g. most Ustilaginales sequences were assigned the *U. maydis* reference for prediction purposes. Protein sequences were then extracted from GFF output via the proteinFromGFF.py script. Putative CYPs were confirmed by motif examination of their predicted protein sequences using MEME v4.11.4 (Bailey et al. 2009, 2015).

The CYP sequences were then categorised according to the amino acid identity with known CYP sequences (more than 55% identity puts sequences into the same sub-family, more than 40% into the same family only). This was done using a blastp (Altschul et al. 1990) search against a database constructed of all known fungal CYP sequences from the David Nelson database (not just yeast). The sequences were filtered by length to eliminate those shorter than 200 amino acids (roughly half the length of a complete protein) to remove short fragments that are not likely to correspond to functional genes, though it is worth noting that interesting candidates may be erroneously thrown out in this step.

The phylogeny built in Chapter 2 was annotated using the International Tree of Life (ITOL) annotation tool at `https://itol.embl.de/`. The annotation was derived from the CYP categorisation step described above and mapped to the leaves of the tree to show how CYP diversity is spread taxonomically.

5.3 Results & Discussion

Cytochrome P450 content of the NCYC collection

Approximately ten thousand (10,568) putative CYP sequences were identified from the initial HMMER search of 961 draft genome assemblies. Visual examination of the motif search results verified that all putative CYPs did contain at least some of the CYP motifs (missing motifs were usually due to short sequences, potentially broken up during the sequencing process). After filtering for length (>200aa), there were 5,068 sequences remaining. Of these, 4,572 sequences could be identified by sequence identity to at least family level, of which 3,792 were identified to subfamily level (see Table C.1 for the identified CYPs per strain, and Figure C.1 for the abundances of identified CYPs.). The remaining 496 share less than 40% identity with any known CYP protein in this dataset and therefore may be members of novel CYP families (Table 5.1).

Maximum classification level	Number CYPs
Subfamily	3792
Family	780
Unclassified	496
Total	5068

Table 5.1: Table showing the numbers of putative CYPs from the NCYC collection, broken down into the classification categories mentioned earlier. Here the number shown is the number of CYPs whose final classification is at the level stated, i.e. the total number classified to family level necessarily includes those further classified to subfamily level.

The 496 unidentified CYP sequences were BLASTed against the nr protein database (limited to fungal sequences to stay under CPU limit) to see if they matched anything outside of the fungal CYP dataset used previously. This reduced the number of truly unidentified sequences to 69, see Table 5.2. What is evident from this table is that the majority of unidentified CYP sequences belong to Basidiomycete species; only *D. anomala*, *K. servazzii*, and *S. lactativora* are Ascomycetes. Table 5.3 shows the 119 unique nr protein entries that were the top matches for the other 427 unidentified CYP sequences. Many of these top hits are hypothetical proteins, while those identified as P450s are not classified. This means that although these sequences clearly have homologs in the public databases, they still cannot be identified to family or subfamily level, though it is clear that they fall into distinct subgroups of their own (i.e. many of the unknown CYPs match to the same database entries).

Species	CYP ID	Top Hit ID	Percentage Identity	Alignment Length
Cryptococcus diffluens	NCYC3983_7215_4369-6081:1-1454_g2	XP_007006284.1	32.292	288
C. diffluens	NCYC3983_7547_1399-2613:2-1215_g3	XP_024742324.1	37.852	391
C. diffluens	NCYC3983_7636_4121-2401:2-1636_g6	BAL05196.1	33.419	389
C. diffluens	NCYC3983_7943_47099-46068:2-1032_g10	ORY21756.1	32.063	315
C. diffluens	NCYC3983_8086_70876-69410:562-1457_g14	SJL15675.1	34.962	266
C. flavescens	NCYC2963_8845_657695-658837:2-1096_g5	RSH93778.1	34.983	303
C. laurentii	NCYC2712_13267_1153086-1154228:2-1143_g21	RSH90033.1	36.364	341
C. longus	NCYC3086_2256_15361-16997:2-1637_g1	XP_028478670.1	38.294	504
C. longus	NCYC3086_2353_156758-158323:2-1566_g13	OCF40419.1	39.511	491
C. longus	NCYC3086_2356_904711-905705:2-995_g17	XP_028478670.1	38.393	336
C. luteolus	NCYC591_2222_48676-47918:2-759_g1	XP_018265046.1	37.688	199
C. luteolus	NCYC591_7655_480460-479081:2-1380_g7	ORY34740.1	37.562	402
C. macerans	NCYC578_40278_1238-341:2-898_g6	AFV26090.1	39.732	224
C. marinus	NCYC784_8708_827831-826684:2-1148_g9	RSH90033.1	39.665	358

C. peneaus	NCYC553_909_4522049-4523774:1-1726_g2	TFK67144.1	39.096	376
C. peneaus	NCYC553_917_212423-210966:2-1453_g5	XP_018998122.1	33.053	357
C. rugosa	NCYC2726_30498_12801-11865:1-937_g17	CCE81469.1	37.895	285
C. saitoi	NCYC3008_9454_17008-15200:417-1718_g2	ABE01888.1	36.441	236
C. saitoi	NCYC3008_9680_81930-83444:2-1505_g5	PBK65183.1	31.436	404
C. terreus	NCYC2442_55731_31481-30339:2-1143_g4	RSH90033.1	36.364	341
C. victoriae	NCYC3109_938_5217-3773:381-1280_g3	CCO36833.1	35.165	273
C. victoriae	NCYC3109_986_29690-28157:2-1296_g7	TPX14092.1	38.997	359
C. victoriae	NCYC3109_998_1119965-1121541:411-1496_g9	XP_012048170.1	32.66	297
Dekkera anomala	NCYC2_5873_2207-1339:1-869_g5	Q9UVC3.1	34.084	311
D. anomala	NCYC2_6467_3084-1774:1-1311_g7	RYO79165.1	39.593	442
Filobasidium uniguttulatum	NCYC2623_7366_73533-71722:1-1812_g13	XP_024512301.1	32.994	491
Kazachstania servazzii	NCYC3716_11162_9123-10779:52-844_g1	XP_025350345.1	36.607	224
Rhodotorula aurantiaca	NCYC138_24294_26953-24980:1-1968_g2	ESK86852.1	33.473	478
R. aurantiaca	NCYC138_25933_137879-139244:1-1221_g9	ORY58823.1	37.113	194
R. aurantiaca	NCYC138_26187_12489-11164:2-1326_g11	KDE04214.1	35.514	428
R. aurantiaca	NCYC138_26187_15296-13780:1-1515_g12	SCV68130.1	29.843	382
R. aurantiaca	NCYC138_26187_17827-16213:1-1607_g13	SCZ89829.1	34.988	403
R. aurantiaca	NCYC138_26873_15807-17461:1-1269_g18	XP_007403905.1	32.448	339
R. aurantiaca	NCYC138_26873_18173-19731:1-1556_g19	KDE04214.1	37.371	388
R. aurantiaca	NCYC138_27265_87320-85674:1-1364_g24	XP_019027000.1	34.772	394
R. aurantiaca	NCYC138_9077_28926-27356:2-1509_g26	SGY81781.1	37.824	386
R. glutinis	NCYC59_9560_4292-5160:2-869_g10	Q9UVC3.1	34.084	311
R. minuta	NCYC2581_28195_14695-13365:2-1249_g8	XP_007404658.1	37.143	280
R. minuta	NCYC2581_28609_42824-41261:1-1293_g14	KIJ61829.1	37.407	270
R. minuta	NCYC2581_28960_42956-44527:1-1200_g18	XP_014564801.1	38.095	294
R. minuta	NCYC2581_29054_26353-27832:1-1349_g19	KIM77629.1	35.952	420
R. minuta	NCYC2581_29242_7459-9138:1-1649_g21	KKY17076.1	30.201	298
R. minuta	NCYC2581_29293_2311-974:31-1252_g23	TFK84945.1	32.273	220
R. minuta	NCYC930_1733_259395-257781:2-995_g2	XP_007313925.1	28.521	284
R. minuta	NCYC930_1792_350169-348267:341-1643_g5	SCV68130.1	39.739	307
R. minuta	NCYC930_1792_352537-351055:40-690_g6	XP_014564801.1	36.744	215
R. minuta	NCYC930_1881_56804-55082:1-1722_g15	OCH86155.1	39.295	369
R. minuta	NCYC930_1881_87207-85614:307-1594_g16	RDW85359.1	37.183	355
R. minuta	NCYC930_1889_876864-875386:1-1377_g17	TNY24476.1	36.636	434
R. minuta	NCYC930_1890_430907-429073:1-1681_g18	PCH35719.1	32.13	277
R. minuta	NCYC930_1895_155438-153979:110-1253_g19	OLN87586.1	30.128	312
R. minuta	NCYC931_40553_73128-71351:2-1541_g14	TEB37484.1	32.297	418
R. minuta	NCYC931_40635_248322-249684:1-1312_g18	SGY81781.1	35.686	255
R. minuta	NCYC931_40635_30752-32185:1-1353_g19	KZP12182.1	36.436	376

R. minuta	NCYC931_40859_96402-97889:2-1356_g21	XP_024686388.1	35.294	340
R. minuta	NCYC931_41020_56686-54879:2-1722_g24	OOQ90429.1	27.614	373
R. minuta	NCYC931_41026_66044-67271:2-1096_g25	KIJ26121.1	29.536	237
R. minuta	NCYC931_41031_74240-72778:306-1433_g26	XP_001886091.1	35.043	351
R. minuta	NCYC931_41057_401231-399617:2-1615_g30	XP_007313925.1	38.057	494
R. minuta	NCYC931_41221_59592-57836:1-1576_g34	ESK85959.1	36.559	372
R. minuta	NCYC931_41221_90257-88703:1-1555_g35	TFK78780.1	37.825	423
Sporopachydermia lactativora	NCYC2866_26798_8463-9399:1-937_g1	CCE81469.1	37.025	316
Trichosporon domesticum	NCYC2665_2381_386157-387698:1-1542_g10	ORY33012.1	34.899	447
T. dulcitum	NCYC2510_9794_718676-720249:1-1493_g10	XP_018279169.1	36.709	395
T. jirovecii	NCYC3254_17661_5217-3773:381-1280_g10	CCO36833.1	35.165	273
T. jirovecii	NCYC3254_17842_86027-87560:2-1296_g12	TPX14092.1	38.997	359
T. jirovecii	NCYC3254_17984_853532-851956:411-1496_g13	XP_012048170.1	32.66	297
T. montevideense	NCYC2638_4604_1035698-1034150:1-1549_g10	OWZ39582.1	34.318	440
T. ovoides	NCYC472_13235_51563-52596:2-1034_g19	ORY33012.1	39.535	344

Table 5.2: Top hits from a blastp search of the nr database (limited to fungi) for sequences matching the 496 unclassified NCYC CYPs. Only those that could not be identified by this method are shown (n=69).

Accession	**Num. hits**	**Description**
ABO09628.1	2	NADPH cytochrome P450 reductase [Starmerella bombicola]
AQZ15421.1	1	PGK1 (YCR012W) [Zygosaccharomyces parabailii]
AQZ18622.1	1	PGK1 (YCR012W) [Zygosaccharomyces parabailii]
BAL05196.1	1	cytochrome P450 [Phanerochaete chrysosporium]
CBQ73997.1	2	conserved hypothetical protein [Sporisorium reilianum SRZ2]
CDF90157.1	22	ZYBA0S06-01970g1_1 [Zygosaccharomyces bailii CLIB 213]
CDI51145.1	4	related to Cytochrome P450 4F8 [Melanopsichium pennsylvanicum 4]
CDI54874.1	3	related to Cytochrome P450 [Melanopsichium pennsylvanicum 4]
CDI55655.1	1	related to Cytochrome P450 [Melanopsichium pennsylvanicum 4]
CDO51727.1	1	similar to S. cerevisiae YBL022C PIM1 ATP-dependent Lon protease [Geotrichum candidum]
CDO53625.1	8	Conserved hypothetical protein. Putative monooxygenase [Geotrichum candidum]
CDO58003.1	8	conserved hypothetical protein [Geotrichum candidum]

CDR39009.1	1	CYFA0S02e10572g1_1 [Cyberlindnera fabianii]
CDR43218.1	15	CYFA0S11e01750g1_1 [Cyberlindnera fabianii]
CDU24917.1	2	related to Cytochrome P450 [Sporisorium scitamineum]
CEP23841.1	2	TRI4 [Cyberlindnera jadinii]
CUA77073.1	1	Alkane hydroxylase MAH1 [Rhizoctonia solani]
EJU01233.1	1	cytochrome P450 [Dacryopinax primogenitus]
EKC98846.1	3	cyclin-dependent protein kinase [Trichosporon asahii var. asahii CBS 8904]
EKD02434.1	2	cytochrome P450 [Trichosporon asahii var. asahii CBS 8904]
EKD05546.1	3	hypothetical protein A1Q2_00160 [Trichosporon asahii var. asahii CBS 8904]
EME43868.1	1	hypothetical protein DOTSEDRAFT_171756 [Dothistroma septosporum NZE10]
ETS60641.1	2	hypothetical protein PaG_05285 [Moesziomyces aphidis DSM 70725]
ETS61408.1	1	hypothetical protein PaG_04438 [Moesziomyces aphidis DSM 70725]
ETS61464.1	1	hypothetical protein PaG_04501 [Moesziomyces aphidis DSM 70725]
ETS61768.1	1	hypothetical protein PaG_03864 [Moesziomyces aphidis DSM 70725]
ETS62300.1	3	hypothetical protein PaG_03377 [Moesziomyces aphidis DSM 70725]
ETS62942.1	1	hypothetical protein PaG_02711 [Moesziomyces aphidis DSM 70725]
ETS63459.1	1	hypothetical protein PaG_01744 [Moesziomyces aphidis DSM 70725]
ETS64552.1	1	hypothetical protein PaG_01017 [Moesziomyces aphidis DSM 70725]
GAC72072.1	1	cytochrome P450 CYP3/CYP5/CYP6/CYP9 subfamilies [Moesziomyces antarcticus T-34]
GAC73194.1	1	cytochrome P450 CYP3/CYP5/CYP6/CYP9 subfamilies [Moesziomyces antarcticus T-34]
GAC76832.1	3	cytochrome P450 CYP4/CYP19/CYP26 subfamilies [Moesziomyces antarcticus T-34]
GAV27818.1	5	hypothetical protein PMKS-001286 [Pichia membranifaciens]
KGK40082.1	11	hypothetical protein JL09_g752 [Pichia kudriavzevii]
KGK40122.1	14	hypothetical protein JL09_g753 [Pichia kudriavzevii]
KII83030.1	2	hypothetical protein PLICRDRAFT_494881, partial [Plicaturopsis crispa FD-325 SS-3]
KWU42266.1	7	cytochrome P450 oxidoreductase [Rhodotorula sp. JG-1b]

KZT50748.1	1	cytochrome P450 [Calocera cornea HHB12733]
OJJ60515.1	1	hypothetical protein ASPSYDRAFT_197967 [Aspergillus sydowii CBS 593.65]
ONH66674.1	5	Isotrichodermin C-15 hydroxylase [Cyberlindnera fabianii]
ONH77793.1	28	Outward-rectifier potassium channel TOK1 [Pichia kudriavzevii]
ORY21756.1	4	cytochrome P450 [Naematelia encephala]
ORY55968.1	2	cytochrome P450 oxidoreductase [Leucosporidium creatinivorum]
ORY81226.1	1	cytochrome P450 [Leucosporidium creatinivorum]
ORY88428.1	1	cytochrome P450 [Leucosporidium creatinivorum]
OUT20005.1	3	hypothetical protein CAS74_004738 [Pichia kudriavzevii]
OUT20805.1	4	hypothetical protein CAS74_004475 [Pichia kudriavzevii]
PRQ72580.1	2	cytochrome P450 oxidoreductase [Rhodotorula toruloides]
PVH77741.1	2	putative benzoate 4-monooxygenase cytochrome P450 [Cadophora sp. DSE1049]
PWY97649.1	1	cytochrome P450 [Testicularia cyperi]
RDL40765.1	1	Uncharacterized protein BP5553_00744 [Phialophora cf. hyalina BP 5553]
RDW75511.1	1	hypothetical protein BP6252_06653 [Coleophoma cylindrospora]
RPD81420.1	1	cytochrome P450 [Lentinus tigrinus ALCF2SS1-7]
RSH93263.1	3	hypothetical protein EHS25_007617 [Saitozyma podzolica]
RXK36384.1	1	hypothetical protein M231_06350 [Tremella mesenterica]
RYO79165.1	2	hypothetical protein DL763_009389 [Monosporascus cannonballus]
RYP59664.1	1	hypothetical protein DL771_010807 [Monosporascus sp. 5C6A]
SCZ89829.1	1	BZ3500_MvSof-1268-A1-R1_Chr1-3g01603 [Microbotryum saponariae]
SJX63839.1	3	uncharacterized protein SRS1_11138 [Sporisorium reilianum f. sp. reilianum]
SMN19201.1	3	similar to S. cerevisiae YMR152W YIM1 PUF [Kazachstania saulgeensis]
TID28319.1	2	hypothetical protein CANINC_002496 [[Candida] inconspicua]
TID30093.1	17	hypothetical protein CANINC_001317 [[Candida] inconspicua]
TKA57515.1	2	hypothetical protein B0A53_00746 [Rhodotorula sp. CCFEE 5036]

TKA58519.1	1	hypothetical protein B0A53_00260 [Rhodotorula sp. CCFEE 5036]
TKY86221.1	3	hypothetical protein EX895_005046 [Sporisorium graminicola]
TKY90424.1	1	hypothetical protein EX895_000422 [Sporisorium graminicola]
XP_007875702.1	1	cytochrome p450 monooxygenase [Anthracocystis flocculosa PF-1]
XP_011391519.1	3	hypothetical protein UMAG_05791 [Ustilago maydis 521]
XP_012186313.1	1	cytochrome P450 [Pseudozyma hubeiensis SY62]
XP_012186345.1	6	isotrichodermin C-15 hydroxylase [Pseudozyma hubeiensis SY62]
XP_012187276.1	1	hypothetical protein PHSY_001254 [Pseudozyma hubeiensis SY62]
XP_012187488.1	1	benzoate 4-monooxygenase cytochrome P450 [Pseudozyma hubeiensis SY62]
XP_012188084.1	3	pisatin demethylase [Pseudozyma hubeiensis SY62]
XP_012188252.1	6	hypothetical protein PHSY_002238 [Pseudozyma hubeiensis SY62]
XP_012188344.1	1	cytochrome P450 monooxygenase [Pseudozyma hubeiensis SY62]
XP_012189265.1	1	RNA-directed RNA polymerase [Pseudozyma hubeiensis SY62]
XP_012190368.1	1	hypothetical protein PHSY_004365 [Pseudozyma hubeiensis SY62]
XP_012191738.1	1	hypothetical protein PHSY_005740 [Pseudozyma hubeiensis SY62]
XP_014176214.1	2	Cytochrome P450 [Trichosporon asahii var. asahii CBS 2479]
XP_014180864.1	1	cytochrome P450 [Trichosporon asahii var. asahii CBS 2479]
XP_014182292.1	1	Cytochrome P450 [Trichosporon asahii var. asahii CBS 2479]
XP_014182466.1	4	cytochrome P450 [Trichosporon asahii var. asahii CBS 2479]
XP_014183143.1	2	Cytochrome P450 [Trichosporon asahii var. asahii CBS 2479]
XP_014564801.1	1	hypothetical protein L969DRAFT_97442 [Mixia osmundae IAM 14324]
XP_014568861.1	1	hypothetical protein L969DRAFT_86923 [Mixia osmundae IAM 14324]
XP_014659110.1	3	cytochrome P450 monooxygenase [Moesziomyces antarcticus]

XP_016270781.1	2	cytochrome P450, family 4, subfamily A [Rhodotorula toruloides NP11]
XP_016292673.1	1	cytochrome P450 [Kalmanozyma brasiliensis GHG001]
XP_016292796.1	2	cytochrome P450 [Kalmanozyma brasiliensis GHG001]
XP_016293162.1	2	hypothetical protein PSEUBRA_SCAF18g04717 [Kalmanozyma brasiliensis GHG001]
XP_016294019.1	2	hypothetical protein PSEUBRA_SCAF13g01973 [Kalmanozyma brasiliensis GHG001]
XP_016294816.1	1	hypothetical protein PSEUBRA_SCAF1g00282 [Kalmanozyma brasiliensis GHG001]
XP_018221826.1	11	NCP1-like protein [Saccharomyces eubayanus]
XP_018265046.1	2	hypothetical protein I303_03228 [Kwoniella dejecticola CBS 10117]
XP_018276022.1	2	cytochrome P450 [Cutaneotrichosporon oleaginosum]
XP_018276981.1	2	cytochrome P450 [Cutaneotrichosporon oleaginosum]
XP_018277590.1	1	cytochrome P450 [Cutaneotrichosporon oleaginosum]
XP_018278631.1	2	cytochrome P450 [Cutaneotrichosporon oleaginosum]
XP_018279169.1	2	cytochrome P450 [Cutaneotrichosporon oleaginosum]
XP_018280211.1	7	cytochrome P450 [Cutaneotrichosporon oleaginosum]
XP_018281194.1	1	cytochrome P450 [Cutaneotrichosporon oleaginosum]
XP_018282904.1	8	cytochrome P450 [Cutaneotrichosporon oleaginosum]
XP_018715005.1	1	cytochrome P450 [Metschnikowia bicuspidata var. bicuspidata NRRL YB-4993]
XP_019016213.1	8	hypothetical protein PICMEDRAFT_17595 [Pichia membranifaciens NRRL Y-2026]
XP_019016827.1	25	hypothetical protein PICMEDRAFT_73229 [Pichia membranifaciens NRRL Y-2026]
XP_019036661.1	17	hypothetical protein WICANDRAFT_64802 [Wickerhamomyces anomalus NRRL Y-366-8]
XP_019041970.1	6	hypothetical protein WICANDRAFT_99132 [Wickerhamomyces anomalus NRRL Y-366-8]
XP_019043506.1	1	hypothetical protein I302_08084 [Kwoniella bestiolae CBS 10118]
XP_019049805.1	1	hypothetical protein I302_00224 [Kwoniella bestiolae CBS 10118]
XP_020067599.1	3	RabGAP/TBC [Suhomyces tanzawaensis NRRL Y-17324]
XP_020068903.1	3	cytochrome P450 [Cyberlindnera jadinii NRRL Y-1542]
XP_020070685.1	1	beta subunit of fatty acid synthase [Cyberlindnera jadinii NRRL Y-1542]
XP_028472465.1	3	hypothetical protein EHS24_003629 [Apiotrichum porosum]
XP_028472874.1	1	hypothetical protein EHS24_002785 [Apiotrichum porosum]

XP_028475057.1	1	hypothetical protein EHS24_009618 [Apiotrichum porosum]
XP_028475766.1	1	hypothetical protein EHS24_008481 [Apiotrichum porosum]
XP_028477865.1	1	hypothetical protein EHS24_005937 [Apiotrichum porosum]
XP_029323478.1	27	uncharacterized protein C5L36_0E00700 [Pichia kudriavzevii]

Table 5.3: The 119 unique sequences forming the top blastp hits for the 427 putative CYPs with hits above 40% ID. PUF = Protein of Unknown Function.

Taxonomic clustering of novel P450s

Given the larger and more complex genomes seen in the Basidiomycetes compared to the Ascomycetes, it was hypothesised that relatively more CYPs would be found in species representing the former phylum. The CYP sequences discovered in the NCYC genome sequences were categorised and mapped to the rDNA phylogeny built in Chapter 2 (Figure 2.5), see Figure 5.4 for the annotated phylogenetic tree showing the taxonomic distribution of classified CYPs across the NCYC genome sequence data. The vast majority of *Saccharomyces* strains appear to contain only the three CYPs that have been previously assigned to them. Meanwhile there are marked peaks in CYP gene number in the *Metschnikowia*, *Rhodotorula*, and the Ustilaginales. In the latter two clades there are clear peaks in the numbers of unclassified CYPs (green), suggesting that these groups are somewhat understudied with regards to CYP content. One thing to note about this figure is that it is confounded by the same issues as discussed in Chapter 2 (Phylogenetics), so some strains appear to contain more, or sometimes fewer, CYPs than might be expected. Some of this is due to the fragmentation issues resulting in single CYP genes being split into two. On a smaller scale it would be possible to manually curate these cases but here this is infeasible.

Given that CYP genes are often associated with metabolic gene clusters, this finding suggests that there might be some merit in the idea of using CYPs as anchor sequences in the search for such gene clusters. In particular, the taxonomic clustering of potentially novel CYP sequences in the Basidiomycetes suggests that there is potential for novel metabolic pathways to be discovered in these strains/ species.

One problem with the methods described in this chapter is that one is searching for and then retrieving nucleotide sequence matches, then using the protein sequence to classify the CYPs. This requires the use of software such as AUGUSTUS to predict the protein sequence

of each of the nucleotide sequence matches, since a straight translation to amino acid is impossible due to the fact that the sequence matches are not guaranteed to be in frame (they may be partial alignments/matches). As a result, some sequences identified at the HMMER search stage are subsequently lost at the protein prediction step in the event that AUGUSTUS cannot predict a protein from the sequence. Without the protein sequence, these matches cannot be classified according to David Nelson's blastp method. Use of the nucleotide sequence to align to known CYP sequences would not be a feasible alternative given the extreme sequence diversity seen at the nucleotide level in CYPs.

5.4 Conclusions

Searching for and predicting cytochrome P450 genes gives some insight into the metabolic potential of the NCYC strains. The CYP sequences are easily identifiable due to their conserved motifs and may therefore act as flags to focus the search for useful gene clusters. In this case, several thousand CYPs have been identified and located, some of which (496) appear to be part of novel protein families, and therefore may be part of unknown metabolic pathways. This information may be useful in targeting the search for gene clusters, as shown in Figure 5.5.

Phylogenetic analysis has shown that the majority of the unclassified CYPs in this dataset belong to species in the Basidiomycetes. These species are known to have larger, more complex genomes and produce a greater complement of secondary metabolites. The next chapter will deal with the search for novel gene clusters, using CYPs and other genes as flags to focus the discovery process.

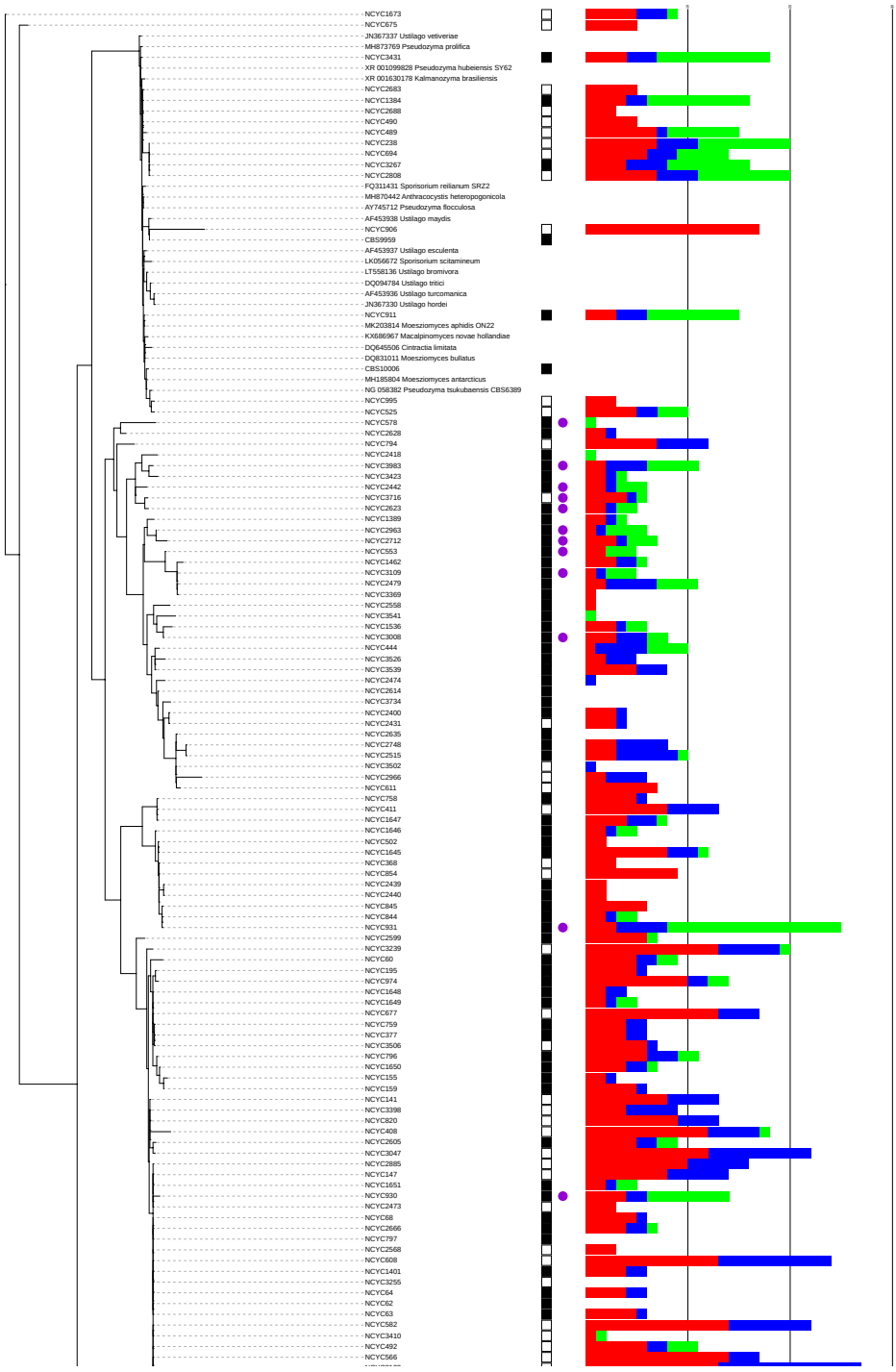

Figure 5.4: Continued on next page.

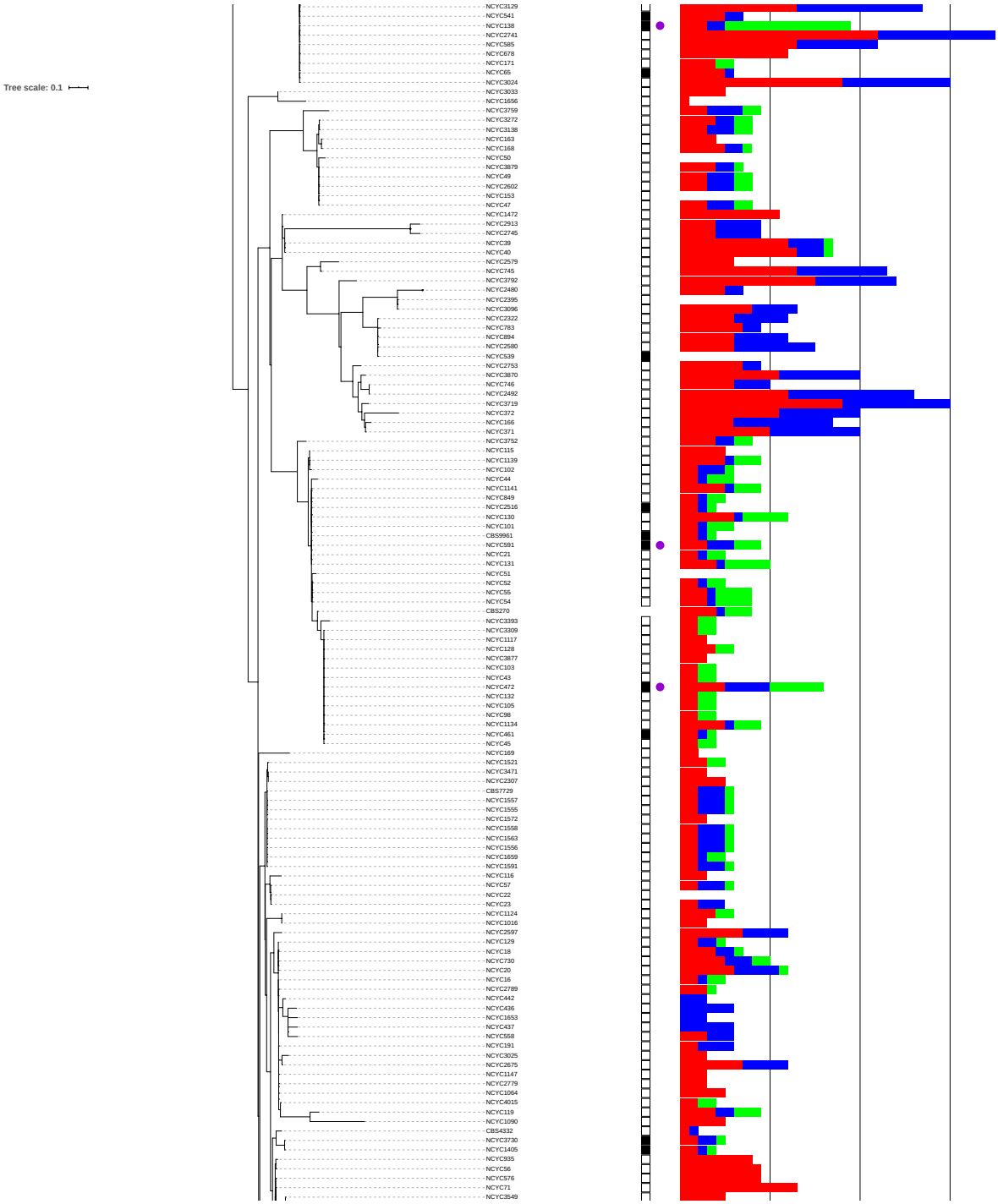

Figure 5.4: Continued on next page.

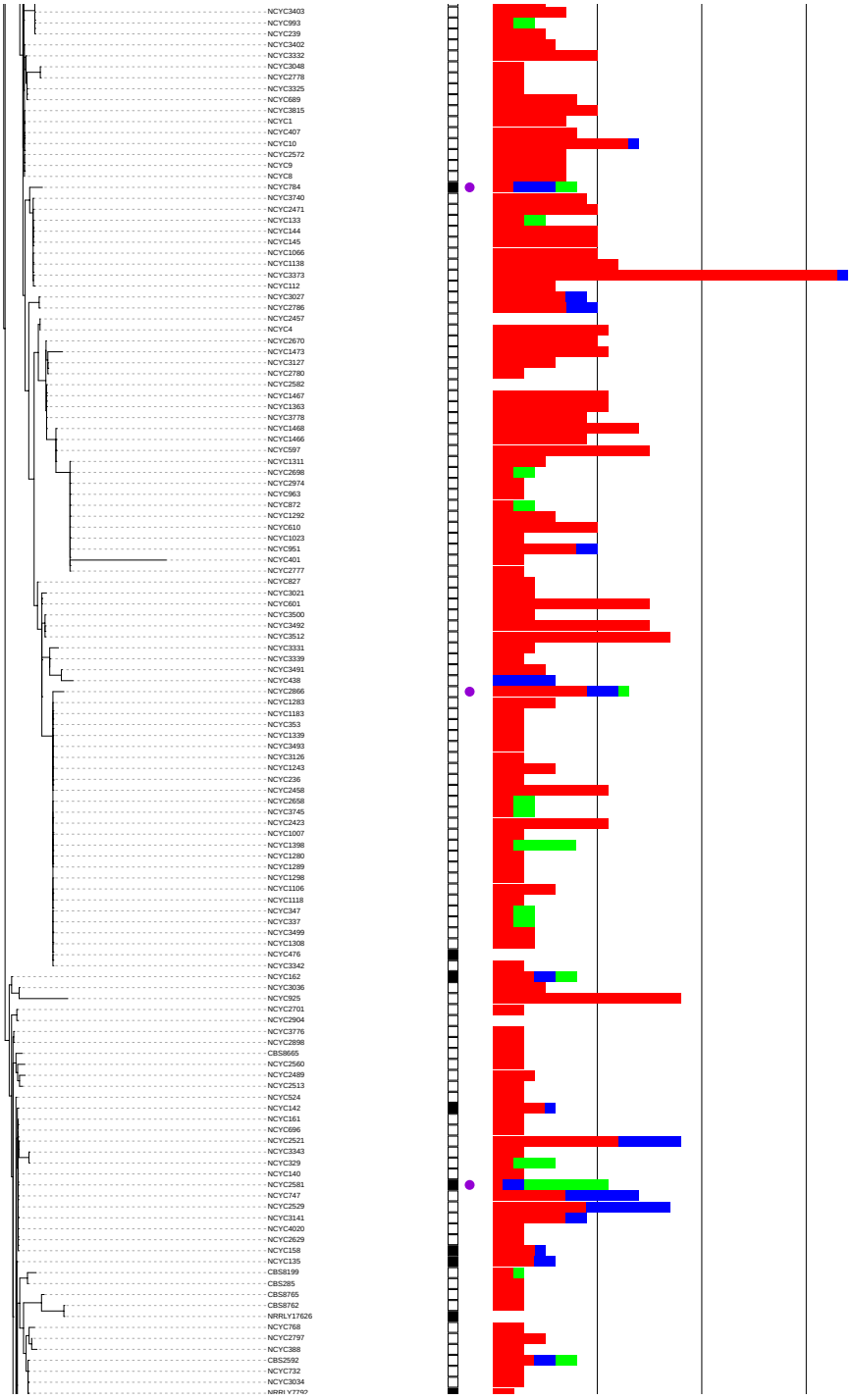

Figure 5.4: Continued on next page.

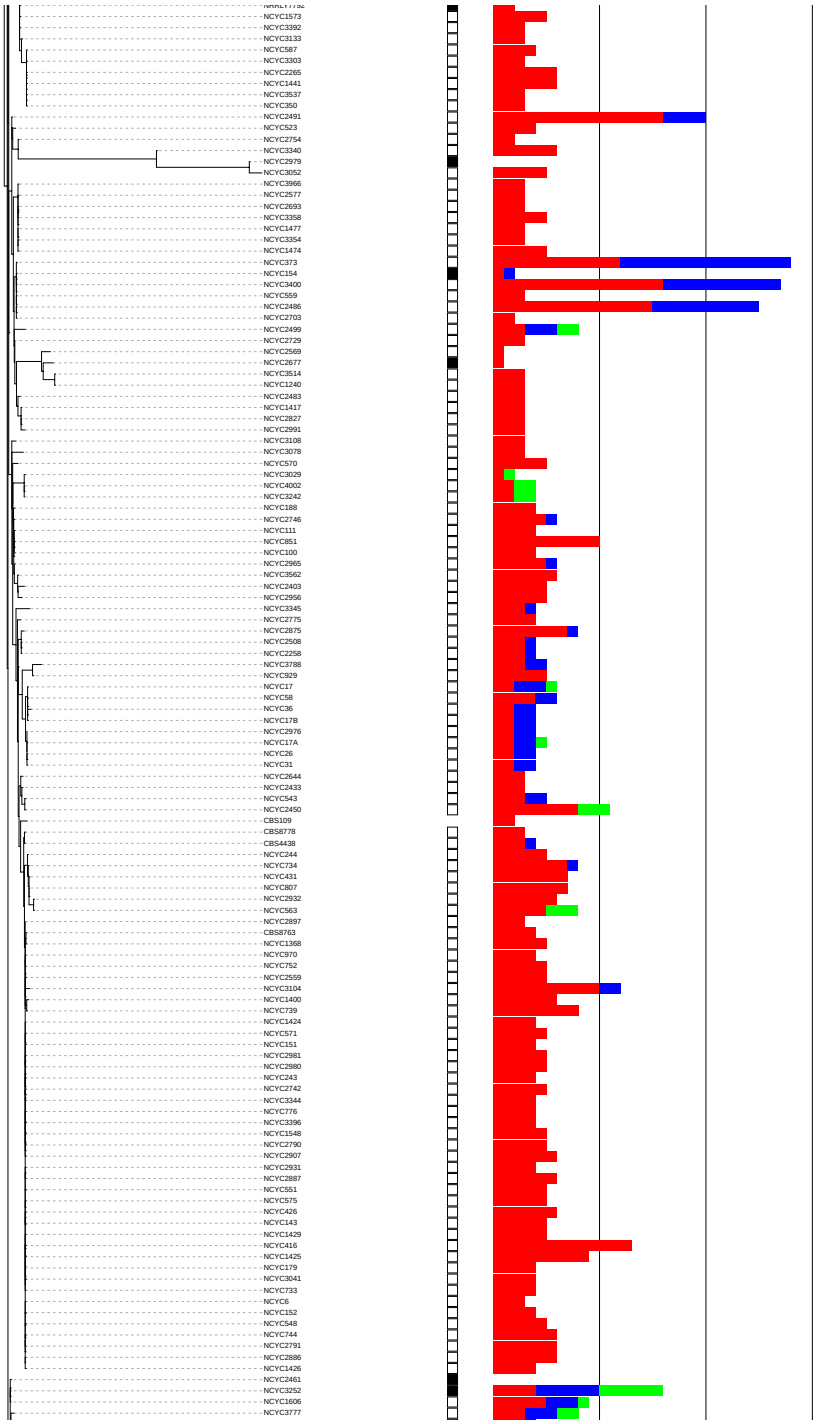

Figure 5.4: Continued on next page.

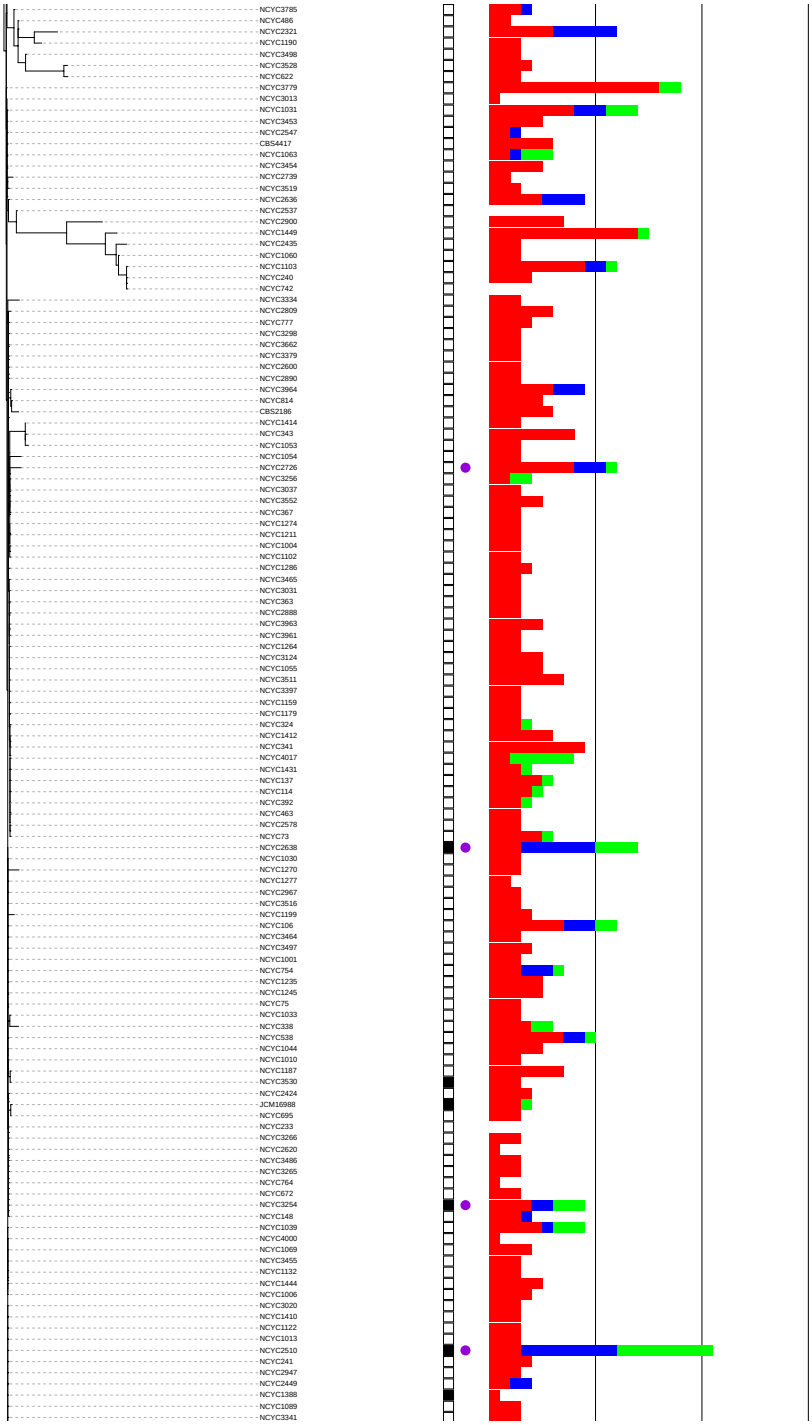

Figure 5.4: Continued on next page.

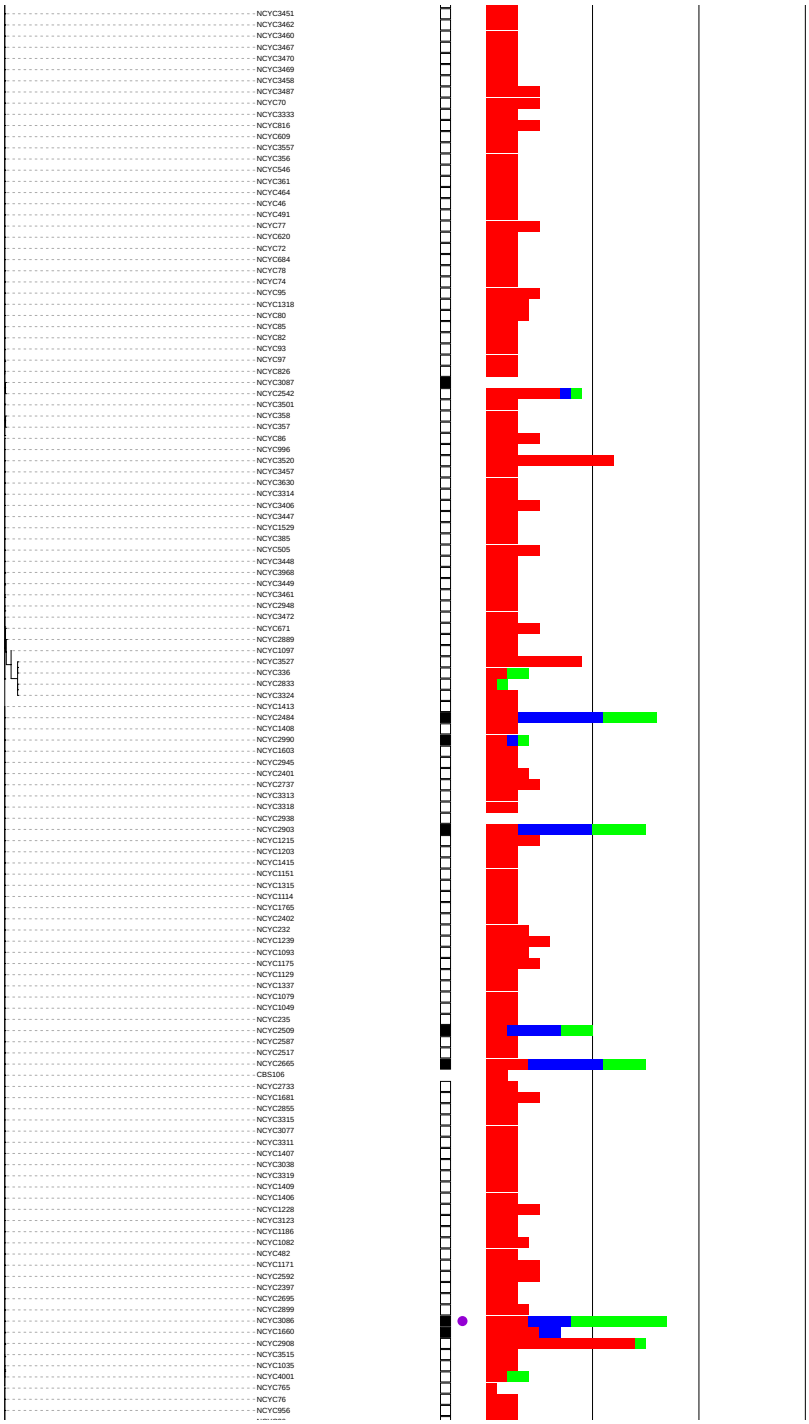

Figure 5.4: Continued on next page.

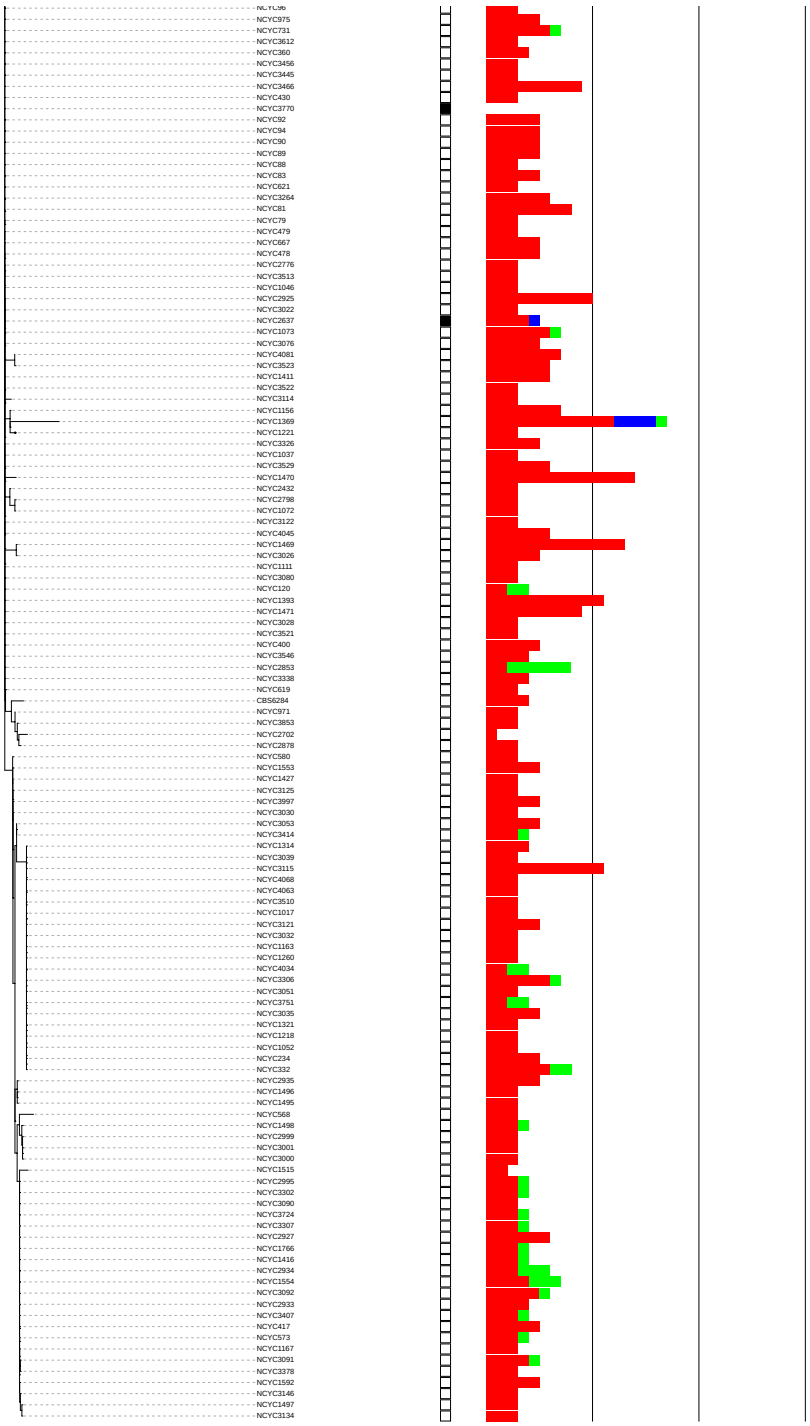

Figure 5.4: Phylogenetic tree constructed from the D1/D2 subunit of the ribosomal DNA of all sequenced NCYC strains and a number of publicly available Ustilaginomycete genomes (annotated version of Figure 2.5). The stacked bar chart shows the numbers of CYPs identified in each strain. Red = CYPs classified to subfamily level, blue = classified only to family level, green = cannot be matched to any known fungal CYP from the Nelson database. Vertical scale lines indicate 10, 20, 30 CYPs. Strains from Table 5.2, indicating CYPs with no identified matches, are marked with purple circles. Basidiomycetes are marked in bold black.

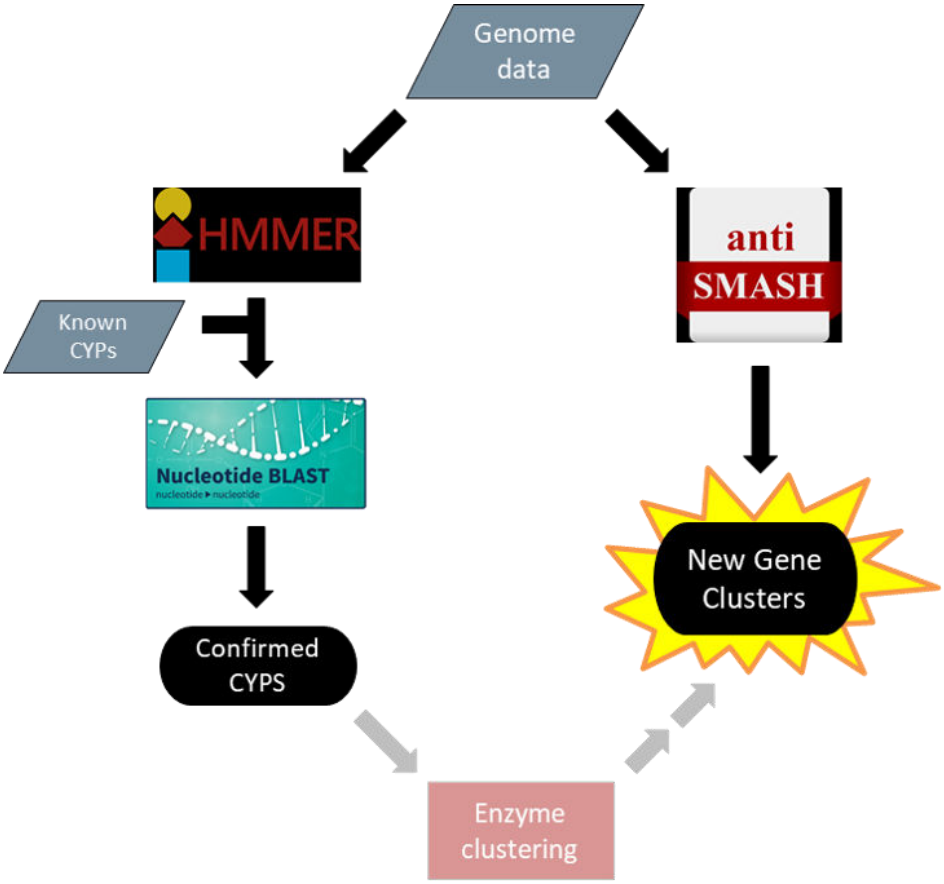

Figure 5.5: Diagrammatic representation of how the CYP discovery work fits into the metabolic gene cluster discovery pipeline discussed in Chapter 6. The process described by left hand side of the diagram could be applied to any gene type deemed to be characteristic of a gene cluster. The enzyme clustering method can form the first step in the process of identifying novel gene clusters based on characteristic genes.

6 Computational methods of finding novel gene clusters in yeast genomes

6.1 Introduction

Discovery of gene cluster variants and entirely novel gene clusters is an active area of research in the natural products community, with emphasis on bioinformatics approaches that circumvent issues of non-culturable organisms and silent biosynthetic pathways. Leading software applications such as antiSMASH (Medema et al. 2011; Weber et al. 2015; Blin et al. 2017) offer a relatively easy way of assessing the metabolic gene cluster content of a genome, in this case of a large number of genomes. However, as discussed in the introduction of this thesis, these methods are limited in scope by the fact that they rely on pattern recognition based on existing gene clusters and common components. A more accurate method of predicting novel gene clusters has yet to be conceived.

Yeasts are relatively understudied in terms of their biosynthetic potential. There are only a handful of previously identified and characterised metabolic gene clusters in yeast species, most of which are described in the introduction of this thesis. Yeasts are incredibly diverse and are estimated to comprise approximately 1% of fungal species (Kurtzman et al. 2006), themselves estimated to number at least 1.5 million species (Hawksworth et al. 2017). Very few of these have been fully described and investigated. One of the reasons for this diversity is that, as alluded to earlier, yeasts are not a monophyletic lineage, with true yeasts (e.g. *Saccharomyes*) in the Ascomycota phylum and a variety of yeast-like fungi in the Basidiomycota. These latter species share the simplified, single-cell lifestyle of the true yeasts but are spread throughout all kinds of divergent fungal groups, each with their own unique evolutionary history and therefore their own potential for metabolite production. As such, yeasts represent a potential goldmine of metabolic variety. The NCYC genome sequence dataset covers a wide variety of species (the collection as a whole contains ~4000 strains from approximately 530 species, of which the sequenced strains have been selected to be an approximately representative sample, with enrichment in certain lineages of particular

bioindustrial interest, such as the *Rhodotorula* strains) from multiple phyla. With all this in mind, it seems prudent to use the current suite of BGC discovery tools to mine the NCYC dataset for metabolic gene clusters.

There was also, to my knowledge, no tool that would search for a known gene cluster across a dataset of many (unannotated) genomes. Tools such as MultiGeneBlast (Medema et al. 2013) search annotated (e.g. GenBank format) genome sequences for a given gene cluster. For this reason, the FindClusters pipeline described below was developed to provide a simple way of checking any number of assembled genomes (in FASTA format) for the presence of full or partial copies of a given metabolic gene cluster. It could also be extended to process genomes from the raw read data in the future, if desired, with the addition of a genome assembly step.

This chapter details work to assess the gene cluster content of the NCYC collection, with a particular focus on the pigmented *Rhodotorula* genus, which is closely related to the glycolipid producing ustilaginales investigated in previous chapters. The aim is to answer the question of what number and type of metabolic gene clusters reside in the genomes sequenced from the NCYC collection. I also ask whether there is a trend towards greater metabolic diversity in certain lineages, with the expectation that this will be true. The dataset contains representatives of both simplified true yeasts (*Saccharomyces*, etc.) and yeast-like Basidiomycete species related to well known secondary metabolite producing filamentous fungi (e.g. *Aspergillus*). The chapter follows three threads, firstly the development of the FindClusters pipeline for describing the prevalence of known gene clusters in novel genome datasets, second is a case study to investigate the unknown gene cluster potential of the *Rhodotorula*, and thirdly the preliminary work in developing an alternative method for identifying novel gene clusters in assembled genome sequences.

6.2 Gene Cluster Finding Pipeline - FindClusters

One of the themes of this thesis is the search for known gene clusters in the NCYC collection. As part of this theme, I developed a computational pipeline that takes a set of assembled genomes and searches them for the presence of genes from a known cluster, provided by the user. The pipeline reports, via a PNG image and a tab-delimited file, the assemblies in which the genes occur as a whole or partial cluster. The pipeline is written in Python, and controlled by a shell script (bash). It relies on HMMER (Eddy 2015) to create profile HMMs for each cluster gene, and then search the assemblies for matches to those HMMs. It then utilises the Biopython (Cock et al. 2009) and Pyfaidx (Shirley et al. 2015) libraries to retrieve the

matching sequences and then sort them into clusters when present as such. The PIL library (see Pillow fork from `https://pillow.readthedocs.io/en/stable/index.html`) is used to draw a colour coded figure that illustrates the configurations of any cluster found. A diagram of the pipeline can be seen in Figure 6.1, with more detailed description of each script below. All code can be found at `https://github.com/chrispyatt/FindClusters`. There is a README file as well as help options (-h) for each script detailing the various options.

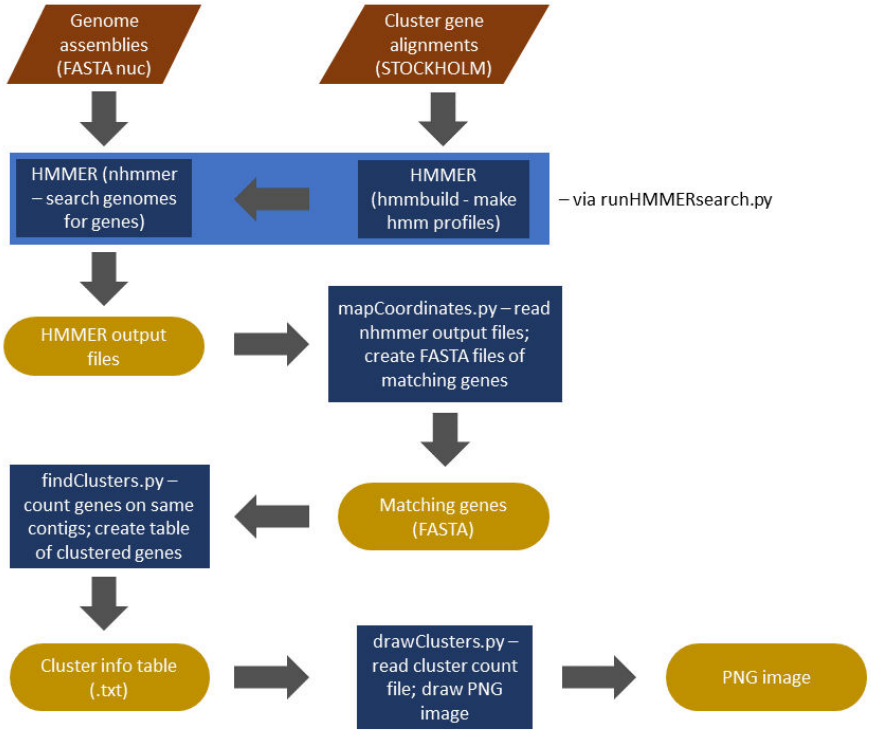

Figure 6.1: Schematic representation of the cluster finding and drawing pipeline.

The first script to be run is runHMMERsearch.py, which takes a list of FASTA formatted target gene files, and STOCKHOLM formatted genome assemblies (to be searched for the target genes). These are separated by file extension and the former used to create profile HMMs (using hmmbuild) with which to search the latter (using nhmmer). The script also takes an output directory option (default being the current directory) in which to deposit the output files.

The second script to run is mapCoordinates.py, which takes as input a single nhmmer output file and a corresponding genome assembly from which to retrieve any good hits (matching sequences). "Good" in this case is defined as being above the default inclusion threshold but this can be changed if desired. Other optional arguments that can be invoked may limit the retrieved hits to a specified number, or retrieve some degree of surrounding sequence

in addition to the hit. The script uses Biopython SeqIO to index and retrieve sequence from the genome assembly. The output is a directory populated by the retrieved sequences in individual FASTA files, with both the filename and FASTA header containing location information.

Next, the findClusters.py script uses the location metadata from the FASTA files (created by mapCoordinates.py) to determine whether any of the desired cluster genes are clustered in the input assembly. The output is a text file describing any gene clusters found in terms of number of genes found, and clustering thereof.

The final script in the pipeline is drawClusters.py, which takes the output file from findClusters.py and creates a PNG image to visualise the gene clusters found in the input genome assemblies. It also has an optional argument that can be used to omit rows that the user considers irrelevant or spurious. If genes are on the same contig but more than 10,000 bp apart, the script will draw a standard red bar with a label showing the actual separation. This prevents images from being very wide, and indicates that the genes may not actually be clustered (in the case of very large contigs). The figure of 10,000 bp was chosen as most documented gene clusters are more tightly grouped than this (see previous chapters for examples).

6.3 Focus on Rhodotorula gene clusters - FindClusters & antiSMASH

The Rhodotorula are a genus of Basidiomycetous yeasts named for their red pigmentation. This pigmentation is due to production of carotenoids, some from a metabolic gene cluster called the CAR gene cluster (mentioned previously in the introduction of this thesis). The gene cluster arrangement is shown in Figure 6.2. The NCYC collection contains roughly 80 examples of *Rhodotorula* strains, the majority of which (76) have had their genomes sequenced. The genus is therefore a particularly well sampled one in which to look for gene cluster diversity. The majority of the 76 strains are attributed to the species *R. mucilaginosa*, *R. glutinis*, and *R. minuta* (36, 10, & 8 strains respectively), but there are also representatives of *R. aurantiaca* (1), *R. creatinovora* (1), *R. cresolica* (1), *R. dairenensis* (2), *R. graminis* (5), *R. laryngis* (5), *R. phylloplana* (1), *R. sloofiae* (1), and *R. vanillica* (1), plus four novel strains (NCYC3056, NCYC3832, NCYC3833, NCYC3835). The three species most abundant in the collection have been known to cause opportunistic infection in humans, particularly of medical inserts such as venal catheters (Zaas et al. 2003). They are also of potential use in bioremediation, with an apparent ability to survive in polluted sediments and neutralise

certain petroleum contaminants (Hesham et al. 2012; MacGillivray et al. 1993).

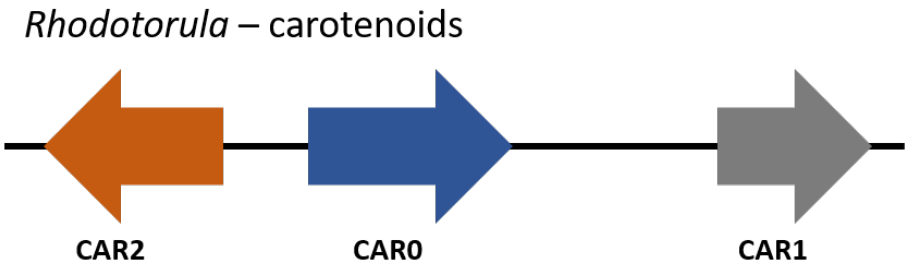

Figure 6.2: CAR gene cluster producing pigmentation carotenoids such as β-carotene. This is the arrangement seen in *R. graminis*, as reported by Landolfo et al. (2018).

Methods

In order to see what gene clusters might be present in this genus, analyses were run using the FindClusters pipeline (to check for variants of the CAR gene cluster; default settings with no strains omitted from final output) and established tools such as antiSMASH (Medema et al. 2011; Weber et al. 2015; Blin et al. 2017) and BiG-SCAPE (Navarro-Muñoz et al. 2018). The antiSMASH analysis included the 'clusterfinder' option to predict potential gene clusters of unknown function/type. BiG-SCAPE (default options) was used primarily to categorise and visualise the output from the antiSMASH analysis (i.e. assess gene cluster families). BUSCO v3 (Waterhouse et al. 2018; Simao et al. 2015) was used to assess genome assembly completeness.

Results

The results from the FindClusters analysis are shown in Figure 6.3. See Table D.1 (Appendix) for more strain information, including genome quality figures. Twenty six (26) full CAR gene clusters were found, along with seven (7) partial gene clusters and twenty (20) fragmented gene clusters (due to short contigs). Twelve (12) assemblies had missing gene clusters (possibly due to fragmentation/ poor assembly). The other eleven *Rhodotorula* genomes in the dataset had not yet been assembled at the time of writing and therefore were not included. There is some variation evident, even in such a small gene cluster. For example the only strains with gene clusters exactly matching that shown in Figure 6.2 are NCYC60 and NCYC3722 (*R. glutinis* and *R. graminis*). Other *R. glutinis* strains seem to follow the *R. mucilaginosa* pattern, as does the other *R. graminis* strain (NCYC1401 only, NCYC502 is fragmented, and the other two were not present in the final analysis). In all examples of this gene cluster, CAR0 and CAR1 are always facing opposite directions. The other variant is

NCYC2605 (*R. vanillica*), where either CAR1 has been moved to the opposite end of the gene cluster, or the other two genes have been switched around. Only one example of this species is present in the dataset so this cannot currently be corroborated by other data points. In general, the majority of the strains in this dataset follow the pattern seen in *R. mucilaginosa* (Landolfo et al. 2018), though it is worth bearing in mind that a significant part of the dataset is made up of that species. According to the *Rhodotorula* phylogeny of Biswas et al. (2001), *R. graminis* is nested within the genus as opposed to being a sister species to the others. This suggests that the gene cluster arrangement seen in this species has arisen (presumably through inversion of the CAR1 gene) since the split from its sister species.

The *Rhodotorula* genomes were also analysed using antiSMASH and BiG-SCAPE (Medema et al. 2011; Navarro-Muñoz et al. 2018). In total, 700 putative gene clusters were predicted, grouped into 341 families, see Table 6.1. This included 127 Terpene based, and 54 non-ribosomally synthesised peptide based biosynthetic pathways, which appear to make up a substantial part of the *Rhodotorula* metabolic complement. Substantial numbers of non-standard gene clusters were also predicted (509), although the 'clusterfinder' algorithm used has a fairly high false positive rate due to the greedy strategy taken when classifying potential gene clusters. 193 of the non-standard gene clusters are singletons (i.e. single member gene cluster family) and therefore may be the less reliable predictions. It may also be that some of the smaller gene cluster families are made up of similar predictions in several strains of the same species. See Table D.1 for summary statistics for the strains discussed here, including number of gene clusters found by antiSMASH. Output files are available at `https://github.com/chrispyatt/PhData`. It remains, however, that there is remarkable potential for novel gene cluster discovery and characterisation in this genus.

A cursory examination of the entire NCYC genome sequence dataset (see Table 6.2) suggests that the *Rhodotorula* possess a slightly larger proportion of the collection's metabolic gene cluster diversity than might be expected if gene clusters were distributed equally (the *Rhodotorula* dataset comprises ~6% of the total number of genomes sequenced, while containing ~7.5% of the predicted gene clusters. This is not unexpected given that Basidiomycetes are known to produce more secondary metabolites than Ascomycetes (*Saccharomyces* strains making up a large part of the dataset and collection). In a similar picture to that seen in the *Rhodotorula*, the collection as a whole contains a large number of putative Terpene based gene clusters, along with a modest number of NRPS and Saccharide based pathways. In the case of the NRPS gene clusters, the *Rhodotorula* seem to contain more gene clusters of this type than is seen in the collection as a whole (7.7% versus 1.2%). Terpenes are also over-

Gene cluster type	Number found	Number of families
NRPS	54	26 (22 singletons)
RiPPs	5	4
PKS-1	0	0
PKS-other	0	0
PKS/NRPS hybrid	0	0
Terpene	127	30 (15 singletons)
Saccharides	5	4
Others	509	277 (193 singletons)

Table 6.1: antiSMASH & BiG-SCAPE results (analysis of all sequenced *Rhodotorula* genomes. NRPS = non-ribosomal peptide synthase, RiPPs = ribosomally synthesised and post-translationally modified peptides, PKS = poly-ketide synthase. Many of the gene clusters in the 'Others' category are 'cf_putative', meaning that they are putative gene clusters of unknown type predicted by the clusterfinder algorithm (high FP rate). See Figures D.1, D.2, and D.3 for network diagrams showing the relationships between the gene cluster families identified.

represented, though to a lesser degree (18% to 13%). This may be related to pigmentation, carotenoids being terpene-based compounds. Again there are a lot of non-standard gene clusters predicted by the 'clusterfinder' algorithm which would require further investigation.

6.4 Using members of common gene super-families as flags for local gene cluster searching

The final thread of this chapter concerns preliminary work around the development of an alternative method of identifying gene clusters in *de novo* sequenced genomes. The approach taken was to combine knowledge of genes commonly found in gene clusters (for example Cytochrome P450 oxidases or Major Facilitator transporters - CYPs are present in approximately 9% of the putative gene clusters identified in the *Rhodotorula*, above, and MFS genes are also well represented) with an enzyme prediction step using DETECT v2 (Hung et al. 2010). The approach is to determine where physical clusters of secondary metabolism enzymes (as categorised by DETECT) occur in proximity to the above mentioned 'flag' genes. Figure 6.4 shows an example (in this case with just CYPs acting as 'flag' genes) of how this approach may be used to complement other gene cluster discovery tools in a wider pipeline.

Gene cluster type	Number found
NRPS	112
RiPPs	10
PKS-1	13
PKS-other	1
PKS/NRPS hybrid	0
Terpene	1277
Saccharides	71
Others	7819

Table 6.2: antiSMASH summary (analysis of all sequenced NCYC genomes. NRPS = non-ribosomal peptide synthase, RiPPs = ribosomally synthesised & post-translationally modified peptides, PKS = poly-ketide synthase. Many of the gene clusters in the 'Others' category are 'cf_putative', meaning that they are putative gene clusters of unknown type predicted by the 'clusterfinder' algorithm (high FP rate).

Since this pipeline is in its early stages, no testing data is presented here. More work needs to be done before this is a fully operational pipeline. Existing code is available at `https://github.com/chrispyatt/flagdown` and work is ongoing to fully automate the analysis.

6.5 Discussion & Conclusions

The FindClusters pipeline appears to be a success. It results in an easily interpretable output table and figure that can be used in further analysis. In this case it has effectively described the extent and appearance of the CAR gene cluster in a small-medium sized genome dataset, without the need to manually piece this information together. The obvious drawback of the program is that it is fooled by poor assembly quality and therefore misses partial gene clusters that may be caught by eye. However, this is more a function of assembly quality than the effectiveness of FindClusters. On a practical note, the image file sizes are too large (in the case of the PNG output) for most applications. This was done to enable the images to be printed clearly and can always be rectified with other software if required. The PDF output is also useful if vector graphics are preferred. There is no reason why the pipeline should be confined to yeast genomes so it might be interesting to take other datasets of related genomes and search for gene cluster variants. For example *Fusarium* are known to produce their own carotenoids from a partially clustered biosynthetic pathway (Avalos et al. 2017). It may be that this pathway is fully clustered in another organism's genome; this would be readily testable using FindClusters.

Specifically thinking about the *Rhodotorula* results, it is encouraging that the output seems to reflect what one instinctively thinks about the secondary metabolite complement of this genus. That is, a substantial part is dedicated to compounds that may be associated with the pigmentation for which the genus is named. Further investigation is needed to assess the 'clusterfinder' predictions that make up the majority of the gene clusters predicted in both the *Rhodotorula* dataset and the collection as a whole. One of the aims of this chapter was to get a rough idea of what types of gene cluster are most prevalent in the NCYC collection. A brief analysis using state-of-the-art computational tools suggested an abundance of Terpenoid gene clusters, and very few PKS gene clusters. This is interesting given that PKS based gene clusters are one of the main backbone types looked for by antiSMASH and other tools. Clearly they are not well represented in yeasts for whatever reason. PKS products are frequently involved in antibiotic or antifungal compounds (Koehn et al. 2005; Wawrik et al. 2005) so it seems surprising that they are underrepresented, but perhaps this function is taken by other types of product in yeasts. What is clear is that there is potentially a large number of gene clusters (and their products) that may be useful to us.

The enzyme clustering based approach for flagging potential gene clusters is in its infancy. I am aware of several other attempts to identify novel metabolic gene clusters through new methods. These were mentioned in the introduction of this thesis, one being a machine learning approach to classify clusters by protein domains, the other a deep learning method that is somewhat more advanced (DeepBGC). These methods may all complement each other with the ideal end result being those that work well being integrated into a single platform. So far, antiSMASH has been that platform, since it is the most advanced and supported.

In summary, the principle problem to be solved in the field of computational gene cluster prediction is how to predict truly novel gene clusters without relying on existing annotations and experimental data. Clearly this is not possible in the case of metagenomic sampling or *de novo* sequencing of unusual organisms. As yet, there is no clear solution to this problem, but there are several methods being developed that could make some inroads. Perhaps in a few years' time, there will something more concrete.

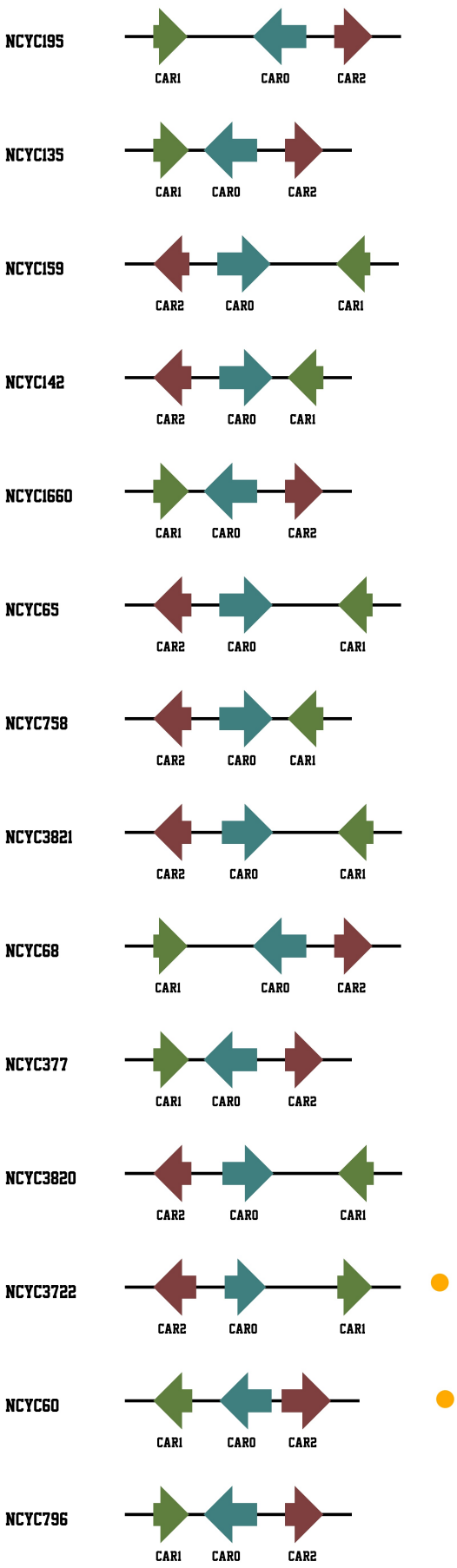
NCYC195
CAR1
CAR0
CAR2
NCYC135
CAR1
CAR0
CAR2
NCYC159
CAR2
CAR0
CAR1
NCYC142
CAR2
CAR0
CAR1
NCYC1660
CAR1
CAR0
CAR2
NCYC65
CAR2
CAR0
CAR1
NCYC758
CAR2
CAR0
CAR1
NCYC3821
CAR2
CAR0
CAR1
NCYC68
CAR1
CAR0
CAR2
NCYC377
CAR1
CAR0
CAR2
NCYC3820
CAR2
CAR0
CAR1
NCYC3722
CAR2
CAR0
CAR1
NCYC60
CAR1
CAR0
CAR2
NCYC796
CAR1
CAR0
CAR2

Figure 6.3

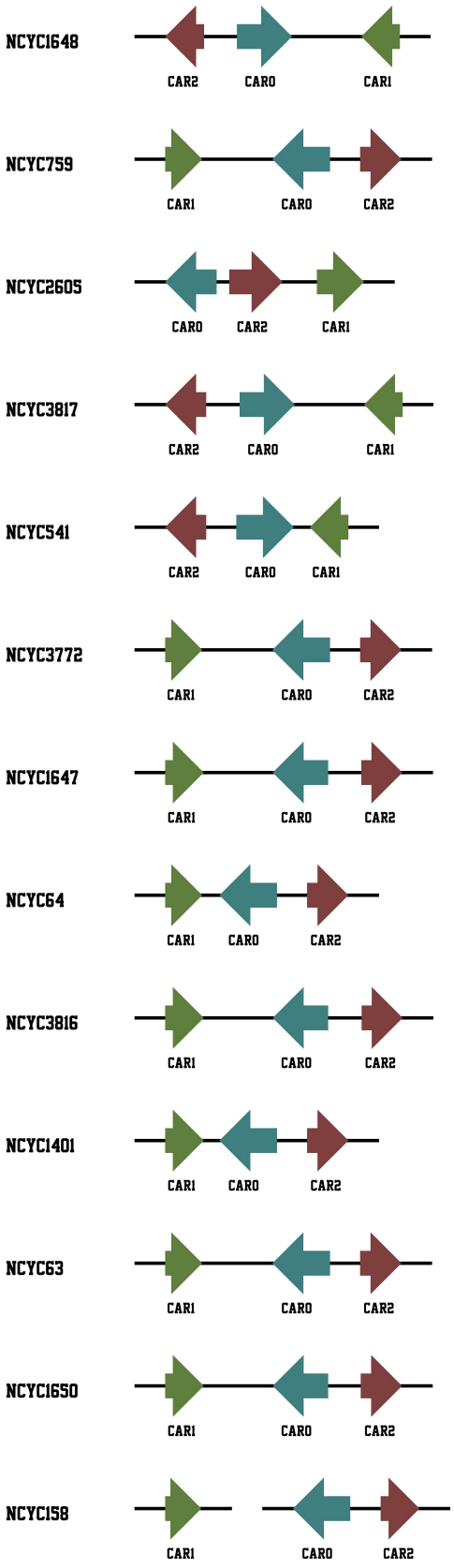
NCYC1648
CAR2
CAR0
CAR1
NCYC759
CAR1
CAR0
CAR2
NCYC2605
CAR0
CAR2
CAR1
NCYC3817
CAR2
CAR0
CAR1
NCYC541
CAR2
CAR0
CAR1
NCYC3772
CAR1
CAR0
CAR2
NCYC1647
CAR1
CAR0
CAR2
NCYC64
CAR1
CAR0
CAR2
NCYC3816
CAR1
CAR0
CAR2
NCYC1401
CAR1
CAR0
CAR2
NCYC63
CAR1
CAR0
CAR2
NCYC1650
CAR1
CAR0
CAR2
NCYC158
CAR1
CAR0
CAR2

Figure 6.3

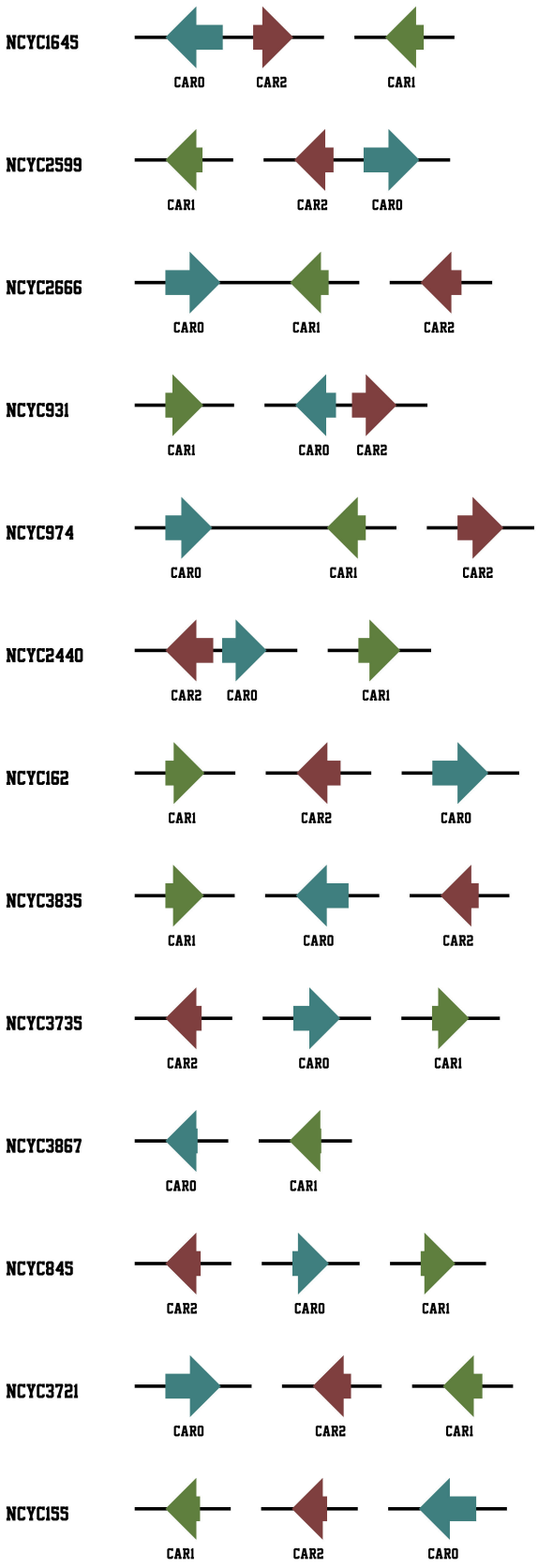
NCYC1645
CAR0
CAR2
CAR1
NCYC2599
CAR1
CAR2
CAR0
NCYC2666
CAR0
CAR1
CAR2
NCYC931
CAR1
CAR0
CAR2
NCYC974
CAR0
CAR1
CAR2
NCYC2440
CAR2
CAR0
CAR1
NCYC162
CAR1
CAR2
CAR0
NCYC3835
CAR1
CAR0
CAR2
NCYC3735
CAR2
CAR0
CAR1
NCYC3867
CAR0
CAR1
NCYC845
CAR2
CAR0
CAR1
NCYC3721
CAR0
CAR2
CAR1
NCYC155
CAR1
CAR2
CAR0

Figure 6.3

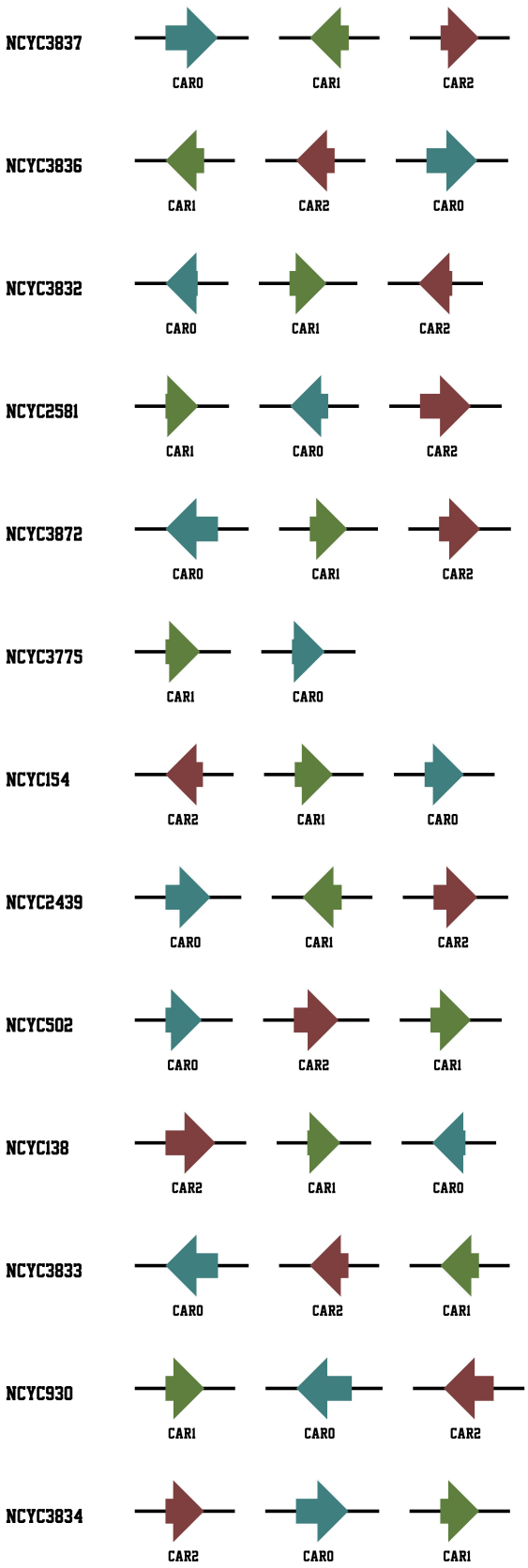

Figure 6.3: Pipeline output for the CAR gene cluster search in *Rhodotorula* genomes. Only NCYC60 and NCYC3722 (*R. glutins var. glutinis* and *R. graminis*; marked by orange dot) can be seen to mirror the configuration seen in Figure 6.2. NCYC2605 (*R. vanillica*; also marked) also varies in its position of CAR1.

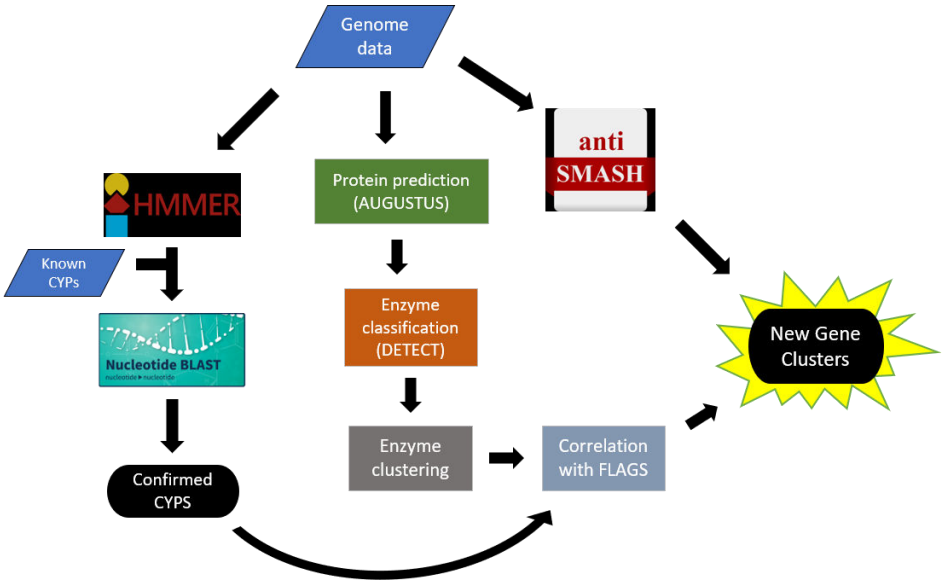

Figure 6.4: Diagrammatic representation of the approach using common gene 'flags' as markers around which to centre enzyme clustering searches. The approach is intended as a complement to existing gene cluster discovery tools such as antiSMASH. In the example shown, CYPs are used as the 'flags' but in theory any gene set could be used if it is deemed to be a gene cluster marker.

7 Discussion

7.1 Main goals

The overarching story of this thesis concerns the search of genomic data, in this case yeast genomes from the NCYC genome sequencing project, for potentially industrially useful or evolutionarily interesting metabolic gene clusters. The key aims of the project were: (1) to assess the gene cluster content of the NCYC collection using the sequenced genomes of a significant subset (1,000 strains) as a proxy, (2) to investigate the evolutionary history of certain gene clusters in the yeast collection through phylogenetic analysis, (3) to assess the suitability of current computational tools for yeast gene cluster discovery and develop new methods if needed.

7.2 Outcomes

Assessing the gene cluster content of the NCYC collection

Referring to the first goal above, we now have a catalogue of computationally predicted gene clusters for strains spanning the full taxonomic breadth of the NCYC collection (see Table 6.2). This has been achieved largely through the use of existing state-of-the-art tools, mainly antiSMASH (Medema et al. 2011). Yeasts in general have never before been investigated in this way on such scale so this represents a substantial cache of information that can be mined further by those looking to exploit such gene clusters. In addition, the information contains a large number of less conclusive predictions courtesy of the 'clusterfinder' algorithm, now included in antiSMASH as of v3 (Blin et al. 2017), which invite further investigation.

In the end, more than 9,000 gene clusters were predicted from the collection. The majority of these are of unknown type, and may be false positives from the 'clusterfinder' algorithm, but there are still at least 1,400 putative gene clusters of established types (see Table 6.2 for details). Therefore the answer to the question of what kinds of metabolite gene clusters exist in the yeast genomes searched here is that the majority of the identified gene clusters are

terpenoid based pathways. Terpenoids are often linked to flavour compounds in spices etc. (Korkmaz et al. 2017) so this may be a useful finding with regards to the food industry, if safe. This represents a useful dataset on which to base future work, particularly given that most natural product genome mining literature seems to concentrate on bacterial genomes.

Evolutionary analysis of known gene clusters

The second goal of the thesis was to describe the evolutionary relationships and taxonomic extent of some of the more established yeast gene clusters. In this case this is the two major biosurfactant gene clusters synthesising cellobiose lipids and mannosylerythritol lipids. The yeast species reported to produce these compounds were fairly well sampled by the genome sequencing project and so represented perfect test beds in which to examine the evolution of gene clusters in yeast. Previously published work had described these two gene clusters in detail in *U. maydis* (Hewald et al. 2006; Teichmann et al. 2007) and *P. flocculosa* (the latter for CBLs only, Teichmann et al. (2011)), with minimal examination of the draft genomes of several other known producers including *P. hubeiensis*, *P. aphidis*, and *P. antarctica* (i.e. simple blast searches confirming presence/absence of the genes without any positional information) (Konishi et al. 2011; Morita et al. 2013a; Saika et al. 2014).

With the results presented in this thesis, this knowledge base has been expanded to include the positional arrangements of the gene clusters in a larger group of species, as well as confirmation of the taxonomic extent of the pathways concerned. It has also added NCYC3267, a novel species, and some *Sporisorium*, *Ustilago*, and *Melanopsichium* strains to the list of potential MEL producers (this would need to be confirmed experimentally but the biosynthetic pathway is certainly there). Questions remain regarding *P. tsukubaensis*, *P. rugulosa*, *P. crassa*, and *P. churashimaensis*, reported producers that do not appear to contain the MEL gene cluster. This result would ideally be confirmed with more sequencing if more time were available, ideally using long read technology. The work presented here has also established that the CBL gene cluster seen in *P. flocculosa* is an unusual form not seen anywhere else. It has also shed light on the question of whether the Ustilaginomycete and Tremellomycete CBL products are evolutionarily related (the biosynthetic pathways appear to have evolved independently and be very differently constructed, ableit from similar gene types, so the answer is likely 'no').

Software assessment and development

An analysis pipeline has been developed, FindClusters, that searches a given set of genome assemblies for the genes making up a given gene cluster. In addition, one more pipeline is in development, Flagdown, that will assess a genome assembly for clusters of metabolic enzymes that coincide with 'flag' genes, gene types commonly found in gene clusters.

Prior to the beginning of this project, computational tools aiming to address this problem were limited both in number and in scope. The principle, field-leading, software tool, antiSMASH (Medema et al. 2011), has been through two major updates since the start of this project in late 2015, with the release of versions 3 and then 4 (Blin et al. 2017; Weber et al. 2015). It has also expanded to include a more diverse set of gene cluster rules with which to search for novel gene clusters, as well as being modified to work better with plant and fungal genomes (where previously it was optimised for bacterial sequences), see Kautsar et al. (2017) and Blin et al. (2017). Apart from antiSMASH and the algorithms that have been incorporated into it, there are no other major players in the computational gene cluster discovery arena.

An examination of the collection's sequenced genomes has established that potentially novel cytochromes P450 exist in fairly large numbers in certain areas of the yeast taxonomy, coinciding with areas in which we might expect greater metabolic potential (i.e. those strains more closely related to the metabolically diverse filamentous fungi). This finding suggests there is potential to use genes from large metabolism-related gene families (such as P450s) as 'flags' for finding new gene clusters in unexplored genomes ('Flagdown', above). Since tools such as antiSMASH are inherently rule-based and can only find gene clusters of types that have been described before, this new approach may allow the discovery of different gene cluster types, although the error rate is likely to be higher.

Other methods attempting to bring machine learning/AI into the gene cluster discovery field are being developed separately but are also still at various stages of early development (by protein domains - Hayda Almeida, Personal Comm., DeepBGC - Chris Woelk, Personal Comm., substrate prediction - Stefan Gunther, Personal Comm.). These all represent distinct and useful additions to the field that can act in complement with the methods described in this thesis.

7.3 Future directions

Gene clusters in the NCYC collection

The most obvious next step regarding the gene cluster catalogue established here is to investigate the predicted gene clusters to find out what specific product they are producing. With the knowledge of approximately what kind of biosynthetic pathway is being investigated, it may be slightly easier to replicate suitable culture conditions that may be amenable to stimulating production. Obviously the major obstacle to this is the sheer scale of the dataset, however it remains possible that better substrate or product prediction methods may appear in the near future (this is an active area of research) and improve the process of choosing targets to investigate.

The specific case of the *Rhodotorula*, mentioned in the final results chapter of this thesis, may be a good place to start. The genus has been shown to contain a sizeable secondary metabolite gene cluster complement which can be screened. There are groups on the Norwich Research Park who have expressed interest in screening the *Rhodotorula* strains within the NCYC collection for metabolites. This could take place in conjunction with association studies to determine which, if any, of the predicted gene clusters are being used to produce any metabolites discovered.

Evolution of MEL & CBL gene clusters

Regarding the MEL gene cluster, as stated above it would be helpful to do some confirmatory sequencing in the cases of those strains that should, according to literature, contain the MEL pathway but do not appear to. In addition there are other strains that would be interesting to add to the analysis (those towards the bottom of Table 3.1) since they are reported producers, in some cases producing unusual types of MEL.

In the case of the CBL gene cluster the main pieces of extra work that would be advantageous to carry out would be a more thorough sampling and sequencing attempt of the Tremallales clade, to determine whether the CBL pathway really is in various states of assembly as suggested in Chapter 4. It would also be good to carry out some RNA-seq, or similar, work in the Moesziomyces strains to confirm whether the genes identified as homologues to the CBL genes are in fact involved in CBL biosynthesis. Since the current hypothesis is that they are not, additional work would aim to discover what genes *are* responsible for CBL biosynthesis in this group, and whether they too form an as-yet undiscovered 3rd CBL gene cluster type.

For both gene clusters it would be interesting to determine whether the gene cluster arrangement has any bearing on the product produced, although at present there may not be

enough examples of each to draw any reliable association conclusions.

One of the main issues encountered over the course of this project has been genome assembly quality. A small number of strains were sequenced multiple times to improve this factor but this was not the case for the majority of strains sequenced, due to the short read nature of the datasets. There were also issues with cross contamination at some unknown stage of the process, meaning that some potentially useful genome sequences were unavailable. It would be worthwhile to work towards improving the assemblies generated by the sequencing project. A change to long read sequencing would, in general, be advantageous for gene cluster finding. Fortunately, with the development of the FindClusters pipeline it would be possible to re-run most of the analyses described in Chapters 3 and 4 in much less time, with better input assemblies. It would be interesting to see what effect this would have, particularly with regard to those strains where gene clusters could not be found despite reports of metabolite production.

Software development

The work described in this thesis lays the groundwork for a more comprehensive methodology to be developed. With a bit more time it would be possible to stitch together the disparate parts of the enzyme clustering pipeline ('Flagdown') and incorporate a greater number of 'flag' gene types into the analysis. The intention is to allow the user of the final application to either choose from a list of pre-loaded 'flag' genes (shipped with the software, CYPs being the first to be established, Major Facilitator transporters next) or supply their own in the form of an alignment of examples. This would then be put through an extra step of HMMER searching the supplied genome(s) for locations containing the given 'flag' gene type, before colocating those with the enzyme clustering predictions.

7.4 Final conclusions

High-throughput gene cluster finding is possible in yeast genomes. There appears to be a significant number of gene clusters to find, particularly in the Basidiomycetes. Good quality genome assemblies, perhaps derived from long read data, are key to producing good results. Current software approaches are not yet optimally tailored to yeast gene clusters. However, this will be achievable with time.

Bibliography

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990). "Basic Local Alignment Search Tool". In: *Journal Of Molecular Biology* 215.3, pp. 403–410. ISSN: 0022-2836. DOI: 10.1006/Jmbi.1990.9999.

Arutchelvi, J. I., S. Bhaduri, P. V. Uppara, and M. Doble (2008). "Mannosylerythritol Lipids: A Review". In: *Journal Of Industrial Microbiology & Biotechnology* 35.12, pp. 1559–1570. ISSN: 1367-5435. DOI: 10.1007/S10295-008-0460-4.

Avalos, J., J. Pardo-Medina, O. Parra-Rivero, M. Ruger-Herreros, R. Rodriguez-Ortiz, D. Hornero-Mendez, and M. Carmen Limon (May 2017). "Carotenoid Biosynthesis In Fusarium". In: *Journal Of Fungi* 3.3. DOI: 10.3390/Jof3030039.

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Y. Ren, W. W. Li, and W. S. Noble (2009). "Meme Suite: Tools For Motif Discovery And Searching". In: *Nucleic Acids Research* 37, W202–W208. ISSN: 0305-1048. DOI: 10.1093/Nar/Gkp335.

Bailey, T. L., J. Johnson, C. E. Grant, and W. S. Noble (2015). "The Meme Suite". In: *Nucleic Acids Research* 43.W1, W39–W49. ISSN: 0305-1048. DOI: 10.1093/Nar/Gkv416.

Besemer, J. and M. Borodovsky (2005). "Genemark: Web Software For Gene Finding In Prokaryotes, Eukaryotes And Viruses". In: *Nucleic Acids Research* 33, W451–W454. ISSN: 0305-1048. DOI: 10.1093/Nar/Gki487.

Biswas, Sk., K. Yokoyama, K. Nishimura, and M. Miyaji (Sept. 2001). "Molecular Phylogenetics Of The Genus Rhodotorula And Related Basidiomycetous Yeasts Inferred From The Mitochondrial Cytochrome B Gene". In: *International Journal Of Systematic And Evolutionary Microbiology* 51.3, pp. 1191–1199. ISSN: 1466-5026. DOI: 10.1099/00207713-51-3-1191.

Blin, K. et al. (2017). "Antismash 4.0-Improvements In Chemistry Prediction And Gene Cluster Boundary Identification". In: *Nucleic Acids Research* 45.W1, W36–W41. ISSN: 0305-1048. DOI: 10.1093/Nar/Gkx319.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel (Apr. 2014). "Trimmomatic: A Flexible Trimmer For Illumina Sequence Data". In: *Bioinformatics* 30.15, pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/Bioinformatics/Btu170. eprint: Http://Oup.Prod.Sis.Lan/Bioinformatics/Article-Pdf/30/15/2114/17143152/Btu170.Pdf.

Brendel, V., L. Xing, and W. Zhu (2004). "Gene Structure Prediction From Consensus Spliced Alignment Of Multiple Ests Matching The Same Genomic Locus". In: *Bioinformatics* 20.7, pp. 1157–69. ISSN: 1367-4803 (Print) 1367-4803 (Linking). DOI: 10.1093/Bioinformatics/Bth058.

Cameotra, S. S. and R. S. Makkar (2004). "Recent Applications Of Biosurfactants As Biological And Immunological Molecules". In: *Current Opinion In Microbiology* 7.3, pp. 262–266. ISSN: 1369-5274. DOI: 10.1016/J.Mib.2004.04.006.

Capella-Gutierrez, Salvador, Jose M. Silla-Martinez, and Toni Gabaldon (2009). "Trimal: A Tool For Automated Alignment Trimming In Large-Scale Phylogenetic Analyses". In: *Bioinformatics* 25.15, pp. 1972–1973. ISSN: 1367-4803. DOI: 10.1093/Bioinformatics/Btp348.

Cary, J. W. and K. C. Ehrlich (2006). "Aflatoxigenicity In Aspergillus: Molecular Genetics, Phylogenetic Relationships And Evolutionary Implications". In: *Mycopathologia* 162.3, pp. 167–177. ISSN: 0301-486x. DOI: 10.1007/S11046-006-0051-8.

Cimermancic, P. et al. (2014). "Insights Into Secondary Metabolism From A Global Analysis Of Prokaryotic Biosynthetic Gene Clusters". In: *Cell* 158.2, pp. 412–421. ISSN: 0092-8674. DOI: 10.1016/J.Cell.2014.06.034.

Clayton, Christine (2019). "Regulation Of Gene Expression In Trypanosomatids: Living With Polycistronic Transcription". In: *Open Biology* 9.6, p. 190072. DOI: 10.1098/rsob.190072.

Cock, Peter J. A. et al. (2009). "Biopython: Freely Available Python Tools For Computational Molecular Biology And Bioinformatics". In: *Bioinformatics* 25.11, pp. 1422–1423. ISSN: 1367-4803. DOI: 10.1093/Bioinformatics/Btp163.

Cohen, Noa E., Roy Shen, and Liran Carmel (July 2011). "The Role Of Reverse Transcriptase In Intron Gain And Loss Mechanisms". In: *Molecular Biology And Evolution* 29.1, pp. 179–186. ISSN: 0737-4038. DOI: 10.1093/Molbev/Msr192. eprint: Http://Oup.Prod.Sis.Lan/Mbe/Article-Pdf/29/1/179/13648489/Msr192.Pdf.

Cook, D. J., J. D. Finnigan, K. Cook, G. W. Black, and S. J. Charnock (2016). "Cytochromes P450: History, Classes, Catalytic Mechanism, And Industrial Application". In: *Insights Into Enzyme Mechanisms And Functions From Experimental And Computational Methods*. Ed. by C. Z. Christov. Vol. 105. Advances In Protein Chemistry And Structural Biology, pp. 105–126. ISBN: 978-0-12-804825-2. DOI: 10.1016/Bs.Apcsb.2016.07.003.

Cruz-Morales, Pablo, Christian E. Martínez-Guerrero, Marco A. Morales-Escalante, Luis A. Yáñez-Guerra, Johannes F. Kopp, Jörg Feldmann, Hilda E. Ramos-Aboites, and Francisco Barona-Gómez (2015). "Recapitulation Of The Evolution Of Biosynthetic Gene Clusters

Reveals Hidden Chemical Diversity On Bacterial Genomes." In: *Biorxiv*. DOI: 10.1101/020503.

Delcher, A. L., K. A. Bratke, E. C. Powers, and S. L. Salzberg (2007). "Identifying Bacterial Genes And Endosymbiont Dna With Glimmer". In: *Bioinformatics* 23.6, pp. 673–679. ISSN: 1367-4803. DOI: 10.1093/Bioinformatics/Btm009.

Eddy, S. R. (1998). "Profile Hidden Markov Models". In: *Bioinformatics* 14.9, pp. 755–763. ISSN: 1367-4803. DOI: 10.1093/Bioinformatics/14.9.755.

— (2015). *Hmmer*. Web Page.

Edgar, R. C. (2004). "Muscle: Multiple Sequence Alignment With High Accuracy And High Throughput". In: *Nucleic Acids Research* 32.5, pp. 1792–1797. ISSN: 0305-1048. DOI: 10.1093/Nar/Gkh340.

Emmanuel, E., K. Hanna, C. Bazin, G. Keck, B. Clement, and Y. Perrodin (2005). "Fate Of Glutaraldehyde In Hospital Wastewater And Combined Effects Of Glutaraldehyde And Surfactants On Aquatic Organisms". In: *Environment International* 31.3, pp. 399–406. ISSN: 0160-4120. DOI: 10.1016/J.Envint.2004.08.011.

Fan, L. L., Y. C. Dong, Y. F. Fan, J. Zhang, and Q. H. Chen (2014). "Production And Identification Of Mannosylerythritol Lipid-A Homologs From The Ustilaginomycetous Yeast Pseudozyma Aphidis Zjudm34". In: *Carbohydrate Research* 392, pp. 1–6. ISSN: 0008-6215. DOI: 10.1016/J.Carres.2014.04.013.

Fukuoka, Tokuma, Toinotake Morita, Masaaki Konishi, Tomohiro Imura, and Dai Kitamoto (2008). "A Basidiomycetous Yeast, Pseudozyma Tsukubaensis, Efficiently Produces A Novel Glycolipid Biosurfactant. The Identification Of A New Diastereomer Of Mannosylerythritol Lipid-B". In: *Carbohydrate Research* 343.3, pp. 555–560. ISSN: 0008-6215. DOI: 10.1016/J.Carres.2007.11.023.

Fukuoka, Tokuma, Tomotake Morita, Masaaki Konishi, Tomohiro Imura, and Dai Kitamoto (2007a). "Characterization Of New Glycolipid Biosurfactants, Tri-Acylated Mannosylerythritol Lipids, Produced By Pseudozyma Yeasts". In: *Biotechnology Letters* 29.7, pp. 1111–1118. ISSN: 0141-5492. DOI: 10.1007/S10529-007-9363-0.

— (2007b). "Characterization Of New Types Of Mannosylerythritol Lipids As Biosurfactants Produced From Soybean Oil By A Basidiomycetous Yeast, Pseudozyma Shanxiensis". In: *Journal Of Oleo Science* 56.8, pp. 435–42.

Fukuoka, Tokuma, Tomotake Morita, Masaaki Konishi, Tomohiro Imura, Hideki Sakai, and Dai Kitamoto (2007c). "Structural Characterization And Surface-Active Properties Of A New Glycolipid Biosurfactant, Mono-Acylated Mannosylerythritol Lipid, Produced From Glucose By Pseudozyma Antarctica". In: *Applied Microbiology And Biotechnology* 76.4, pp. 801–810. ISSN: 0175-7598. DOI: 10.1007/S00253-007-1051-4.

Gardiner, Donald M. and Barbara J. Howlett (July 2005). "Bioinformatic And Expression Analysis Of The Putative Gliotoxin Biosynthetic Gene Cluster Of Aspergillus Fumigatus". In: *Fems Microbiology Letters* 248.2, pp. 241–248. ISSN: 0378-1097. DOI: 10.1016/J.Femsle.2005.05.046. eprint: Http://Oup.Prod.Sis.Lan/Femsle/Article-Pdf/248/2/241/19122742/248-2-241.Pdf.

El-Gebali, Sara et al. (2019). "The Pfam Protein Families Database In 2019". In: *Nucleic Acids Research* 47.D1, pp. D427–D432. ISSN: 0305-1048. DOI: 10.1093/Nar/Gky995.

Golubev, W. I., T. V. Kulakovskaya, A. S. Shashkov, E. V. Kulakovskaya, and N. V. Golubev (2008). "Antifungal Cellobiose Lipid Secreted By The Epiphytic Yeast Pseudozyma Graminicola". In: *Microbiology* 77.2, pp. 171–175. ISSN: 0026-2617. DOI: 10.1134/S00262617080200

Haskins, R. H. (1950). "Biochemistry Of The Ustilaginales .1. Preliminary Cultural Studies Of Ustilago Zeae". In: *Canadian Journal Of Research Section C-Botanical Sciences* 28.2, pp. 213–&. ISSN: 0366-7405. DOI: 10.1139/Cjr50c-012.

Hawksworth, David L. and Robert Lücking (2017). "Fungal Diversity Revisited: 2.2 To 3.8 Million Species". In: *The Fungal Kingdom*. American Society Of Microbiology, pp. 79–95.

Hesham, A. E. L., S. Khan, Y. Tao, D. Li, Y. Zhang, and M. Yang (2012). "Biodegradation Of High Molecular Weight Pahs Using Isolated Yeast Mixtures: Application Of Meta-Genomic Methods For Community Structure Analyses". In: *Environmental Science And Pollution Research* 19.8, pp. 3568–3578. ISSN: 0944-1344. DOI: 10.1007/S11356-012-0919-8.

Hewald, S., U. Linne, M. Scherer, M. A. Marahiel, J. Kamper, and M. Bolker (2006). "Identification Of A Gene Cluster For Biosynthesis Of Mannosylerythritol Lipids In The Basidiomycetous Fungus Ustilago Maydis". In: *Appl Environ Microbiol* 72.8, pp. 5469–77. ISSN: 0099-2240 (Print) 0099-2240 (Linking). DOI: 10.1128/Aem.00506-06.

Hittinger, Chris Todd, Antonis Rokas, and Sean B. Carroll (2004). "Parallel Inactivation Of Multiple GAL Pathway Genes And Ecological Diversification In Yeasts". In: *Proceedings of the National Academy of Sciences* 101.39, pp. 14144–14149. ISSN: 0027-8424. DOI: 10.1073/pnas.0404319101. eprint: https://www.pnas.org/content/101/39/14144.full.pdf.

Hung, S. S., J. Wasmuth, C. Sanford, and J. Parkinson (2010). "Detect-A Density Estimation Tool For Enzyme Classification And Its Application To Plasmodium Falciparum". In: *Bioinformatics* 26.14, pp. 1690–1698. ISSN: 1367-4803. DOI: 10.1093/Bioinformatics/Btq266.

Inoh, Y., D. Kitamoto, N. Hirashima, and M. Nakanishi (2004). "Biosurfactant Mel-A Dramatically Increases Gene Transfection Via Membrane Fusion". In: *Journal Of Controlled Release* 94.2-3, pp. 423–431. ISSN: 0168-3659. DOI: 10.1016/J.Jconrel.2003.10.020.

Jezierska, S., S. Claus, and I. Van Bogaert (2018). "Yeast Glycolipid Biosurfactants". In: *Febs Letters* 592.8, pp. 1312–1329. ISSN: 0014-5793. DOI: 10.1002/1873-3468.12888.

Kautsar, Satria A., Hernando G. Suarez Duran, Kai Blin, Anne Osbourn, and Marnix H. Medema (Apr. 2017). "Plantismash: Automated Identification, Annotation And Expression Analysis Of Plant Biosynthetic Gene Clusters". In: *Nucleic Acids Research* 45.W1, W55–W63. ISSN: 0305-1048. DOI: 10.1093/Nar/Gkx305. eprint: Http://Oup.Prod.Sis.Lan/Nar/Article-Pdf/45/W1/W55/18137272/Gkx305.Pdf.

Kearse, Matthew et al. (2012). "Geneious Basic: An Integrated And Extendable Desktop Software Platform For The Organization And Analysis Of Sequence Data". In: *Bioinformatics* 28.12, pp. 1647–1649. ISSN: 1367-4803. DOI: 10.1093/Bioinformatics/Bts199.

Keller, N. P. and T. M. Hohn (1997). "Metabolic Pathway Gene Clusters In Filamentous Fungi". In: *Fungal Genetics And Biology* 21.1, pp. 17–29. ISSN: 1087-1845. DOI: 10.1006/Fgbi.1997.0970.

Kelly, D. E., N. Krasevec, J. Mullins, and D. R. Nelson (2009). "The Cypome (Cytochrome P450 Complement) Of Aspergillus Nidulans". In: *Fungal Genetics And Biology* 46, S53–S61. ISSN: 1087-1845. DOI: 10.1016/J.Fgb.2008.08.010.

Khaldi, N., J. Collemare, M. H. Lebrun, and K. H. Wolfe (2008). "Evidence For Horizontal Transfer Of A Secondary Metabolite Gene Cluster Between Fungi". In: *Genome Biology* 9.1. ISSN: 1474-760x. DOI: 10.1186/Gb-2008-9-1-R18.

Khaldi, N., F. T. Seifuddin, G. Turner, D. Haft, W. C. Nierman, K. H. Wolfe, and N. D. Fedorova (2010). "Smurf: Genomic Mapping Of Fungal Secondary Metabolite Clusters". In: *Fungal Genetics And Biology* 47.9, pp. 736–741. ISSN: 1087-1845. DOI: 10.1016/J.Fgb.2010.06.003.

Kitamoto, D., K. Haneishi, T. Nakahara, and T. Tabuchi (1990). "Production Of Mannosylerythritol Lipids By Candida Antarctica From Vegetable Oils". In: *Agricultural And Biological Chemistry* 54.1, pp. 37–40. ISSN: 0002-1369.

Kitamoto, Dai, Tomotake Morita, Tokuma Fukuoka, Masa-Aki Konishi, and Tomohiro Imura (2009). "Self-Assembling Properties Of Glycolipid Biosurfactants And Their Potential Applications". In: *Current Opinion In Colloid & Interface Science* 14.5, pp. 315–328. ISSN: 1359-0294. DOI: 10.1016/J.Cocis.2009.05.009.

Koehn, F. E. and G. T. Carter (2005). "The Evolving Role Of Natural Products In Drug Discovery". In: *Nature Reviews Drug Discovery* 4.3, pp. 206–220. ISSN: 1474-1776. DOI: 10.1038/Nrd1657.

Konishi, M., Y. Hatada, and J. Horiuchi (2013). "Draft Genome Sequence Of The Basidiomycetous Yeast-Like Fungus Pseudozyma Hubeiensis Sy62, Which Produces An Abun-

dant Amount Of The Biosurfactant Mannosylerythritol Lipids". In: *Genome Announc* 1.4. ISSN: 2169-8287 (Electronic). DOI: 10.1128/Genomea.00409-13.

Konishi, M., T. Nagahama, T. Fukuoka, T. Morita, T. Imura, D. Kitamoto, and Y. Hatada (2011). "Yeast Extract Stimulates Production Of Glycolipid Biosurfactants, Mannosylerythritol Lipids, By Pseudozyma Hubeiensis Sy62". In: *Journal Of Bioscience And Bioengineering* 111.6, pp. 702–705. ISSN: 1389-1723. DOI: 10.1016/J.Jbiosc.2011.02.004.

Korkmaz, Aziz, Ali Adnan Hayaloglu, and Ahmet Ferit Atasoy (2017). "Evaluation Of The Volatile Compounds Of Fresh Ripened Capsicum Annuum And Its Spice Pepper (Dried Red Pepper Flakes And Isot)". In: *Lwt* 84, pp. 842–850. ISSN: 0023-6438. DOI: Https://Doi.Org/10.1016/J.Lwt.2017.06.058.

Krause, David J. et al. (Oct. 2018). "Functional And Evolutionary Characterization Of A Secondary Metabolite Gene Cluster In Budding Yeasts". In: *Proceedings Of The National Academy Of Sciences Of The Unites States Of America* 115.43, pp. 11030–11035. ISSN: 0027-8424. DOI: 10.1073/Pnas.1806268115.

Kulakovskaya, E. V., T. V. Kulakovskaya, V. I. Golubev, A. S. Shashkov, A. A. Grachev, and N. E. Nifantiev (2007). "Fungicidal Activity Of Cellobiose Lipids From Culture Broth Of Yeast Cryptococcus Humicola And Pseudozyma Fusiformata". In: *Russian Journal Of Bioorganic Chemistry* 33.1, pp. 156–160. ISSN: 1068-1620. DOI: 10.1134/S1068162007010189.

Kulakovskaya, T. V., W. I. Golubev, M. A. Tomashevskaya, E. V. Kulakovskaya, A. S. Shashkov, A. A. Grachev, A. S. Chizhov, and N. E. Nifantiev (2010). "Production Of Antifungal Cellobiose Lipids By Trichosporon Porosum". In: *Mycopathologia* 169.2, pp. 117–123. ISSN: 0301-486x. DOI: 10.1007/S11046-009-9236-2.

Kulakovskaya, T. V., A. S. Shashkov, E. V. Kulakovskaya, and W. I. Golubev (2004). "Characterization Of An Antifungal Glycolipid Secreted By The Yeast Sympodiomycopsis Paphiopedili". In: *Fems Yeast Research* 5.3, pp. 247–252. ISSN: 1567-1356. DOI: 10.1016/J.Femysr.2004.07.008.

— (2005). "Ustilagic Acid Secretion By Pseudozyma Fusiformata Strains". In: *Fems Yeast Research* 5.10, pp. 919–923. ISSN: 1567-1356. DOI: 10.1016/J.Femsyr.2005.04.006.

Kurtzman, Cletus P. and Jure Piskur (2006). "Taxonomy And Phylogenetic Diversity Among The Yeasts". In: *Comparative Genomics: Using Fungi As Models*. Ed. by Per Sunnerhagen and Jure Piskur. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 29–46. ISBN: 978-3-540-31495-0. DOI: 10.1007/B106654.

Landolfo, S., G. Ianiri, S. Camiolo, A. Porceddu, G. Mulas, R. Chessa, G. Zara, and I. Mannazzu (2018). "Car Gene Cluster And Transcript Levels Of Carotenogenic Genes In

Rhodotorula Mucilaginosa". In: *Microbiology-Sgm* 164.1, pp. 78–87. ISSN: 1350-0872. DOI: 10.1099/Mic.0.000588.

Lanfear, Robert, Paul B. Frandsen, April M. Wright, Tereza Senfeld, and Brett Calcott (2017). "Partitionfinder 2: New Methods For Selecting Partitioned Models Of Evolution For Molecular And Morphological Phylogenetic Analyses". In: *Molecular Biology And Evolution* 34.3, pp. 772–773. ISSN: 0737-4038. DOI: 10.1093/Molbev/Msw260.

Lawrence, J. (1999). "Selfish Operons: The Evolutionary Impact Of Gene Clustering In Prokaryotes And Eukaryotes". In: *Current Opinion In Genetics & Development* 9.6, pp. 642–648. ISSN: 0959-437x. DOI: 10.1016/S0959-437x(99)00025-8.

Lawrence, J. G. and J. R. Roth (1996). "Selfish Operons: Horizontal Transfer May Drive The Evolution Of Gene Clusters". In: *Genetics* 143.4, pp. 1843–1860. ISSN: 0016-6731.

Lee, J. M. and E. L. L. Sonnhammer (2003). "Genomic Gene Clustering Analysis Of Pathways In Eukaryotes". In: *Genome Research* 13.5, pp. 875–882. ISSN: 1088-9051. DOI: 10.1101/Gr.737703.

Lefebvre, F., D. L. Joly, C. Labbe, B. Teichmann, R. Linning, F. Belzile, G. Bakkeren, and R. R. Belanger (2013). "The Transition From A Phytopathogenic Smut Ancestor To An Anamorphic Biocontrol Agent Deciphered By Comparative Whole-Genome Analysis". In: *Plant Cell* 25.6, pp. 1946–59. ISSN: 1532-298x (Electronic) 1040-4651 (Linking). DOI: 10.1105/Tpc.113.113969.

Li, M. H. T., P. M. U. Ung, J. Zajkowski, S. Garneau-Tsodikova, and D. H. Sherman (2009). "Automated Genome Mining For Natural Products". In: *Bmc Bioinformatics* 10. ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-185.

Lorenz, S., M. Guenther, C. Grumaz, S. Rupp, S. Zibek, and K. Sohn (2014). "Genome Sequence Of The Basidiomycetous Fungus Pseudozyma Aphidis Dsm70725, An Efficient Producer Of Biosurfactant Mannosylerythritol Lipids". In: *Genome Announc* 2.1. ISSN: 2169-8287 (Electronic). DOI: 10.1128/Genomea.00053-14.

MacGillivray, A. R. and M. P. Shiaris (1993). "Biotransformation Of Polycyclic Aromatic Hydrocarbons By Yeasts Isolated From Coastal Sediments". In: *Appl Environ Microbiol* 59.5, pp. 1613–8. ISSN: 0099-2240 (Print) 0099-2240.

Makkar, R. S. and S. S. Cameotra (2002). "An Update On The Use Of Unconventional Substrates For Biosurfactant Production And Their New Applications". In: *Applied Microbiology And Biotechnology* 58.4, pp. 428–434. ISSN: 0175-7598. DOI: 10.1007/S00253-001-0924-1.

Marchand, G., W. Remus-Borel, F. Chain, W. Hammami, F. Belzile, and R. R. Belanger (2009). "Identification Of Genes Potentially Involved In The Biocontrol Activity Of Pseu-

dozyma Flocculosa". In: *Phytopathology* 99.10, pp. 1142–1149. ISSN: 0031-949x. DOI: 10.1094/Phyto-99-10-1142.

Medema, M. H., K. Blin, P. Cimermancic, V. De Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano, and R. Breitling (2011). "Antismash: Rapid Identification, Annotation And Analysis Of Secondary Metabolite Biosynthesis Gene Clusters In Bacterial And Fungal Genome Sequences". In: *Nucleic Acids Research* 39, W339–W346. ISSN: 0305-1048. DOI: 10.1093/Nar/Gkr466.

Medema, M. H., E. Takano, and R. Breitling (2013). "Detecting Sequence Homology At The Gene Cluster Level With Multigeneblast". In: *Molecular Biology And Evolution* 30.5, pp. 1218–1223. ISSN: 0737-4038. DOI: 10.1093/Molbev/Mst025.

Medema, M. H. et al. (2015). "Minimum Information About A Biosynthetic Gene Cluster". In: *Nat Chem Biol* 11.9, pp. 625–31. ISSN: 1552-4469 (Electronic) 1552-4450 (Linking). DOI: 10.1038/Nchembio.1890.

Mimee, B., B. Labbe, R. Pelletier, and R. R. Belanger (2005). "Antifungal Activity Of Flocculosin, A Novel Glycolipid Isolated From Pseudozyma Flocculosa". In: *Antimicrobial Agents And Chemotherapy* 49.4, pp. 1597–1599. ISSN: 0066-4804. DOI: 10.1128/Aac.49.4.1597-1599.2005.

Mimee, B., R. Pelletier, and R. R. Belanger (2009). "In Vitro Antibacterial Activity And Antifungal Mode Of Action Of Flocculosin, A Membrane-Active Cellobiose Lipid". In: *Journal Of Applied Microbiology* 107.3, pp. 989–996. ISSN: 1364-5072. DOI: 10.1111/J.1365-2672.2009.04280.X.

Morita, T., T. Fukuoka, T. Imura, and D. Kitamoto (2013a). "Accumulation Of Cellobiose Lipids Under Nitrogen-Limiting Conditions By Two Ustilaginomycetous Yeasts, Pseudozyma Aphidis And Pseudozyma Hubeiensis". In: *Fems Yeast Research* 13.1, pp. 44–49. ISSN: 1567-1356. DOI: 10.1111/1567-1364.12005.

— (2013b). "Production Of Mannosylerythritol Lipids And Their Application In Cosmetics". In: *Applied Microbiology And Biotechnology* 97.11, pp. 4691–4700. ISSN: 0175-7598. DOI: 10.1007/S00253-013-4858-1.

Morita, T., H. Koike, Y. Koyama, H. Hagiwara, E. Ito, T. Fukuoka, T. Imura, M. Machida, and D. Kitamoto (2013c). "Genome Sequence Of The Basidiomycetous Yeast Pseudozyma Antarctica T-34, A Producer Of The Glycolipid Biosurfactants Mannosylerythritol Lipids". In: *Genome Announc* 1.2, E0006413. ISSN: 2169-8287 (Electronic). DOI: 10.1128/Genomea.00064-13.

Morita, T., M. Konishi, T. Fukuoka, T. Imura, and D. Kitamoto (2006). "Discovery Of Pseudozyma Rugulosa Nbrc 10877 As A Novel Producer Of The Glycolipid Biosurfactants, Mannosylerythritol Lipids, Based On Rdna Sequence". In: *Applied Microbiology*

And Biotechnology 73.2, pp. 305–313. ISSN: 0175-7598. DOI: 10.1007/S00253-006-0466-7.

Morita, T., M. Konishi, T. Fukuoka, T. Imura, and D. Kitamoto (2007a). "Physiological Differences In The Formation Of The Glycolipid Biosurfactants, Mannosylerythritol Lipids, Between Pseudozyma Antarctica And Pseudozyma Aphidis". In: *Applied Microbiology And Biotechnology* 74.2, pp. 307–315. ISSN: 0175-7598. DOI: 10.1007/S00253-006-0672-3.

— (2008a). "Identification Of Ustilago Cynodontis As A New Producer Of Glycolipid Biosurfactants, Mannosylerythritol Lipids, Based On Ribosomal Dna Sequences". In: *Journal Of Oleo Science* 57.10, pp. 549–556. ISSN: 1345-8957. DOI: 10.5650/Jos.57.549.

Morita, T., M. Konishi, T. Fukuoka, T. Imura, S. Yamamoto, M. Kitagawa, A. Sogabe, and D. Kitamoto (2008b). "Identification Of Pseudozyma Graminicola Cbs 10092 As A Producer Of Glycolipid Biosurfactants, Mannosylerythritol Lipids". In: *Journal Of Oleo Science* 57.2, pp. 123–131. ISSN: 1345-8957. DOI: 10.5650/Jos.57.123.

Morita, T., M. Takashima, T. Fukuoka, M. Konishi, T. Imura, and D. Kitamoto (2010). "Isolation Of Basidiomycetous Yeast Pseudozyma Tsukubaensis And Production Of Glycolipid Biosurfactant, A Diastereomer Type Of Mannosylerythritol Lipid-B". In: *Applied Microbiology And Biotechnology* 88.3, pp. 679–688. ISSN: 0175-7598. DOI: 10.1007/S00253-010-2762-5.

Morita, Tomotake, Masaaki Konishi, Tokuma Fukuoka, Tomohiro Imura, Hiroko K. Kitamoto, and Dai Kitamoto (2007b). "Characterization Of The Genus Pseudozyma By The Formation Of Glycolipid Biosurfactants, Mannosylerythritol Lipids". In: *Fems Yeast Research* 7.2, pp. 286–292. ISSN: 1567-1356. DOI: 10.1111/J.1567-1364.2006.00154.X.

Morita, Tomotake, Masaaki Konishi, Tokuma Fukuoka, Tomohiro Imura, Hideki Sakai, and Dai Kitamoto (2008c). "Efficient Production Of Di- And Tri-Acylated Mannosylerythritol Lipids As Glycolipid Biosurfactants By Pseudozyma Parantarctica Jcm 11752(T)". In: *Journal Of Oleo Science* 57.10, pp. 557–565. ISSN: 1345-8957. DOI: 10.5650/Jos.57.557.

Morita, Tomotake, Yuki Ogura, Masako Takashima, Naoto Hirose, Tokuma Fukuoka, Tomohiro Imura, Yukishige Kondo, and Dai Kitamoto (2011). "Isolation Of Pseudozyma Churashimaensis Sp Nov., A Novel Ustilaginomycetous Yeast Species As A Producer Of Glycolipid Biosurfactants, Mannosylerythritol Lipids". In: *Journal Of Bioscience And Bioengineering* 112.2, pp. 137–144. ISSN: 1389-1723. DOI: 10.1016/J.Jbiosc.2011.04.008.

Navarro-Muñoz, Jorge C. et al. (2018). "A Computational Framework For Systematic Exploration Of Biosynthetic Diversity From Large-Scale Genomic Data". In: *Biorxiv*. DOI:

10.1101/445270. eprint: Https://Www.Biorxiv.Org/Content/Early/2018/10/17/445270.Full.Pdf.

Nelson, D. R. (2018). "Cytochrome P450 Diversity In The Tree Of Life". In: *Biochimica Et Biophysica Acta-Proteins And Proteomics* 1866.1, pp. 141–154. ISSN: 1570-9639. DOI: 10.1016/J.Bbapap.2017.05.003.

Nützmann, Hans-Wilhelm, Ancheng Huang, and Anne Osbourn (2016). "Plant Metabolic Clusters – From Genetics To Genomics". In: *New Phytologist* 211.3, pp. 771–789. DOI: 10.1111/Nph.13981. eprint: Https://Nph.Onlinelibrary.Wiley.Com/Doi/Pdf/10.1111/Nph.13981.

Ochi, K. and T. Hosaka (2013). "New Strategies For Drug Discovery: Activation Of Silent Or Weakly Expressed Microbial Gene Clusters". In: *Appl Microbiol Biotechnol* 97.1, pp. 87–98. ISSN: 1432-0614 (Electronic) 0175-7598 (Linking). DOI: 10.1007/S00253-012-4551-9.

Oliveira, J. V. D., T. A. Borges, R. A. C. Dos Santos, L. F. D. Freitas, C. A. Rosa, G. H. Goldman, and D. M. Riano-Pachon (2014). "Pseudozyma Brasiliensis Sp Nov., A Xylanolytic, Ustilaginomycetous Yeast Species Isolated From An Insect Pest Of Sugarcane Roots". In: *International Journal Of Systematic And Evolutionary Microbiology* 64, pp. 2159–2168. ISSN: 1466-5026. DOI: 10.1099/Ijs.0.060103-0.

Osbourn, A. (2010). "Secondary Metabolic Gene Clusters: Evolutionary Toolkits For Chemical Innovation". In: *Trends In Genetics* 26.10, pp. 449–457. ISSN: 0168-9525. DOI: 10.1016/J.Tig.2010.07.001.

Patron, N. J., R. F. Waller, A. J. Cozijnsen, D. C. Straney, D. M. Gardiner, W. C. Nierman, and B. J. Howlett (2007). "Origin And Distribution Of Epipolythiodioxopiperazine (Etp) Gene Clusters In Filamentous Ascomycetes". In: *Bmc Evolutionary Biology* 7. ISSN: 1471-2148. DOI: 10.1186/1471-2148-7-174.

Proctor, R. H., S. P. Mccormick, N. J. Alexander, and A. E. Desjardins (2009). "Evidence That A Secondary Metabolic Biosynthetic Gene Cluster Has Grown By Gene Relocation During Evolution Of The Filamentous Fungus Fusarium". In: *Molecular Microbiology* 74.5, pp. 1128–1142. ISSN: 0950-382x. DOI: 10.1111/J.1365-2958.2009.06927.X.

Puchkov, E. O., U. Zahringer, B. Lindner, T. V. Kulakovskaya, U. Seydel, and A. Wiese (2002). "The Mycocidal, Membrane-Active Complex Of Cryptococcus Humicola Is A New Type Of Cellobiose Lipid With Detergent Features". In: *Biochimica Et Biophysica Acta-Biomembranes* 1558.2, pp. 161–170. ISSN: 0005-2736. DOI: 10.1016/S0005-2736(01)00428-X.

Rau, U., L. A. Nguyen, H. Roeper, H. Koch, and S. Lang (2005). "Fed-Batch Bioreactor Production Of Mannosylerythritol Lipids Secreted By Pseudozyma Aphidis". In: *Appl*

Microbiol Biotechnol 68.5, pp. 607–13. ISSN: 0175-7598 (Print) 0175-7598 (Linking). DOI: 10.1007/S00253-005-1906-5.

Reddy, B. V. B., A. Milshteyn, Z. Charlop-Powers, and S. F. Brady (2014). "Esnapd: A Versatile, Web-Based Bioinformatics Platform For Surveying And Mining Natural Product Biosynthetic Diversity From Metagenomes". In: *Chemistry & Biology* 21.8, pp. 1023–1033. ISSN: 1074-5521. DOI: 10.1016/J.Chembiol.2014.06.007.

Rodrigues, L., I. M. Banat, J. Teixeira, and R. Oliveira (2006). "Biosurfactants: Potential Applications In Medicine". In: *Journal Of Antimicrobial Chemotherapy* 57.4, pp. 609–618. ISSN: 0305-7453. DOI: 10.1093/Jac/Dkl024.

Roelants, S. L., S. L. De Maeseneire, K. Ciesielska, I. N. Van Bogaert, and W. Soetaert (2014). "Biosurfactant Gene Clusters In Eukaryotes: Regulation And Biotechnological Potential". In: *Appl Microbiol Biotechnol* 98.8, pp. 3449–61. ISSN: 1432-0614 (Electronic) 0175-7598 (Linking). DOI: 10.1007/S00253-014-5547-4.

Ronquist, Fredrik, Maxim Teslenko, Paul Van Der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Hohna, Bret Larget, Liang Liu, Marc A. Suchard, and John P. Huelsenbeck (2012). "Mrbayes 3.2: Efficient Bayesian Phylogenetic Inference And Model Choice Across A Large Model Space". In: *Systematic Biology* 61.3, pp. 539–542. ISSN: 1063-5157. DOI: 10.1093/Sysbio/Sys029.

Roze, L. V., M. Laivenieks, S. Y. Hong, J. Wee, S. S. Wong, B. Vanos, D. Awad, K. C. Ehrlich, and J. E. Linz (2015). "Aflatoxin Biosynthesis Is A Novel Source Of Reactive Oxygen Species-A Potential Redox Signal To Initiate Resistance To Oxidative Stress?" In: *Toxins* 7.5, pp. 1411–1430. ISSN: 2072-6651. DOI: 10.3390/Toxins7051411.

Saika, A., H. Koike, T. Fukuoka, S. Yamamoto, T. Kishimoto, and T. Morita (2016). "A Gene Cluster For Biosynthesis Of Mannosylerythritol Lipids Consisted Of 4-O-Beta-D-Mannopyranosyl-(2r,3s)-Erythritol As The Sugar Moiety In A Basidiomycetous Yeast Pseudozyma Tsukubaensis". In: *Plos One* 11.6, E0157858. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/Journal.Pone.0157858.

Saika, A., H. Koike, T. Hori, T. Fukuoka, S. Sato, H. Habe, D. Kitamoto, and T. Morita (2014). "Draft Genome Sequence Of The Yeast Pseudozyma Antarctica Type Strain Jcm10317, A Producer Of The Glycolipid Biosurfactants, Mannosylerythritol Lipids". In: *Genome Announc* 2.5. ISSN: 2169-8287 (Electronic). DOI: 10.1128/Genomea.00878-14.

Sayers, E. W., M. Cavanaugh, K. Clark, J. Ostell, K. D. Pruitt, and I. Karsch-Mizrachi (2019). "Genbank". In: *Nucleic Acids Research* 47.D1, pp. D94–D99. ISSN: 0305-1048. DOI: 10.1093/Nar/Gky989.

Scott, M. J. and M. N. Jones (2000). "The Biodegradation Of Surfactants In The Environment". In: *Biochimica Et Biophysica Acta-Biomembranes* 1508.1-2, pp. 235–251. ISSN: 0005-2736. DOI: 10.1016/S0304-4157(00)00013-7.

Sélem-Mojica, Nelly, César Aguilar, Karina Gutiérrez-García, Christian E. Martínez-Guerrero, and Francisco Barona-Gómez (2018). "Evomining Reveals The Origin And Fate Of Natural Products Biosynthetic Enzymes". In: *Biorxiv*. DOI: 10.1101/482273. eprint: Https://Www.Biorxiv.Org/Content/Early/2018/11/29/482273.Full.Pdf.

Shaaban, M., J. M. Palmer, W. A. El-Naggar, M. A. El-Sokkary, E. E. Habib, and N. P. Keller (2010). "Involvement Of Transposon-Like Elements In Penicillin Gene Cluster Regulation". In: *Fungal Genetics And Biology* 47.5, pp. 423–432. ISSN: 1087-1845. DOI: 10.1016/J.Fgb.2010.02.006.

Shimura, K. et al. (2007). "Identification Of A Biosynthetic Gene Cluster In Rice For Momilactones". In: *Journal Of Biological Chemistry* 282.47, pp. 34013–34018. ISSN: 0021-9258. DOI: 10.1074/Jbc.M703344200.

Shin, J., J. E. Kim, Y. W. Lee, and H. Son (2018). "Fungal Cytochrome P450s And The P450 Complement (Cypome) Of Fusarium Graminearum". In: *Toxins* 10.3. ISSN: 2072-6651. DOI: 10.3390/Toxins10030112.

Shirley, M. D., Z. Ma, B.S. Pedersen, and Wheelan S. J. (2015). "Efficient "Pythonic" Access To Fasta Files Using Pyfaidx." In: *Peerj Preprints* 3:E970v1. DOI: 10.7287/Peerj.Preprints.970v1.

Simao, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov (2015). "Busco: Assessing Genome Assembly And Annotation Completeness With Single-Copy Orthologs". In: *Bioinformatics* 31.19, pp. 3210–3212. ISSN: 1367-4803. DOI: 10.1093/Bioinformatics/Btv351.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J.M. Jones, and I. Birol (Feb. 2009). "Abyss: A Parallel Assembler For Short Read Sequence Data". In: *Genome Research* 19.6, pp. 1117–1123. DOI: 10.1101/Gr.089532.108.

Slot, J. C. and D. S. Hibbett (2007). "Horizontal Transfer Of A Nitrate Assimilation Gene Cluster And Ecological Transitions In Fungi: A Phylogenetic Study". In: *Plos One* 2.10. ISSN: 1932-6203. DOI: 10.1371/Journal.Pone.0001097.

Slot, J. C. and A. Rokas (2010). "Multiple Gal Pathway Gene Clusters Evolved Independently And By Different Mechanisms In Fungi". In: *Proceedings Of The National Academy Of Sciences Of The United States Of America* 107.22, pp. 10136–10141. ISSN: 0027-8424. DOI: 10.1073/Pnas.0914418107.

Stamatakis, Alexandros (2014). "Raxml Version 8: A Tool For Phylogenetic Analysis And Post-Analysis Of Large Phylogenies". In: *Bioinformatics* 30.9, pp. 1312–1313. ISSN: 1367-4803. DOI: 10.1093/Bioinformatics/Btu033.

Stanke, Mario, Mark Diekhans, Robert Baertsch, and David Haussler (2008). "Using Native And Syntenically Mapped Cdna Alignments To Improve De Novo Gene Finding". In: *Bioinformatics* 24.5, pp. 637–644. ISSN: 1367-4803. DOI: 10.1093/Bioinformatics/Btn013.

Starcevic, A., J. Zucko, J. Simunkovic, P. F. Long, J. Cullum, and D. Hranueli (2008). "Clustscan: An Integrated Program Package For The Semi-Automatic Annotation Of Modular Biosynthetic Gene Clusters And In Silico Prediction Of Novel Chemical Structures". In: *Nucleic Acids Research* 36.21, pp. 6882–6892. ISSN: 0305-1048. DOI: 10.1093/Nar/Gkn685.

Takeda, I., M. Umemura, H. Koike, K. Asai, and M. Machida (2014). "Motif-Independent Prediction Of A Secondary Metabolism Gene Cluster Using Comparative Genomics: Application To Sequenced Genomes Of Aspergillus And Ten Other Filamentous Fungal Species". In: *Dna Research* 21.4, pp. 447–457. ISSN: 1340-2838. DOI: 10.1093/Dnares/Dsu010.

Teichmann, B., C. Labbe, F. Lefebvre, M. Bolker, U. Linne, and R. R. Belanger (2011). "Identification Of A Biosynthesis Gene Cluster For Flocculosin A Cellobiose Lipid Produced By The Biocontrol Agent Pseudozyma Flocculosa". In: *Mol Microbiol* 79.6, pp. 1483–95. ISSN: 1365-2958 (Electronic) 0950-382x (Linking). DOI: 10.1111/J.1365-2958.2010.07533.X.

Teichmann, B., U. Linne, S. Hewald, M. A. Marahiel, and M. Bolker (2007). "A Biosynthetic Gene Cluster For A Secreted Cellobiose Lipid With Antifungal Activity From Ustilago Maydis". In: *Mol Microbiol* 66.2, pp. 525–33. ISSN: 0950-382x (Print) 0950-382x (Linking). DOI: 10.1111/J.1365-2958.2007.05941.X.

Van Bogaert, I. N., K. Holvoet, S. L. Roelants, B. Li, Y. C. Lin, Y. Van De Peer, and W. Soetaert (2013). "The Biosynthetic Gene Cluster For Sophorolipids: A Biotechnological Interesting Biosurfactant Produced By Starmerella Bombicola". In: *Mol Microbiol* 88.3, pp. 501–9. ISSN: 1365-2958 (Electronic) 0950-382x (Linking). DOI: 10.1111/Mmi.12200.

Van Bogaert, I. N. A., J. X. Zhang, and W. Soetaert (2011). "Microbial Synthesis Of Sophorolipids". In: *Process Biochemistry* 46.4, pp. 821–833. ISSN: 1359-5113. DOI: 10.1016/J.Procbio.2011.01.010.

Walton, J. D. (2000). "Horizontal Gene Transfer And The Evolution Of Secondary Metabolite Gene Clusters In Fungi: An Hypothesis". In: *Fungal Genetics And Biology* 30.3, pp. 167–171. ISSN: 1087-1845. DOI: 10.1006/Fgbi.2000.1224.

Wang, Q. M., D. Begerow, M. Groenewald, X. Z. Liu, B. Theelen, F. Y. Bai, and T. Boekhout (2015). "Multigene Phylogeny And Taxonomic Revision Of Yeasts And Related Fungi In

The Ustilaginomycotina". In: *Studies In Mycology* 81, pp. 55–83. ISSN: 01660616. DOI: 10.1016/J.Simyco.2015.10.004.

Waterhouse, R. M., M. Seppey, F. A. Simao, M. Manni, P. Ioannidis, G. Klioutchnikov, E. V. Kriventseva, and E. M. Zdobnov (2018). "Busco Applications From Quality Assessments To Gene Prediction And Phylogenomics". In: *Molecular Biology And Evolution* 35.3, pp. 543–548. ISSN: 0737-4038. DOI: 10.1093/Molbev/Msx319.

Wawrik, Boris, Lee Kerkhof, Gerben J. Zylstra, and Jerome J. Kukor (2005). "Identification Of Unique Type Ii Polyketide Synthase Genes In Soil". In: *Applied And Environmental Microbiology* 71.5, pp. 2232–2238. ISSN: 0099-2240. DOI: 10.1128/Aem.71.5.2232-2238.2005. eprint: Https://Aem.Asm.Org/Content/71/5/2232.Full.Pdf.

Weber, T., C. Rausch, P. Lopez, I. Hoof, V. Gaykova, D. H. Huson, and W. Wohlleben (2009). "Clusean: A Computer-Based Framework For The Automated Analysis Of Bacterial Secondary Metabolite Biosynthetic Gene Clusters". In: *Journal Of Biotechnology* 140.1-2, pp. 13–17. ISSN: 0168-1656. DOI: 10.1016/J.Jbiotec.2009.01.007.

Weber, T. et al. (2015). "Antismash 3.0-A Comprehensive Resource For The Genome Mining Of Biosynthetic Gene Clusters". In: *Nucleic Acids Research* 43.W1, W237–W243. ISSN: 0305-1048. DOI: 10.1093/Nar/Gkv437.

Wong, S. and K. H. Wolfe (2005). "Birth Of A Metabolic Gene Cluster In Yeast By Adaptive Gene Relocation". In: *Nat Genet* 37.7, pp. 777–82. ISSN: 1061-4036 (Print) 1061-4036 (Linking). DOI: 10.1038/Ng1584.

Wood, D. E. and S. L. Salzberg (2014). "Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments". In: *Genome Biology* 15.3. ISSN: 1465-6906. DOI: 10.1186/Gb-2014-15-3-R46.

Yamamoto, S., T. Fukuoka, T. Imura, T. Morita, S. Yanagidani, D. Kitamoto, and M. Kitagawa (2013). "Production Of A Novel Mannosylerythritol Lipid Containing A Hydroxy Fatty Acid From Castor Oil By Pseudozyma Tsukubaensis". In: *Journal Of Oleo Science* 62.6, pp. 381–389. ISSN: 1345-8957. DOI: 10.5650/Jos.62.381.

Yu, Mingda, Zhifeng Liu, Guangming Zeng, Hua Zhong, Yang Liu, Yongbing Jiang, Min Li, Xiaoxiao He, and Yan He (2015). "Characteristics Of Mannosylerythritol Lipids And Their Environmental Potential". In: *Carbohydrate Research* 407, pp. 63–72. ISSN: 0008-6215. DOI: 10.1016/J.Carres.2014.12.012.

Yu, N. et al. (2016). "Delineation Of Metabolic Gene Clusters In Plant Genomes By Chromatin Signatures". In: *Nucleic Acids Res.* ISSN: 1362-4962 (Electronic) 0305-1048 (Linking). DOI: 10.1093/Nar/Gkw100.

Zaas, A. K., M. Boyce, W. Schell, B. A. Lodge, J. L. Miller, and J. R. Perfect (2003). "Risk Of Fungemia Due To Rhodotorula And Antifungal Susceptibility Testing Of Rhodotorula

Isolates". In: *Journal Of Clinical Microbiology* 41.11, pp. 5233–5235. ISSN: 0095-1137. DOI: 10.1128/Jcm.41.11.5233-5235.2003.

Ziemert, N., S. Podell, K. Penn, J. H. Badger, E. Allen, and P. R. Jensen (2012). "The Natural Product Domain Seeker Napdos: A Phylogeny Based Bioinformatic Tool To Classify Secondary Metabolite Gene Diversity". In: *Plos One* 7.3. ISSN: 1932-6203. DOI: 10.1371/Journal.Pone.0034064.

A Appendix A

Plate	Provider	Machine	Library	Insert size (bp)	Read length (bp)
1	TGAC	Illumina HiSeq	TruSeq	500	2 x 100
2	TGAC	Illumina HiSeq	TruSeq	475	2 x 125
3	TGAC	Illumina HiSeq	TruSeq	475	2 x 125
3B	TGAC	Illumina HiSeq	LITE	430	2 x 250
4	Eurofins	Illumina HiSeq 2500	TruSeq	300	2 x 125
5	Eurofins	Illumina HiSeq 2500	TruSeq	300	2 x 125
6	Eurofins	Illumina HiSeq 2500	TruSeq	300	2 x 125
7	EI	Illumina HiSeq	LITE	430	2 x 250
8	EI	Illumina HiSeq	LITE	430	2 x 250
9	EI	Illumina HiSeq	LITE	430	2 x 250
10	Eurofins	Illumina HiSeq 2500	TruSeq	300	2 x 100
11	WTSI	Illumina X10	NEB Ultra	450	2 x 150

Table A.1: **NCYC yeast genome sequencing project structure.** Yeast genomes were sequenced in eleven batches of 96 strains (in 96-well plate format). Sequencing providers were either TGAC (The Genome Analysis Centre, Norwich, UK; now EI), Eurofins (Eurofins Genomics, Germany), EI (The Earlham Institute, Norwich, UK) or WTSI (Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK).

NCYC strain	Species	Pass 1	Pass 2	Pass 3
1	*Debaryomyces hansenii*	Plate 1		
2	*Dekkera anomala*	Plate 1		
4	*Candida tropicalis*	Plate 1		
6	*Kluyveromyces marxianus*	Plate 1		
8	*Debaryomyces hansenii*	Plate 1		
9	*Debaryomyces hansenii*	Plate 1		
10	*Debaryomyces hansenii*	Plate 1		
16	*Wickerhamomyces anomalus*	Plate 1		
17	*Hanseniaspora valbyensis*	Plate 4		
17A	*Hanseniaspora valbyensis*	Plate 1		
17B	*Hanseniaspora valbyensis*	Plate 1		
18	*Wickerhamomyces anomalus*	Plate 1		
20	*Wickerhamomyces anomalus*	Plate 1		
21	*Pichia membranifaciens*	Plate 1		
22	*Lindnera saturnus*	Plate 1		
23	*Lindnera saturnus*	Plate 1		
26	*Hanseniaspora osmophila*	Plate 1		
31	*Hanseniaspora osmophila*	Plate 1		
36	*Hanseniaspora vineae*	Plate 1		
39	*Candida catenulata*	Plate 1		
40	*Saccharomycopsis selenospora*	Plate 1		
43	*Pichia kudriavzevii*	Plate 1		
44	*Pichia membranifaciens*	Plate 1		
45	*Pichia kudriavzevii*	Plate 1		
46	*Nadsonia fulvescens var. fulvescens*	Plate 1		
47	*Galactomyces candidus like sp.*	Plate 10		
49	*Galactomyces candidus*	Plate 1		
50	*Galactomyces candidus*	Plate 10		
51	*Pichia membranifaciens*	Plate 1		
52	*Pichia membranifaciens*	Plate 1		
54	*Pichia membranifaciens*	Plate 1		
55	*Issatchenkia orientalis*	Plate 1		
56	*Schwanniomyces polymorphus*	Plate 1		
57	*Lindnera saturnus*	Plate 1		
58	*Hanseniaspora osmophila*	Plate 1		
59	*Rhodotorula glutinis var. glutinis*	Plate 1		
60	*Rhodotorula glutinis var. glutinis*	Plate 1		
61	*Rhodosporidium kratochvilovae*	Plate 1		
62	*Rhodotorula minuta var. minuta*	Plate 1	Plate 2	
63	*Rhodotorula mucilaginosa*	Plate 1		
64	*Rhodotorula mucilaginosa*	Plate 1		

65	*Rhodotorula mucilaginosa*	Plate 1		
68	*Rhodotorula mucilaginosa*	Plate 1		
70	*Saccharomyces cerevisiae*	Plate 1		
71	*Debaryomyces fabryi*	Plate 1		
72	*Saccharomyces cerevisiae*	Plate 1		
73	*Saccharomyces pastorianus*	Plate 1		
74	*Saccharomyces cerevisiae*	Plate 1		
75	*Saccharomyces cerevisiae*	Plate 1		
76	*Saccharomyces cerevisiae*	Plate 1		
77	*Saccharomyces cerevisiae*	Plate 1		
78	*Saccharomyces cerevisiae*	Plate 1		
79	*Saccharomyces cerevisiae*	Plate 1		
80	*Saccharomyces cerevisiae*	Plate 1		
81	*Saccharomyces cerevisiae*	Plate 1		
82	*Saccharomyces cerevisiae*	Plate 1		
83	*Saccharomyces cerevisiae*	Plate 1		
84	*Saccharomyces cerevisiae*	Plate 1	Plate 10	
85	*Saccharomyces cerevisiae*	Plate 1		
86	*Saccharomyces cerevisiae*	Plate 1		
87	*Saccharomyces cerevisiae*	Plate 1	Plate 10	
88	*Saccharomyces cerevisiae*	Plate 1		
89	*Saccharomyces cerevisiae*	Plate 1		
90	*Saccharomyces cerevisiae*	Plate 1		
91	*Saccharomyces cerevisiae*	Plate 1	Plate 10	
92	*Saccharomyces cerevisiae*	Plate 1		
93	*Saccharomyces cerevisiae*	Plate 1		
94	*Saccharomyces cerevisiae*	Plate 1		
95	*Saccharomyces cerevisiae*	Plate 1		
96	*Saccharomyces cerevisiae*	Plate 1		
97	*Saccharomyces cerevisiae*	Plate 1		
98	*Pichia kudriavzevii*	Plate 10		
99	*Saccharomyces pastorianus*	Plate 3	Plate 3B	Plate 11
100	*Kluyveromyces marxianus*	Plate 5		
101	*Pichia membranifaciens*	Plate 10		
102	*Debaryomyces hansenii*	Plate 10		
103	*Debaryomyces hansenii*	Plate 10		
104	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
105	*Pichia fermentans*	Plate 10		
106	*Saccharomyces bayanus*	Plate 10		
107	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
108	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
109	*Saccharomyces kudriavzevii*	Plate 3	Plate 3B	
110	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
111	*Kluyveromyces marxianus*	Plate 5		

112	*Saccharomyces pastorianus*	Plate 10	
113	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
114	*Saccharomyces bayanus*	Plate 10	
115	*Saccharomyces pastorianus*	Plate 10	
116	*Pichia membranifaciens*	Plate 10	
117	*Pichia membranifaciens*	Plate 10	
118	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
119	*Wickerhamomyces subpelliculosus*	Plate 8	
120	*Pichia kudriavzevii*	Plate 8	
121	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
122	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
124	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
125	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
126	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
127	*Saccharomycopsis capsularis*	Plate 10	
128	*Zygosaccharomyces bailii*	Plate 6	
129	*Wickerhamomyces anomalus*	Plate 10	
130	*Candida rugosa*	Plate 10	
131	*Schizosaccharomyces octosporus*	Plate 10	
132	*Schizosaccharomyces pombe*	Plate 10	
133	*Schwanniomyces occidentalis*	Plate 10	
135	*Rhodotorula mucilaginosa*	Plate 2	
137	*Kazachstania exigua*	Plate 10	
138	*Rhodotorula aurantiaca*	Plate 2	
140	*Torulaspora delbrueckii*	Plate 2	Plate 3B
141	*Torulaspora delbrueckii*	Plate 2	
142	*Rhodotorula mucilaginosa*	Plate 2	
143	*Kluyveromyces marxianus*	Plate 5	
144	*Meyerozyma guilliermondii*	Plate 10	
145	*Meyerozyma guilliermondii*	Plate 10	
147	*Naumovozyma castellii*	Plate 2	Plate 3B
148	*Candida gropengiesseri*	Plate 11	
151	*Kluyveromyces marxianus*	Plate 5	
152	*Kluyveromyces marxianus*	Plate 5	
153	*Yarrowia lipolytica*	Plate 10	
154	*Rhodotorula glutinis var. glutinis*	Plate 2	
155	*Rhodotorula glutinis var. glutinis*	Plate 2	
158	*Rhodotorula mucilaginosa*	Plate 2	
159	*Rhodotorula mucilaginosa*	Plate 2	
161	*Torulaspora delbrueckii*	Plate 2	Plate 3B
162	*Rhodotorula glutinis var. glutinis*	Plate 2	
163	*Saccharomyces cerevisiae*	Plate 10	
166	*Metschnikowia pulcherrima*	Plate 2	
167	*Saccharomyces cerevisiae*	Plate 3	Plate 3B

168	*Lindnera jadinii*	Plate 10		
169	*Pichia membranifaciens*	Plate 10		
171	*Zygosaccharomyces bisporus*	Plate 6		
176	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
177	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
179	*Kluyveromyces marxianus*	Plate 5		
181	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
182	*Saccharomyces cerevisiae*	Plate 3B	Plate 3	
183	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
185	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
186	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
187	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
188	*Kluyveromyces marxianus*	Plate 5		
190	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
191	*Wickerhamomyces subpelliculosus*	Plate 8		
192	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
195	*Rhodotorula mucilaginosa*	Plate 2	Plate 3B	
196	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
197	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
198	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
199	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
200	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
201	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
202	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
205	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
206	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
207	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
208	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	Plate 11
209	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
210	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
211	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
212	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
213	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
214	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	Plate 11
215	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
216	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
217	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
218	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
219	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
220	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
221	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
222	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
223	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
224	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	

225	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
226	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
227	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
228	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
229	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
230	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
231	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
232	*Saccharomyces cerevisiae*	Plate 4	
233	*Saccharomyces cerevisiae*	Plate 8	
234	*Saccharomyces cerevisiae*	Plate 7	
235	*Saccharomyces cerevisiae*	Plate 4	
236	*Saccharomyces cerevisiae*	Plate 8	
238	*Saccharomyces cerevisiae*	Plate 5	
239	*Saccharomyces cerevisiae*	Plate 8	
240	*Saccharomyces cerevisiae*	Plate 11	
241	*Saccharomyces cerevisiae*	Plate 5	
243	*Kluyveromyces marxianus*	Plate 5	
244	*Kluyveromyces marxianus*	Plate 5	
324	*Saccharomyces pastorianus*	Plate 8	
329	*Pichia kudriavzevii*	Plate 8	
332	*Pichia kudriavzevii*	Plate 8	
336	*Pichia kudriavzevii*	Plate 8	
337	*Pichia kudriavzevii*	Plate 8	
338	*Pichia kudriavzevii*	Plate 8	
341	*Saccharomyces cerevisiae*	Plate 5	
343	*Saccharomyces cerevisiae*	Plate 8	
347	*Pichia kudriavzevii*	Plate 8	
350	*Candida glabrata*	Plate 5	
353	*Saccharomyces cerevisiae*	Plate 8	
356	*Saccharomyces cerevisiae*	Plate 5	
357	*Saccharomyces cerevisiae*	Plate 5	
358	*Saccharomyces cerevisiae*	Plate 5	
360	*Saccharomyces cerevisiae*	Plate 4	
361	*Saccharomyces cerevisiae*	Plate 4	
363	*Saccharomyces cerevisiae*	Plate 8	
367	*Saccharomyces cerevisiae*	Plate 8	
368	*Saccharomyces cerevisiae*	Plate 8	
371	*Metschnikowia pulcherrima*	Plate 2	
372	*Metschnikowia pulcherrima*	Plate 2	
373	*Metschnikowia pulcherrima*	Plate 2	
377	*Rhodotorula glutinis var. glutinis*	Plate 2	
385	*Zygosaccharomyces bailii*	Plate 6	
388	*Candida glabrata*	Plate 1	
392	*Saccharomyces pastorianus*	Plate 4	

400	*Saccharomyces cerevisiae*	Plate 8	
401	*Saccharomyces cerevisiae*	Plate 8	
407	*Candida etchellsii*	Plate 7	
408	*Torulaspora delbrueckii*	Plate 2	
411	*Torulaspora microellipsoides*	Plate 2	Plate 3B
416	*Kluyveromyces lactis*	Plate 1	
417	*Zygosaccharomyces bailii*	Plate 6	
426	*Kluyveromyces marxianus*	Plate 5	
430	*Saccharomyces cerevisiae*	Plate 5	
431	*Saccharomyces cerevisiae*	Plate 4	
436	*Wickerhamomyces subpelliculosus*	Plate 8	
437	*Wickerhamomyces subpelliculosus*	Plate 8	
438	*Wickerhamomyces subpelliculosus*	Plate 8	
442	*Wickerhamomyces subpelliculosus*	Plate 8	
444	*Trichosporon cutaneum var. cutaneum*	Plate 7	
461	*Cryptococcus humicola*	Plate 11	
463	*Saccharomyces cerevisiae*	Plate 5	
464	*Zygosaccharomyces bailii*	Plate 6	
469	*Kluyveromyces lactis*	Plate 6	Plate 10
472	*Trichosporon ovoides*	Plate 11	
476	*Cryptococcus curvatus*	Plate 7	
478	*Saccharomyces cerevisiae*	Plate 5	
479	*Saccharomyces cerevisiae*	Plate 5	
482	*Saccharomyces cerevisiae*	Plate 5	
486	*Candida stellata*	Plate 11	
489	*Saccharomyces cerevisiae*	Plate 5	
490	*Saccharomyces cerevisiae*	Plate 5	
491	*Saccharomyces cerevisiae*	Plate 5	
492	*Torulaspora delbrueckii*	Plate 2	Plate 3B
502	*Rhodotorula graminis*	Plate 2	
505	*Saccharomyces cerevisiae*	Plate 1	
523	*Vanderwaltozyma polyspora*	Plate 1	
524	*Torulaspora pretoriensis*	Plate 2	
525	*Saccharomyces cerevisiae*	Plate 5	
538	*Kluyveromyces dobzhanskii*	Plate 4	
539	*Rhodotorula minuta var. minuta*	Plate 2	Plate 3B
541	*Rhodotorula minuta var. minuta*	Plate 2	
543	*Lachancea kluyveri*	Plate 4	
546	*Kluyveromyces wickerhamii*	Plate 4	
548	*Kluyveromyces lactis*	Plate 6	
551	*Kluyveromyces lactis*	Plate 6	
553	*Cryptococcus peneaus*	Plate 11	
558	*Wickerhamomyces subpelliculosus*	Plate 8	
559	*Torulaspora delbrueckii*	Plate 2	

563	*Zygosaccharomyces bailii*	Plate 6	
566	*Torulaspora delbrueckii*	Plate 2	Plate 3B
568	*Zygosaccharomyces rouxii*	Plate 1	
570	*Kluyveromyces lactis*	Plate 6	
571	*Kluyveromyces lactis*	Plate 6	
573	*Zygosaccharomyces bailii*	Plate 6	
575	*Kluyveromyces lactis*	Plate 6	
576	*Debaryomyces hansenii*	Plate 7	
578	*Cryptococcus macerans*	Plate 11	
580	*Zygosaccharomyces bailii*	Plate 6	
582	*Torulaspora delbrueckii*	Plate 2	
585	*Torulaspora delbrueckii*	Plate 2	
587	*Kluyveromyces marxianus*	Plate 5	
591	*Cryptococcus luteolus*	Plate 11	
597	*Candida albicans*	Plate 7	Plate 9
601	*Candida parapsilosis*	Plate 7	
608	*Torulaspora delbrueckii*	Plate 2	
609	*Saccharomyces cerevisiae*	Plate 6	
610	*Candida albicans*	Plate 7	
611	*Debaryomyces hansenii*	Plate 11	
619	*Saccharomyces cerevisiae*	Plate 5	
620	*Saccharomyces cerevisiae*	Plate 5	
621	*Saccharomyces cerevisiae*	Plate 5	
622	*Saccharomyces cerevisiae*	Plate 8	
667	*Saccharomyces cerevisiae*	Plate 4	
671	*Saccharomyces cerevisiae*	Plate 5	
672	*Saccharomyces cerevisiae*	Plate 5	
675	*Candida gropengiesseri*	Plate 11	
677	*Torulaspora delbrueckii*	Plate 2	
678	*Torulaspora delbrueckii*	Plate 2	
684	*Saccharomyces cerevisiae*	Plate 5	
689	*Candida gropengiesseri*	Plate 11	
694	*Saccharomyces cerevisiae*	Plate 5	
695	*Saccharomyces cerevisiae*	Plate 4	
696	*Torulaspora delbrueckii*	Plate 2	
730	*Saccharomycodes ludwigii*	Plate 10	
731	*Saccharomycodes ludwigii*	Plate 1	
732	*Saccharomycodes ludwigii*	Plate 10	
733	*Saccharomycodes ludwigii/ Kluyveromyces marxianus*	Plate 10	
734	*Saccharomycodes ludwigii*	Plate 10	
739	*Saccharomyces cerevisiae*	Plate 4	
742	*Candida magnoliae*	Plate 11	
744	*Kluyveromyces marxianus*	Plate 5	
745	*Metschnikowia reukaufii*	Plate 2	

746	*Metschnikowia reukaufii*	Plate 2
747	*Metschnikowia pulcherrima*	Plate 2
752	*Kluyveromyces lactis*	Plate 6
754	*Saccharomyces cerevisiae*	Plate 4
758	*Rhodotorula mucilaginosa*	Plate 2
759	*Rhodotorula mucilaginosa*	Plate 2
764	*Candida magnoliae*	Plate 11
765	*Candida magnoliae*	Plate 11
768	*Nakaseomyes delphensis*	Plate 1
776	*Kluyveromyces lactis*	Plate 6
777	*Naumovozyma dairenensis*	Plate 1
783	*Metschnikowia zobellii*	Plate 2
784	*Cryptococcus marinus*	Plate 11
794	*Metschnikowia zobellii*	Plate 2
796	*Rhodotorula mucilaginosa*	Plate 2
797	*Rhodotorula mucilaginosa*	Plate 2
807	*Saccharomyces cerevisiae*	Plate 6
814	*Kazachstania exigua*	Plate 4
816	*Saccharomyces cerevisiae*	Plate 5
820	*Torulaspora globosa*	Plate 2
826	*Saccharomyces cerevisiae*	Plate 4
827	*Kluyveromyces marxianus*	Plate 5
844	*Rhodotorula minuta var. minuta*	Plate 2
845	*Rhodotorula minuta var. minuta*	Plate 2
849	*Saccharomycodes ludwigii*	Plate 10
851	*Kluyveromyces marxianus*	Plate 5
854	*Candida albicans*	Plate 7
872	*Pichia kudriavzevii*	Plate 8
894	*Metschnikowia lunata*	Plate 2
906	*Kluyveromyces marxianus*	Plate 5
911	*Pseudozyma aphidis*	Plate 5
925	*Yarrowia lipolytica*	Plate 7
929	*Kluyveromyces lactis*	Plate 6
930	*Rhodotorula minuta var. minuta*	Plate 2
931	*Rhodotorula minuta var. minuta*	Plate 2
935	*Saccharomyces cerevisiae*	Plate 6
951	*Wickerhamomyces onychis*	Plate 7
956	*Saccharomyces cerevisiae*	Plate 4
963	*Saccharomyces cerevisiae*	Plate 8
970	*Kluyveromyces marxianus*	Plate 5
971	*Kazachstania unispora*	Plate 4
974	*Rhodotorula glutinis var. glutinis*	Plate 2
975	*Saccharomyces pastorianus*	Plate 4
993	*Pichia kudriavzevii*	Plate 8

995	*Saccharomyces cerevisiae*	Plate 5	
996	*Saccharomyces cerevisiae*	Plate 5	
1001	*Saccharomyces cerevisiae*	Plate 8	
1004	*Saccharomyces cerevisiae*	Plate 8	
1006	*Saccharomyces cerevisiae*	Plate 1	
1007	*Saccharomyces cerevisiae*	Plate 8	
1010	*Saccharomyces cerevisiae*	Plate 8	
1013	*Saccharomyces cerevisiae*	Plate 11	
1016	*Saccharomyces cerevisiae*	Plate 11	
1017	*Saccharomyces cerevisiae*	Plate 9	
1020	*Saccharomyces cerevisiae*	Plate 10	
1023	*Saccharomyces cerevisiae*	Plate 9	
1026	*Saccharomyces cerevisiae*	Plate 1	Plate 10
1030	*Saccharomyces cerevisiae*	Plate 9	
1031	*Saccharomyces cerevisiae*	Plate 11	
1033	*Saccharomyces cerevisiae*	Plate 9	
1035	*Saccharomyces cerevisiae*	Plate 11	
1037	*Saccharomyces cerevisiae*	Plate 9	
1039	*Saccharomyces cerevisiae*	Plate 11	
1040	*Saccharomyces cerevisiae*	Plate 10	
1044	*Saccharomyces cerevisiae*	Plate 9	
1046	*Saccharomyces cerevisiae*	Plate 11	
1049	*Saccharomyces cerevisiae*	Plate 10	
1052	*Saccharomyces cerevisiae*	Plate 9	
1053	*Saccharomyces cerevisiae*	Plate 11	
1054	*Saccharomyces cerevisiae*	Plate 11	
1055	*Saccharomyces cerevisiae*	Plate 9	
1060	*Saccharomyces cerevisiae*	Plate 11	
1063	*Saccharomyces cerevisiae*	Plate 4	
1064	*Saccharomyces cerevisiae*	Plate 4	
1066	*Saccharomyces cerevisiae*	Plate 10	
1069	*Saccharomyces cerevisiae*	Plate 10	
1072	*Saccharomyces cerevisiae*	Plate 9	
1073	*Saccharomyces pastorianus*	Plate 11	
1076	*Saccharomyces cerevisiae*	Plate 10	
1079	*Saccharomyces cerevisiae*	Plate 9	Plate 10
1082	*Saccharomyces cerevisiae*	Plate 10	
1085	*Saccharomyces cerevisiae*	Plate 9	
1089	*Saccharomyces cerevisiae*	Plate 10	
1090	*Saccharomyces cerevisiae*	Plate 11	
1093	*Saccharomyces cerevisiae*	Plate 10	
1097	*Saccharomyces cerevisiae*	Plate 9	
1102	*Saccharomyces cerevisiae*	Plate 9	
1103	*Saccharomyces cerevisiae*	Plate 11	

1106	*Saccharomyces cerevisiae*	Plate 9	
1111	*Saccharomyces cerevisiae*	Plate 9	
1114	*Saccharomyces cerevisiae*	Plate 10	
1117	*Saccharomyces cerevisiae*	Plate 11	
1118	*Saccharomyces cerevisiae*	Plate 9	
1122	*Saccharomyces cerevisiae*	Plate 10	
1124	*Saccharomyces cerevisiae*	Plate 11	
1126	*Saccharomyces cerevisiae*	Plate 10	
1129	*Saccharomyces cerevisiae*	Plate 10	
1132	*Saccharomyces cerevisiae*	Plate 9	
1134	*Saccharomyces cerevisiae*	Plate 11	
1138	*Saccharomyces cerevisiae*	Plate 10	
1139	*Saccharomyces cerevisiae*	Plate 11	
1141	*Saccharomyces cerevisiae*	Plate 11	
1147	*Saccharomyces cerevisiae*	Plate 9	
1151	*Saccharomyces cerevisiae*	Plate 4	
1156	*Saccharomyces cerevisiae*	Plate 9	
1159	*Saccharomyces cerevisiae*	Plate 9	
1161	*Saccharomyces cerevisiae*	Plate 11	
1163	*Saccharomyces cerevisiae*	Plate 9	
1167	*Saccharomyces cerevisiae*	Plate 9	
1171	*Saccharomyces cerevisiae*	Plate 11	
1175	*Saccharomyces cerevisiae*	Plate 10	
1179	*Saccharomyces cerevisiae*	Plate 9	
1183	*Saccharomyces cerevisiae*	Plate 9	
1186	*Saccharomyces cerevisiae*	Plate 10	
1187	*Saccharomyces cerevisiae*	Plate 1	
1190	*Saccharomyces cerevisiae*	Plate 9	
1199	*Saccharomyces cerevisiae*	Plate 9	
1203	*Saccharomyces cerevisiae*	Plate 11	
1210	*Saccharomyces cerevisiae*	Plate 11	
1211	*Saccharomyces cerevisiae*	Plate 9	
1215	*Saccharomyces cerevisiae*	Plate 9	Plate 10
1218	*Saccharomyces cerevisiae*	Plate 9	
1221	*Saccharomyces cerevisiae*	Plate 9	
1228	*Saccharomyces cerevisiae*	Plate 1	
1235	*Saccharomyces cerevisiae*	Plate 11	
1239	*Saccharomyces pastorianus*	Plate 11	
1240	*Saccharomyces cerevisiae*	Plate 9	
1243	*Saccharomyces cerevisiae*	Plate 9	
1245	*Saccharomyces cerevisiae*	Plate 1	
1260	*Saccharomyces cerevisiae*	Plate 9	
1264	*Saccharomyces cerevisiae*	Plate 9	
1270	*Saccharomyces cerevisiae*	Plate 9	

1274	*Saccharomyces cerevisiae*	Plate 9
1277	*Saccharomyces cerevisiae*	Plate 9
1280	*Saccharomyces cerevisiae*	Plate 9
1283	*Saccharomyces cerevisiae*	Plate 9
1286	*Saccharomyces cerevisiae*	Plate 9
1289	*Saccharomyces cerevisiae*	Plate 9
1292	*Saccharomyces cerevisiae*	Plate 9
1298	*Saccharomyces cerevisiae*	Plate 9
1308	*Saccharomyces cerevisiae*	Plate 9
1311	*Saccharomyces cerevisiae*	Plate 9
1314	*Saccharomyces cerevisiae*	Plate 9
1315	*Saccharomyces cerevisiae*	Plate 6
1318	*Saccharomyces cerevisiae*	Plate 9
1321	*Saccharomyces cerevisiae*	Plate 9
1337	*Saccharomyces cerevisiae*	Plate 4
1339	*Saccharomyces cerevisiae*	Plate 9
1363	*Candida albicans*	Plate 7
1368	*Kluyveromyces lactis*	Plate 6
1369	*Candida catenulata*	Plate 7
1384	*Pseudozyma fusiformata*	Plate 5
1388	*Cryptococcus albidus var. aerius*	Plate 11
1389	*Cryptococcus laurentii var. laurentii*	Plate 11
1393	*Candida tropicalis*	Plate 7
1398	*Pichia kudriavzevii*	Plate 8
1400	*Zygosaccharomyces bailii*	Plate 6
1401	*Rhodotorula graminis*	Plate 2
1405	*Cryptococcus vishniacii var. vishniacii*	Plate 11
1406	*Saccharomyces cerevisiae*	Plate 5
1407	*Saccharomyces cerevisiae*	Plate 5
1408	*Saccharomyces cerevisiae*	Plate 5
1409	*Saccharomyces cerevisiae*	Plate 5
1410	*Saccharomyces cerevisiae*	Plate 5
1411	*Saccharomyces cerevisiae*	Plate 5
1412	*Saccharomyces cerevisiae*	Plate 5
1413	*Saccharomyces cerevisiae*	Plate 5
1414	*Saccharomyces cerevisiae*	Plate 5
1415	*Saccharomyces cerevisiae*	Plate 5
1416	*Zygosaccharomyces bailii*	Plate 4
1417	*Kazachstania lodderae*	Plate 4
1424	*Kluyveromyces marxianus*	Plate 5
1425	*Kluyveromyces marxianus*	Plate 5
1426	*Kluyveromyces marxianus*	Plate 5
1427	*Zygosaccharomyces bailii*	Plate 6
1429	*Kluyveromyces marxianus*	Plate 5

1431	*Saccharomyces cerevisiae*	Plate 5	
1441	*Kluyveromyces marxianus*	Plate 5	
1444	*Saccharomyces cerevisiae*	Plate 4	
1449	*Starmerella bombicola*	Plate 4	
1462	*Cryptococcus laurentii var. laurentii*	Plate 11	
1466	*Candida albicans*	Plate 7	
1467	*Candida albicans*	Plate 7	
1468	*Candida albicans*	Plate 7	
1469	*Candida albicans*	Plate 7	
1470	*Candida albicans*	Plate 7	
1471	*Candida albicans*	Plate 7	
1472	*Candida albicans*	Plate 7	
1473	*Candida albicans*	Plate 7	
1474	*Naumovozyma dairenensis*	Plate 6	
1477	*Naumovozyma dairenensis*	Plate 6	
1495	*Zygosaccharomyces bisporus*	Plate 4	
1496	*Zygosaccharomyces bisporus*	Plate 6	
1497	*Zygosaccharomyces bisporus*	Plate 6	
1498	*Zygosaccharomyces bisporus*	Plate 6	
1510	*Pseudozyma tsukubaensis*	Plate 5	Plate 10
1515	*Zygosaccharomyces bisporus*	Plate 6	
1520	*Zygosaccharomyces bailii*	Plate 6	Plate 10
1521	*Zygosaccharomyces bailii*	Plate 6	
1529	*Saccharomyces cerevisiae*	Plate 5	
1530	*Saccharomyces cerevisiae*	Plate 5	Plate 10
1536	*Cryptococcus heveanensis*	Plate 11	
1548	*Kluyveromyces lactis*	Plate 6	
1553	*Zygosaccharomyces bailii*	Plate 6	
1554	*Zygosaccharomyces bailii*	Plate 6	
1555	*Zygosaccharomyces bisporus*	Plate 6	
1556	*Zygosaccharomyces bailii*	Plate 6	
1557	*Zygosaccharomyces bailii*	Plate 6	
1558	*Zygosaccharomyces bailii*	Plate 6	
1563	*Nematospora coryli*	Plate 10	
1572	*Zygosaccharomyces bailii*	Plate 6	
1573	*Zygosaccharomyces bailii*	Plate 6	
1591	*Zygosaccharomyces bailii*	Plate 6	
1592	*Zygosaccharomyces bailii*	Plate 6	
1603	*Saccharomyces cerevisiae*	Plate 4	
1606	*Saccharomyces cerevisiae*	Plate 4	
1645	*Rhodotorula mucilaginosa*	Plate 2	
1646	*Rhodotorula mucilaginosa*	Plate 2	
1647	*Rhodotorula mucilaginosa*	Plate 2	
1648	*Rhodotorula mucilaginosa*	Plate 2	

1649	*Rhodotorula mucilaginosa*	Plate 2	
1650	*Rhodotorula mucilaginosa*	Plate 2	
1651	*Rhodotorula mucilaginosa*	Plate 2	
1653	*Wickerhamomyces subpelliculosus*	Plate 8	
1656	*Naumovozyma dairenensis*	Plate 6	
1659	*Pichia fermentans*	Plate 2	
1660	*Rhodotorula mucilaginosa*	Plate 2	
1673	*Saccharomyces cerevisiae*	Plate 6	
1681	*Saccharomyces cerevisiae*	Plate 1	
1765	*Saccharomyces cerevisiae*	Plate 5	
1766	*Zygosaccharomyces bailii*	Plate 6	
2258	*Kluyveromyces lactis*	Plate 6	
2265	*Kluyveromyces marxianus*	Plate 5	
2307	*Kazachstania servazzii*	Plate 6	
2321	*Metschnikowia pulcherrima*	Plate 2	
2322	*Metschnikowia pulcherrima*	Plate 2	
2395	*Metschnikowia hawaiiensis*	Plate 2	Plate 3B
2396	*Metschnikowia hawaiiensis*	Plate 2	
2397	*Saccharomyces cerevisiae*	Plate 4	
2400	*Trichosporon gracile*	Plate 11	
2401	*Saccharomyces cerevisiae*	Plate 5	
2402	*Saccharomyces cerevisiae*	Plate 5	
2403	*Zygosaccharomyces mellis*	Plate 4	
2418	*Cryptococcus gastricus*	Plate 11	
2423	*Candida tropicalis*	Plate 7	
2424	*Candida apis var. galacta*	Plate 11	
2431	*Candida bombi*	Plate 11	
2432	*Candida etchellsii*	Plate 7	
2433	*Lachancea thermotolerans*	Plate 1	
2435	*Candida etchellsii*	Plate 10	
2439	*Rhodotorula glutinis*	Plate 2	
2440	*Rhodotorula glutinis*	Plate 2	
2442	*Cryptococcus terreus*	Plate 11	
2449	*Kazachstania telluris*	Plate 4	
2450	*Candida humilis*	Plate 4	
2457	*Candida tropicalis*	Plate 7	
2458	*Candida tropicalis*	Plate 7	
2461	*Cryptococcus magnus*	Plate 11	
2471	*Galactomyces geotrichum*	Plate 10	
2473	*Torulaspora delbrueckii*	Plate 2	Plate 3B
2474	*Trichosporon mucoides*	Plate 10	
2479	*Trichosporon jirovecii*	Plate 7	
2480	*Metschnikowia agaves*	Plate 2	
2483	*Kazachstania piceae*	Plate 4	

2484	*Trichosporon moniliiforme*	Plate 11
2486	*Metschnikowia agaves*	Plate 2
2489	*Zygotorulaspora mrakii*	Plate 1
2491	*Metschnikowia gruessii*	Plate 2
2492	*Metschnikowia gruessii*	Plate 2
2499	*Galactomyces geotrichum*	Plate 8
2508	*Lachancea fermentati*	Plate 4
2509	*Trichosporon cutaneum*	Plate 11
2510	*Trichosporon dulcitum*	Plate 11
2513	*Zygotorulaspora florentina*	Plate 4
2515	*Trichosporon inkin*	Plate 7
2516	*Trichosporon cutaneum*	Plate 10
2517	*Saccharomyces cerevisiae*	Plate 5
2521	*Metschnikowia bicuspidata*	Plate 2
2529	*Metschnikowia bicuspidata*	Plate 2
2537	*Candida gropengiesseri*	Plate 11
2542	*Candida norvegica*	Plate 11
2547	*Candida apicola*	Plate 11
2558	*Cryptococcus amylolentus*	Plate 11
2559	*Kluyveromyces dobzhanskii*	Plate 5
2560	*Kazachstania sinensis*	Plate 4
2568	*Torulaspora microellipsoides*	Plate 2
2569	*Candida magnoliae*	Plate 11
2572	*Debaryomyces hansenii var. hansenii*	Plate 1
2577	*Kazachstania servazzii*	Plate 1
2578	*Saccharomyces bayanus*	Plate 1
2579	*Clavispora lusitaniae*	Plate 7
2580	*Metschnikowia pulcherrima*	Plate 2
2581	*Rhodotorula minuta var. minuta*	Plate 2
2582	*Candida albicans*	Plate 7
2587	*Saccharomyces cerevisiae*	Plate 5
2592	*Saccharomyces cerevisiae*	Plate 4
2597	*Kluyveromyces marxianus var. marxianus*	Plate 5
2599	*Rhodotorula mucilaginosa*	Plate 2
2600	*Saccharomyces paradoxus*	Plate 1
2602	*Galactomyces geotrichum*	Plate 10
2605	*Rhodotorula vanillica*	Plate 2
2614	*Trichosporon loubieri*	Plate 11
2620	*Candida magnoliae*	Plate 11
2623	*Filobasidium uniguttulatum*	Plate 11
2628	*Guehomyces pullulans*	Plate 10
2629	*Torulaspora delbrueckii*	Plate 1
2635	*Trichosporon aquatile*	Plate 11
2636	*Candida valdiviana*	Plate 11

2637	*Cryptococcus podzolicus*	Plate 11	
2638	*Trichosporon montevideense*	Plate 11	
2644	*Lachancea waltii*	Plate 4	
2658	*Pichia kudriavzevii*	Plate 8	
2665	*Trichosporon domesticum*	Plate 11	
2666	*Rhodotorula glutinis var. glutinis*	Plate 2	
2670	*Candida dublinensis*	Plate 7	
2675	*Kluyveromyces marxianus*	Plate 5	
2677	*Trichosporon asahii*	Plate 10	
2683	*Saccharomyces cerevisiae*	Plate 5	
2688	*Saccharomyces cerevisiae*	Plate 5	
2693	*Kazachstania servazzii*	Plate 6	
2695	*Saccharomyces cerevisiae*	Plate 5	
2698	*Pichia kudriavzevii*	Plate 8	
2701	*Kazachstania viticola*	Plate 1	
2702	*Kazachstania kunashirensis*	Plate 4	
2703	*Kazachstania martiniae*	Plate 4	
2712	*Cryptococcus laurentii*	Plate 11	
2726	*Candida rugosa*	Plate 7	
2729	*Kazachstania africana*	Plate 4	
2733	*Saccharomyces cerevisiae*	Plate 4	
2737	*Saccharomyces cerevisiae*	Plate 4	
2739	*Hanseniaspora uvarum*	Plate 4	
2741	*Torulaspora delbrueckii*	Plate 2	
2742	*Kluyveromyces lactis*	Plate 6	
2745	*Candida pararugosa*	Plate 7	
2746	*Candida zeylanoides*	Plate 7	
2748	*Trichosporon inkin*	Plate 7	
2752	*Rhodotorula cresolica*	Plate 3	
2753	*Metschnikowia zobellii*	Plate 2	
2754	*Vanderwaltozyma yarrowii*	Plate 4	
2775	*Kazachstania servazzii*	Plate 6	
2776	*Saccharomyces cerevisiae*	Plate 7	
2777	*Saccharomyces cerevisiae*	Plate 7	
2778	*Saccharomyces cerevisiae*	Plate 7	
2779	*Saccharomyces cerevisiae*	Plate 7	
2780	*Saccharomyces cerevisiae*	Plate 7	
2786	*Candida aaseri*	Plate 7	
2789	*Zygosaccharomyces lentus*	Plate 4	
2790	*Zygosaccharomyces bailii*	Plate 6	
2791	*Kluyveromyces marxianus*	Plate 1	
2797	*Kluyveromyces lactis*	Plate 6	
2798	*Saccharomyces cerevisiae*	Plate 7	
2804	*Saccharomyces bayanus var. uvarum*	Plate 1	Plate 5

2808	*Saccharomyces cerevisiae*	Plate 5	
2809	*Saccharomyces cerevisiae*	Plate 5	
2826	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
2827	*Kazachstania rosinii*	Plate 4	
2833	*Pichia norvegensis*	Plate 7	
2853	*Pichia kudriavzevii*	Plate 8	
2855	*Saccharomyces cerevisiae*	Plate 5	
2864	*Rhodotorula mucilaginosa*	Plate 3	
2866	*Sporopachydermia lactativora*	Plate 7	
2873	*Rhodotorula creatinovora*	Plate 3	
2875	*Lachancea cidri*	Plate 4	
2878	*Kazachstania barnettii*	Plate 4	
2885	*Torulaspora delbrueckii*	Plate 2	
2886	*Kluyveromyces marxianus*	Plate 5	
2887	*Kluyveromyces marxianus*	Plate 5	
2888	*Saccharomyces mikatae*	Plate 1	
2889	*Saccharomyces kudriavzevii*	Plate 1	
2890	*Saccharomyces cariocanus*	Plate 1	
2897	*Zygosaccharomyces kombuchaensis*	Plate 6	
2898	*Naumovozyma castellii*	Plate 6	
2899	*Candida geochares*	Plate 11	
2900	*Candida batistae*	Plate 11	
2903	*Trichosporon asahii*	Plate 11	
2904	*Yarrowia lipolytica*	Plate 1	Plate 4
2907	*Kluyveromyces marxianus*	Plate 5	
2908	*Starmerella bombicola*	Plate 4	
2913	*Saccharomyces cerevisiae*	Plate 7	
2925	*Saccharomyces cerevisiae*	Plate 11	
2927	*Zygosaccharomyces bailii*	Plate 6	
2931	*Zygosaccharomyces bailii*	Plate 6	
2932	*Zygosaccharomyces bailii*	Plate 6	
2933	*Zygosaccharomyces bailii*	Plate 6	
2934	*Zygosaccharomyces bailii*	Plate 6	
2935	*Zygosaccharomyces bisporus*	Plate 6	
2938	*Candida sorbosivorans*	Plate 11	
2945	*Saccharomyces cerevisiae*	Plate 4	
2947	*Saccharomyces cerevisiae*	Plate 5	
2948	*Saccharomyces cerevisiae*	Plate 5	
2956	*Kluyveromyces lactis*	Plate 6	
2963	*Cryptococcus flavescens*	Plate 11	
2965	*Saccharomyces cerevisiae*	Plate 7	
2966	*Saccharomyces cerevisiae*	Plate 7	
2967	*Saccharomyces cerevisiae*	Plate 7	
2972	*Rhodotorula graminis*	Plate 3	

2974	*Saccharomyces cerevisiae*	Plate 7	
2976	*Hanseniaspora osmophila*	Plate 4	
2979	*Cryptococcus gastricus*	Plate 7	
2980	*Kluyveromyces lactis var. lactis*	Plate 6	
2981	*Kluyveromyces lactis var. drosophilarum*	Plate 6	
2990	*Cryptococcus albidus var. albidus*	Plate 11	
2991	*Kazachstania spencerorum*	Plate 4	
2995	*Zygosaccharomyces bailii*	Plate 6	
2999	*Zygosaccharomyces kombuchaensis*	Plate 4	
3000	*Zygosaccharomyces kombuchaensis*	Plate 6	
3001	*Zygosaccharomyces kombuchaensis*	Plate 6	
3008	*Cryptococcus saitoi*	Plate 11	
3013	*Candida davenportii*	Plate 11	
3020	*Saccharomyces cerevisiae*	Plate 5	
3021	*Saccharomyces cerevisiae*	Plate 5	
3022	*Saccharomyces cerevisiae*	Plate 5	
3024	*Torulaspora microellipsoides*	Plate 2	Plate 3B
3025	*Saccharomyces cerevisiae*	Plate 7	
3026	*Saccharomyces cerevisiae*	Plate 7	
3027	*Saccharomyces cerevisiae*	Plate 7	
3028	*Saccharomyces cerevisiae*	Plate 7	
3029	*Saccharomyces cerevisiae*	Plate 7	
3030	*Saccharomyces cerevisiae*	Plate 7	
3031	*Saccharomyces cerevisiae*	Plate 7	
3032	*Saccharomyces cerevisiae*	Plate 7	
3033	*Saccharomyces cerevisiae*	Plate 7	
3034	*Saccharomyces cerevisiae*	Plate 6	
3035	*Saccharomyces cerevisiae*	Plate 7	
3036	*Saccharomyces cerevisiae*	Plate 7	
3037	*Saccharomyces cerevisiae*	Plate 7	
3038	*Saccharomyces cerevisiae*	Plate 7	
3039	*Saccharomyces cerevisiae*	Plate 7	
3041	*Kluyveromyces lactis*	Plate 6	
3047	*Metschnikowia pulcherrima*	Plate 2	
3048	*Saccharomyces cerevisiae*	Plate 7	
3051	*Saccharomyces cerevisiae*	Plate 7	
3052	*Saccharomyces cerevisiae*	Plate 7	
3053	*Kazachstania servazzii*	Plate 6	
3056	*Rhodotorula sp. nov.*	Plate 3	
3057	*Rhodotorula mucilaginosa*	Plate 3	
3072	*Rhodotorula laryngis*	Plate 3	
3076	*Saccharomyces cerevisiae*	Plate 7	
3077	*Saccharomyces cerevisiae*	Plate 7	
3078	*Saccharomyces cerevisiae*	Plate 7	

3080	*Saccharomyces cerevisiae*	Plate 7	
3086	*Cryptococcus longus*	Plate 11	
3087	*Cryptococcus gilvescens*	Plate 11	
3090	*Zygosaccharomyces bailii*	Plate 6	
3091	*Zygosaccharomyces bailii*	Plate 6	
3092	*Zygosaccharomyces bailii*	Plate 6	
3096	*Metschnikowia fructicola*	Plate 2	
3104	*Candida pseudointermedia*	Plate 5	
3108	*Naumovozyma castellii*	Plate 4	
3109	*Cryptococcus victoriae*	Plate 11	
3114	*Saccharomyces cerevisiae*	Plate 7	
3115	*Candida albicans*	Plate 7	
3120	*Rhodotorula phylloplana*	Plate 3	
3121	*Saccharomyces cerevisiae*	Plate 7	
3122	*Saccharomyces cerevisiae*	Plate 7	
3123	*Saccharomyces cerevisiae*	Plate 7	
3124	*Saccharomyces cerevisiae*	Plate 7	
3125	*Saccharomyces cerevisiae*	Plate 7	
3126	*Saccharomyces cerevisiae*	Plate 7	
3127	*Saccharomyces cerevisiae*	Plate 7	
3129	*Metschnikowia sp.*	Plate 2	
3133	*Candida bracarensis*	Plate 7	
3134	*Zygosaccharomyces bisporus*	Plate 6	
3138	*Geotrichum vulgare*	Plate 10	
3141	*Torulaspora delbrueckii*	Plate 2	Plate 3B
3146	*Zygosaccharomyces bailii*	Plate 6	
3239	*Torulaspora delbrueckii*	Plate 2	
3242	*Pichia kudriavzevii*	Plate 8	
3252	*Trichosporon coremiiforme*	Plate 11	
3254	*Trichosporon jirovecii*	Plate 11	
3255	*Torulaspora delbrueckii*	Plate 2	
3256	*Pichia kudriavzevii*	Plate 8	
3264	*Saccharomyces cerevisiae*	Plate 4	
3265	*Saccharomyces cerevisiae*	Plate 4	
3266	*Saccharomyces cerevisiae*	Plate 4	
3267	*Pseudozyma sp.*	Plate 5	
3272	*Geotrichum vulgare*	Plate 10	
3284	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
3290	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
3298	*Zygosaccharomyces bisporus*	Plate 6	
3302	*Zygosaccharomyces bailii*	Plate 6	
3303	*Candida glabrata*	Plate 5	
3306	*Saccharomyces cerevisiae*	Plate 9	
3307	*Zygosaccharomyces bailii*	Plate 6	

3309	*Pichia kudriavzevii*	Plate 8	
3311	*Saccharomyces cerevisiae*	Plate 4	
3312	*Saccharomyces cerevisiae*	Plate 3	Plate 3B
3313	*Saccharomyces cerevisiae*	Plate 4	
3314	*Saccharomyces cerevisiae*	Plate 4	
3315	*Saccharomyces cerevisiae*	Plate 4	
3318	*Saccharomyces cerevisiae*	Plate 4	
3319	*Saccharomyces cerevisiae*	Plate 4	
3324	*Saccharomyces cerevisiae*	Plate 7	
3325	*Saccharomyces cerevisiae*	Plate 7	
3326	*Saccharomyces cerevisiae*	Plate 7	
3331	*Saccharomyces cerevisiae*	Plate 7	
3332	*Saccharomyces cerevisiae*	Plate 7	
3333	*Saccharomyces cerevisiae*	Plate 7	
3334	*Saccharomyces cerevisiae*	Plate 7	
3338	*Saccharomyces cerevisiae*	Plate 7	
3339	*Saccharomyces cerevisiae*	Plate 7	
3340	*Saccharomyces cerevisiae*	Plate 8	
3341	*Saccharomyces cerevisiae*	Plate 8	
3342	*Saccharomyces cerevisiae*	Plate 8	
3343	*Saccharomyces cerevisiae*	Plate 8	
3344	*Kluyveromyces marxianus*	Plate 5	
3345	*Saccharomycodes ludwigii*	Plate 10	
3354	*Kazachstania servazzii*	Plate 6	
3358	*Saccharomyces cerevisiae*	Plate 6	
3369	*Cryptococcus carnescens*	Plate 11	
3373	*Candida palmioleophila*	Plate 8	
3378	*Zygosaccharomyces bailii*	Plate 6	
3379	*Zygosaccharomyces bailii*	Plate 6	
3392	*Candida cf. glabrata*	Plate 8	
3393	*Pichia kudriavzevii*	Plate 8	
3396	*Kluyveromyces marxianus*	Plate 5	
3397	*Candida bracarensis*	Plate 8	
3398	*Metschnikowia aff. fructicola*	Plate 2	
3400	*Metschnikowia sp. nov.*	Plate 2	
3401	*Rhodotorula graminis*	Plate 3	
3402	*Debaryomyces hansenii*	Plate 8	
3403	*Saccharomyces cerevisiae*	Plate 8	
3406	*Saccharomyces cerevisiae*	Plate 4	
3407	*Zygosaccharomyces bailii*	Plate 6	
3410	*Zygosaccharomyces bailii*	Plate 6	
3411	*Rhodotorula mucilaginosa*	Plate 3	
3414	*Zygosaccharomyces bailii*	Plate 6	
3423	*Cryptococcus uzbekistanensis*	Plate 11	

3431	*Pseudozyma hubeiensis*	Plate 5		
3444	*Rhodotorula dairenensis*	Plate 3		
3445	*Saccharomyces cerevisiae*	Plate 4		
3447	*Saccharomyces cerevisiae*	Plate 4		
3448	*Saccharomyces cerevisiae*	Plate 4		
3449	*Saccharomyces cerevisiae*	Plate 4		
3451	*Saccharomyces cerevisiae*	Plate 4		
3452	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
3453	*Saccharomyces cerevisiae*	Plate 4	Plate 10	
3454	*Saccharomyces cerevisiae*	Plate 4		
3455	*Saccharomyces cerevisiae*	Plate 4		
3456	*Saccharomyces cerevisiae*	Plate 4		
3457	*Saccharomyces cerevisiae*	Plate 4		
3458	*Saccharomyces cerevisiae*	Plate 4		
3460	*Saccharomyces cerevisiae*	Plate 4		
3461	*Saccharomyces cerevisiae*	Plate 4		
3462	*Saccharomyces cerevisiae*	Plate 4		
3464	*Saccharomyces cerevisiae*	Plate 8		
3465	*Saccharomyces cerevisiae*	Plate 8		
3466	*Saccharomyces cerevisiae*	Plate 4		
3467	*Saccharomyces cerevisiae*	Plate 4		
3468	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	Plate 11
3469	*Saccharomyces cerevisiae*	Plate 4		
3470	*Saccharomyces cerevisiae*	Plate 4		
3471	*Saccharomyces cerevisiae*	Plate 4		
3472	*Saccharomyces cerevisiae*	Plate 4		
3486	*Saccharomyces cerevisiae*	Plate 4		
3487	*Saccharomyces cerevisiae*	Plate 4		
3491	*Saccharomyces cerevisiae*	Plate 8		
3492	*Candida parapsilosis*	Plate 8		
3493	*Saccharomyces cerevisiae*	Plate 8		
3497	*Saccharomyces cerevisiae*	Plate 8		
3498	*Saccharomyces cerevisiae*	Plate 8		
3499	*Saccharomyces cerevisiae*	Plate 8		
3500	*Saccharomyces cerevisiae*	Plate 8		
3501	*Candida parapsilosis*	Plate 8		
3502	*Candida glabrata*	Plate 5		
3504	*Rhodotorula mucilaginosa*	Plate 3		
3506	*Torulaspora delbrueckii*	Plate 2		
3510	*Saccharomyces cerevisiae*	Plate 8		
3511	*Saccharomyces cerevisiae*	Plate 8		
3512	*Saccharomyces cerevisiae*	Plate 8		
3513	*Saccharomyces cerevisiae*	Plate 8		
3514	*Saccharomyces cerevisiae*	Plate 8		

3515	*Saccharomyces cerevisiae*	Plate 8	
3516	*Saccharomyces cerevisiae*	Plate 8	
3519	*Candida glabrata*	Plate 5	
3520	*Candida albicans*	Plate 8	
3521	*Saccharomyces cerevisiae*	Plate 8	
3522	*Saccharomyces cerevisiae*	Plate 8	
3523	*Saccharomyces cerevisiae*	Plate 8	
3526	*Cryptococcus sp. nov.*	Plate 11	
3527	*Candida albicans*	Plate 8	
3528	*Saccharomyces cerevisiae*	Plate 8	
3529	*Saccharomyces cerevisiae*	Plate 8	
3530	*Cryptococcus albidus*	Plate 11	
3536	*Rhodotorula mucilaginosa*	Plate 3	Plate 3B
3537	*Candida glabrata*	Plate 5	
3539	*Cryptococcus sp. nov.*	Plate 11	
3541	*Cryptococcus shivajii*	Plate 11	
3546	*Saccharomyces cerevisiae*	Plate 9	
3549	*Saccharomyces cerevisiae*	Plate 9	
3552	*Saccharomyces cerevisiae*	Plate 9	
3557	*Saccharomyces cerevisiae*	Plate 6	
3562	*Saccharomyces cerevisiae*	Plate 6	
3612	*Saccharomyces cerevisiae*	Plate 4	
3630	*Saccharomyces cerevisiae*	Plate 4	
3662	*Saccharomyces paradoxus*	Plate 4	
3716	*Kazachstania servazzii*	Plate 6	
3719	*Metschnikowia chrysoperlae*	Plate 2	
3721	*Rhodotorula slooffiae*	Plate 3	Plate 3B
3722	*Rhodotorula graminis*	Plate 3	Plate 3B
3724	*Zygosaccharomyces bailii*	Plate 6	
3725	*Rhodotorula dairenensis*	Plate 3	Plate 3B
3730	*Cryptococcus wieringae*	Plate 11	
3734	*Trichosporon scarabaeorum*	Plate 11	
3735	*Rhodotorula mucilaginosa*	Plate 3	Plate 3B
3740	*Pichia kudriavzevii*	Plate 8	
3745	*Pichia kudriavzevii*	Plate 8	
3751	*Pichia kudriavzevii*	Plate 8	
3752	*Pichia kudriavzevii*	Plate 10	
3759	*Geotrichum fermentans*	Plate 10	
3770	*Cryptococcus albidosimilis*	Plate 11	
3772	*Rhodotorula mucilaginosa*	Plate 3	Plate 3B
3775	*Rhodotorula mucilaginosa*	Plate 3B	Plate 3
3776	*Naumovozyma castellii*	Plate 6	
3777	*Geotrichum/Galactomyces sp. nov.*	Plate 10	
3778	*Candida albicans*	Plate 8	

3779	*Candida albicans*	Plate 8	
3785	*Candida bombi*	Plate 11	
3788	*Hanseniaspora guilliermondii*	Plate 4	
3792	*Metschnikowia koreensis*	Plate 2	
3815	*Candida lactis-condensi*	Plate 11	
3816	*Rhodotorula mucilaginosa*	Plate 3	Plate 3B
3817	*Rhodotorula mucilaginosa*	Plate 3	Plate 3B
3820	*Rhodotorula mucilaginosa*	Plate 3	Plate 3B
3821	*Rhodotorula mucilaginosa*	Plate 3	Plate 3B
3832	*Rhodotorula sp. nov.*	Plate 3	Plate 3B
3833	*Rhodotorula sp. nov.*	Plate 3	Plate 3B
3834	*Rhodotorula laryngis*	Plate 3	Plate 3B
3835	*Rhodotorula sp. nov.*	Plate 3	Plate 3B
3836	*Rhodotorula laryngis*	Plate 3	Plate 3B
3837	*Rhodotorula laryngis*	Plate 3	Plate 3B
3838	*Rhodotorula laryngis*	Plate 3	Plate 3B
3853	*Kazachstania bulderi*	Plate 4	
3867	*Rhodotorula mucilaginosa*	Plate 3	Plate 3B
3870	*Metschnikowia koreensis*	Plate 2	
3872	*Rhodotorula mucilaginosa*	Plate 3	Plate 3B
3877	*Torulaspora franciscae*	Plate 2	
3879	*Geotrichum candidum*	Plate 8	
3961	*Kazachstania servazzii*	Plate 6	
3963	*Kazachstania servazzii*	Plate 6	
3964	*Kazachstania servazzii*	Plate 6	
3966	*Kazachstania servazzii*	Plate 6	
3968	*Kazachstania servazzii*	Plate 6	
3983	*Cryptococcus diffluens*	Plate 11	
3997	*Saccharomyces cerevisiae*	Plate 9	
4000	*Kazachstania yasuniensis*	Plate 4	
4001	*Issatchenkia orientalis*	Plate 8	
4002	*Issatchenkia orientalis*	Plate 8	
4015	*Pichia kudriavzevii*	Plate 8	
4017	*Pichia kudriavzevii*	Plate 8	
4020	*Torulaspora delbrueckii*	Plate 6	
4034	*Issatchenkia orientalis*	Plate 8	
4045	*Saccharomyces cerevisiae*	Plate 9	
4051	*Saccharomyces cerevisiae*	Plate 9	Plate 10
4063	*Saccharomyces cerevisiae*	Plate 9	
4068	*Saccharomyces cerevisiae*	Plate 9	
4081	*Saccharomyces cerevisiae*	Plate 9	
4144	*Candida albicans*	Plate 9	
4145	*Candida albicans*	Plate 9	
4146	*Candida albicans*	Plate 9	

Table A.2: Yeast strains whose genomes were sequenced within the NCYC yeast genome sequencing project. Some genomes were sequenced multiple times (maximum of three times), either for quality control purposes or where a sequencing failure had occurred. Blue shading denotes a sequencing failure, either at the sequencing library construction or sequencing run stages.

Strain	**Species**	**Sequencing plate**
CBS 10006	*Pseudozyma thailandica*	Plate 10
CBS 9959	*Pseudozyma crassa*	Plate 10
CBS 9961	*Pseudozyma alboarmeniaca*	Plate 10
JCM 16988	*Pseudozyma churashimaensis*	Plate 10
NRRL Y-17626	*Pseudozyma rugulosa*	Plate 10
NRRL Y-7792	*Pseudozyma tsukubaensis*	Plate 10

Table A.3: Sequenced strains from other international yeast collections.

B Appendix B

Strain	MaxCluster	NumGenes	Gene	Contig	Strand	Start	End	Length
Um521	13	13	ahd1	759001946	-	343089	344233	1144
			fat1	759001946	-	326581	328121	1540
			orf1	759001946	+	334453	335582	1129
			fhd1	759001946	+	336607	337418	811
			cyp2	759001946	-	307265	308749	1484
			orf2	759001946	-	341789	342136	347
			fas2	759001946	-	309909	320980	11071
			fgt1	759001946	-	338216	339945	1729
			rfl1	759001946	+	304317	306563	2246
			fat3	759001946	-	326630	328133	1503
			fat2	759001946	-	332365	333755	1390
			cyp1	759001946	-	329618	331406	1788
			atr1	759001946	+	322143	326281	4138
NCYC3431	12	13	ahd1	1052	-	36278	37413	1135
			fat1	1052	-	23627	25161	1534
			orf1	1052	+	30664	31734	1070
			fhd1	1052	+	32277	33090	813
			rfl1	1052	+	1218	3485	2267
			orf2	285	-	30026	30070	44
			fas2	1052	-	6992	18073	11081
			fgt1	1052	-	33507	35221	1714
			cyp2	1052	-	4609	6105	1496
			fat3	1052	-	23675	25083	1408
			fat2	1052	-	28587	29971	1384
			cyp1	1052	-	25808	27599	1791
			atr1	1052	+	19264	23372	4108
NCYC1384	11	13	ahd1	577	+	314598	315689	1091
			fat1	577	+	328847	330387	1540
			fhd1	577	-	318794	319606	812
			cyp2	577	+	347944	349420	1476
			orf1	538	-	344571	345488	917
			orf2	250	+	10145	10229	84
			fas2	577	+	335861	346887	11026
			fgt1	577	+	316676	318363	1687

			rfl1	577	-	349844	351935	2091
			fat3	577	+	328911	330370	1459
			fat2	577	+	324150	325525	1375
			cyp1	577	+	326455	328249	1794
			atr1	577	-	330625	334728	4103
NCYC489	4	13	ahd1	42379	+	1377	2468	1091
			fat1	78995	+	2106	3646	1540
			orf1	84515	+	3849	4887	1038
			orf2	68448	+	910	994	84
			cyp2	86097	+	3	1085	1082
			fhd1	64308	-	3	534	531
			fas2	85113	+	1911	5710	3799
			fgt1	62642	+	59	1746	1687
			rfl1	86097	-	1509	3599	2090
			fat3	78995	+	2170	3629	1459
			fat2	84515	-	1257	2632	1375
			cyp1	78995	+	1	1508	1507
			atr1	78995	-	3884	5161	1277
NCYC2808	12	12	ahd1	69483	-	40839	41995	1156
			fat1	69483	-	25805	27343	1538
			fhd1	69483	+	36337	37148	811
			rfl1	69483	+	2032	4299	2267
			orf1	69483	+	33200	34282	1082
			fas2	69483	-	8640	19692	11052
			fgt1	69483	-	38044	39771	1727
			cyp2	69483	-	5670	7177	1507
			fat3	69483	-	25853	27275	1422
			fat2	69483	-	30957	32356	1399
			cyp1	69483	-	28032	29823	1791
			atr1	69483	+	20992	25138	4146
NCYC3267	12	12	ahd1	6141	-	48248	49404	1156
			fat1	6141	-	33214	34752	1538
			fhd1	6141	+	43746	44557	811
			cyp2	6141	-	13079	14586	1507
			orf1	6141	+	40609	41691	1082
			fas2	6141	-	16049	27101	11052
			fgt1	6141	-	45453	47180	1727
			rfl1	6141	+	9441	11708	2267
			fat3	6141	-	33262	34684	1422
			fat2	6141	-	38366	39765	1399
			cyp1	6141	-	35441	37232	1791
			atr1	6141	+	28401	32547	4146
NCYC238	5	12	ahd1	15832	-	1691	2847	1156
			fat1	15380	-	17696	19234	1538

			fhd1	15127	+	1793	2604	811
			cyp2	15715	-	4583	6090	1507
			orf1	15368	-	286	1368	1082
			fas2	15380	-	531	11583	11052
			fgt1	15127	-	3500	4629	1129
			rfl1	15715	+	945	3212	2267
			fat3	15380	-	17744	19166	1422
			fat2	15368	+	2212	3611	1399
			cyp1	15380	-	19923	21714	1791
			atr1	15380	+	12883	17029	4146
NCYC1510	3	12	ahd1	109419	+	1558	2505	947
			fat1	104729	+	2967	4505	1538
			fhd1	77821	-	1583	2394	811
			rfl1	96932	+	3	405	402
			orf1	109479	+	87	1169	1082
			fas2	95195	+	1	4728	4727
			fgt1	38652	+	1288	2049	761
			cyp2	67580	-	19	884	865
			fat3	104729	+	3035	4457	1422
			fat2	98684	+	61	809	748
			cyp1	104729	+	487	2278	1791
			atr1	100913	-	2	2008	2006
NCYC694	2	12	ahd1	38337	+	642	1725	1083
			fat1	71506	+	156	1694	1538
			orf1	32261	-	2	810	808
			rfl1	21129	+	37	1145	1108
			fhd1	58820	+	408	1219	811
			fas2	70551	-	2	3747	3745
			fgt1	52012	-	3	1192	1189
			cyp2	2125	+	1	653	652
			fat3	71506	+	224	1646	1422
			fat2	29959	-	1	1103	1102
			cyp1	34975	-	2	1726	1724
			atr1	62370	-	1	1626	1625
NCYC2683	1	12	ahd1	31784	-	1	176	175
			fat1	7513	+	3	191	188
			fhd1	44348	+	761	1127	366
			cyp2	9408	-	1	173	172
			orf1	38008	-	6	151	145
			orf2	33694	+	5	88	83
			fas2	35318	+	1	951	950
			fgt1	13890	+	1	163	162
			rfl1	37125	-	4	131	127
			fat2	25678	-	19	164	145

			cyp1	37013	-	2	296	294
			atr1	32194	-	2	259	257
NCYC2903	7	11	ahd1	7646	+	93513	94437	924
			fas2	7549	-	897670	905065	7395
			fgt1	7646	-	79037	79457	420
			cyp2	7549	-	1084570	1085036	466
			fat3	7646	+	84222	84717	495
			fat1	7646	+	83506	84733	1227
			fat2	7646	+	82179	82479	300
			rfl1	7632	-	2308981	2309172	191
			cyp1	7646	-	91072	92624	1552
			orf1	7646	+	87996	88514	518
			atr1	7404	-	93856	96733	2877
NCYC3086	6	10	ahd1	2359	+	1347869	1348777	908
			fas2	2354	-	168831	176297	7466
			fgt1	2354	-	2687	3160	473
			cyp2	2359	-	1339147	1340669	1522
			fat3	2359	+	1341971	1342869	898
			fat1	2359	+	1341671	1343053	1382
			fat2	2359	-	1349184	1350472	1288
			cyp1	2359	-	1345398	1346946	1548
			orf1	2353	+	160904	161695	791
			atr1	2297	+	128696	132520	3824
NCYC3252	4	10	ahd1	4396	-	195249	195994	745
			fas2	4403	-	981571	988959	7388
			fgt1	4275	-	71721	72130	409
			cyp2	4408	+	163229	163332	103
			fat3	4275	+	76342	77189	847
			fat1	4275	+	75975	77193	1218
			fat2	4275	+	74893	75166	273
			cyp1	4208	+	271741	272235	494
			orf1	4384	-	58947	59120	173
			atr1	4388	+	902360	905237	2877
NCYC3502	4	10	ahd1	2765	+	22460	23205	745
			fas2	2394	-	23851	31234	7383
			fgt1	2930	+	13103	13512	409
			atr1	2769	-	19426	22303	2877
			fat3	2930	-	8044	8891	847
			fat1	2930	-	8040	9258	1218
			fat2	2930	-	10067	10340	273
			cyp1	2436	+	4071	4565	494
			orf1	2935	+	7696	7869	173
			cyp2	2830	-	37213	37316	103
NCYC2484	3	7	ahd1	8100	-	935052	935343	291

			fas2	8100	-	697594	705039	7445
			cyp2	8100	-	403921	404038	117
			cyp1	8099	-	200013	200570	557
			fhd1	8085	-	428171	428689	518
			atr1	8086	-	473493	476298	2805
			orf1	8075	-	186422	186580	158
NCYC3013	3	7	ahd1	2041	+	201481	201617	136
			fas2	2048	-	155147	159218	4071
			cyp2	2003	+	1260199	1260329	130
			cyp1	2049	-	224172	224328	156
			orf1	2048	+	523024	523155	131
			atr1	2048	+	207990	209542	1552
			fhd1	1968	+	930508	930981	473
NCYC3721	2	7	ahd1	1373	-	276468	276575	107
			fas2	1330	+	55076	58106	3030
			cyp2	1330	-	130424	130569	145
			cyp1	1295	+	321575	321739	164
			fhd1	1403	-	240335	240517	182
			atr1	1325	-	33632	34432	800
			orf1	1411	-	443752	443902	150
NCYC2966	2	7	ahd1	58673	-	21827	22433	606
			fas2	57258	-	245	7633	7388
			fgt1	57537	+	42396	42451	55
			atr1	58704	+	935	3802	2867
			cyp1	58430	+	34945	35090	145
			orf1	58525	-	5979	6491	512
			cyp2	58430	+	34070	34280	210
NCYC2515	2	7	ahd1	25298	-	21825	22431	606
			fas2	21317	-	279	7667	7388
			fgt1	25432	-	32706	32761	55
			cyp2	1087	-	22811	23021	210
			cyp1	1087	-	22001	22146	145
			orf1	9396	-	5984	6496	512
			atr1	23817	-	7445	10312	2867
NCYC472	2	7	ahd1	13206	+	85460	86244	784
			fas2	12926	-	135934	143496	7562
			atr1	12810	-	360415	363609	3194
			cyp2	13028	+	309504	309617	113
			cyp1	13028	+	308873	309618	745
			orf1	13134	-	283447	283916	469
			rfl1	12884	+	91330	91525	195
NCYC931	2	7	ahd1	41231	+	49105	49212	107
			fas2	40635	-	105164	108188	3024
			atr1	40172	+	75410	76210	800

			cyp1	12579	-	2416	2591	175
			orf1	41056	-	219530	219680	150
			cyp2	40635	+	32038	32183	145
			fhd1	40515	+	336454	336636	182
NCYC591	2	7	ahd1	7825	+	464656	464760	104
			fas2	7679	+	672267	679610	7343
			cyp2	7942	+	443682	443762	80
			cyp1	7895	+	211417	211490	73
			orf1	7829	+	32540	32998	458
			atr1	7825	+	324612	325946	1334
			fhd1	7536	+	34166	34854	688
NCYC2748	2	7	ahd1	47967	-	21846	22452	606
			fas2	37189	-	250	7635	7385
			fgt1	46620	-	5658	5713	55
			atr1	47676	-	17537	20402	2865
			cyp1	27878	-	4806	4951	145
			orf1	48108	+	21617	22129	512
			cyp2	27878	-	5616	5826	210
NCYC930	2	7	ahd1	1901	+	132057	132200	143
			fas2	1863	+	425237	427716	2479
			atr1	1880	+	705288	706636	1348
			cyp1	1890	-	429130	429293	163
			orf1	1890	+	41683	41800	117
			cyp2	1808	-	146340	146398	58
			fhd1	1764	+	63787	63865	78
NCYC2638	3	6	ahd1	4661	-	641803	642326	523
			fas2	4624	-	215527	222958	7431
			cyp2	4409	-	286404	286620	216
			cyp1	4661	+	887158	887572	414
			orf1	4661	-	890663	891255	592
			atr1	4437	-	11634	15141	3507
NCYC3833	3	6	fas2	14291	-	297221	299716	2495
			cyp2	14291	-	15386	15530	144
			cyp1	14423	-	931666	931927	261
			fhd1	14423	-	581511	581714	203
			atr1	14405	-	164763	165554	791
			orf1	14423	+	363261	363462	201
NCYC9	2	6	ahd1	3119	-	162039	162150	111
			fas2	3070	-	485798	487048	1250
			cyp1	3205	+	11385	11575	190
			fhd1	3040	+	153470	153727	257
			atr1	3070	-	260292	260465	173
			orf1	3076	+	227670	227808	138
NCYC3109	2	6	ahd1	994	+	143140	143265	125

			fas2	935	-	751218	758683	7465
			cyp2	920	+	80457	80532	75
			cyp1	998	+	1121325	1121417	92
			orf1	998	+	1332357	1332516	159
			atr1	1014	-	195139	197049	1910
NCYC2509	2	6	ahd1	18419	+	20972	21561	589
			fas2	18416	-	142188	149633	7445
			cyp2	18313	+	189723	189827	104
			cyp1	18952	+	108481	108694	213
			orf1	18323	-	278460	278633	173
			atr1	18323	+	205320	209120	3800
NCYC784	2	6	fas2	8667	+	388087	395645	7558
			fat3	8648	-	3684743	3685041	298
			cyp1	8619	+	466792	466970	178
			orf1	8542	+	1775515	1775635	120
			atr1	8705	+	1040130	1042467	2337
			fhd1	8648	+	10846	11520	674
NCYC2424	2	6	ahd1	8432	+	76774	76904	130
			fas2	8442	-	499618	505290	5672
			cyp1	8440	+	469752	470365	613
			fhd1	8384	+	4065	4139	74
			atr1	8398	-	2514656	2516330	1674
			orf1	8398	+	360352	360473	121
NCYC2572	2	6	ahd1	3627	+	155429	155540	111
			fas2	3449	-	463707	464957	1250
			cyp1	3636	+	11491	11679	188
			fhd1	3584	+	153723	153827	104
			atr1	3449	-	243964	244137	173
			orf1	3647	-	297449	297526	77
NCYC2913	2	6	ahd1	13914	+	25245	25395	150
			fas2	13943	+	25054	30908	5854
			rfl1	13845	+	21094	21228	134
			cyp1	2515	-	18707	19187	480
			orf1	13914	+	74516	74737	221
			atr1	13904	+	8372	9945	1573
NCYC689	2	6	ahd1	5030	+	155391	155502	111
			fas2	5075	-	670981	672231	1250
			cyp1	5125	+	11583	11771	188
			orf1	5041	+	12258	12342	84
			atr1	5075	-	451251	451423	172
			fhd1	4727	+	152778	152882	104
NCYC3740	2	6	ahd1	1615	+	480447	480849	402
			fas2	1631	-	337780	342017	4237
			cyp1	1636	+	626808	627304	496

			fhd1	1645	+	273108	273418	310
			atr1	1618	+	158098	158815	717
			orf1	1631	+	96780	96969	189
NCYC2745	2	6	ahd1	17168	-	367459	367609	150
			fas2	17144	+	25054	30908	5854
			atr1	17163	+	182180	183753	1573
			cyp1	17162	-	18703	19183	480
			orf1	17168	-	318117	318338	221
			rfl1	17158	-	214290	214424	134
NCYC71	2	6	ahd1	2984	-	159241	159352	111
			fas2	2985	+	189699	190949	1250
			cyp1	3035	-	1567	1884	317
			orf1	2960	+	171257	171387	130
			atr1	2985	+	413118	413296	178
			fhd1	2942	-	185453	185704	251
NCYC2665	3	5	ahd1	2393	-	1196281	1196516	235
			fas2	2304	-	393296	400728	7432
			cyp1	2393	+	1441587	1442040	453
			orf1	2393	-	1445455	1446020	565
			atr1	2390	-	77345	80851	3506
NCYC3530	2	5	ahd1	14271	+	465579	465704	125
			fas2	14271	-	335937	337189	1252
			orf1	14263	-	22653	22737	84
			atr1	14020	-	467195	467525	330
			fhd1	13666	-	220946	221050	104
CBS6284	2	5	ahd1	60750	+	485521	485910	389
			fas2	58419	+	76438	78694	2256
			orf1	60822	-	493245	493320	75
			atr1	60822	+	316875	317510	635
			fhd1	58768	+	319392	320024	632
NCYC764	2	5	ahd1	1673	-	811492	811587	95
			fas2	1688	-	90867	96540	5673
			cyp1	1691	+	235247	235987	740
			orf1	1582	+	287659	287812	153
			atr1	1673	-	542709	544379	1670
NCYC23	2	5	ahd1	5057	+	792820	792956	136
			fas2	5299	+	528458	530798	2340
			orf1	5036	+	55973	56536	563
			atr1	5293	+	457791	458453	662
			fhd1	5299	+	1053273	1053345	72
NCYC3983	2	5	ahd1	8067	+	32691	32770	79
			fas2	8035	-	34791	42700	7909
			cyp1	7547	+	1917	2565	648
			orf1	7547	+	72299	72606	307

			atr1	7824	+	7573	10789	3216
NCYC2569	2	5	ahd1	1844	-	495266	495361	95
			fas2	1832	-	91773	97446	5673
			cyp1	1833	-	187582	188322	740
			orf1	1695	+	495815	495968	153
			atr1	1844	-	226515	228185	1670
CBS285	2	5	ahd1	4024	+	837461	837850	389
			fas2	4100	-	28198	30454	2256
			orf1	3984	-	493204	493279	75
			atr1	3984	+	316912	317547	635
			fhd1	3966	-	234488	235120	632
NCYC568	2	5	ahd1	3019	-	803813	804175	362
			fas2	2921	-	587468	589725	2257
			orf1	3023	+	188080	188255	175
			atr1	2932	-	114822	115271	449
			fhd1	3019	+	1348566	1348825	259
NCYC3541	2	5	ahd1	1424	-	234119	234194	75
			fas2	1398	+	424092	431737	7645
			cyp1	1424	-	658095	658179	84
			orf1	1398	-	57483	57641	158
			atr1	1436	-	186899	188629	1730
NCYC1141	2	5	ahd1	49408	-	216706	217313	607
			fas2	49948	+	5900	8174	2274
			fhd1	51598	+	261715	262034	319
			atr1	51726	-	204272	205140	868
			orf1	51726	-	647149	647272	123
NCYC2431	2	5	ahd1	11896	+	245088	245239	151
			fas2	11908	-	70833	76502	5669
			cyp1	11896	-	185039	185344	305
			orf1	11770	+	75695	75843	148
			atr1	11912	-	295231	296955	1724
NCYC36	2	5	ahd1	8562	+	113075	113200	125
			fas2	8864	+	1111818	1114649	2831
			fhd1	8562	-	167355	167599	244
			atr1	8542	+	62520	63191	671
			orf1	8615	-	281688	281768	80
NCYC2400	2	5	ahd1	12281	-	285941	286092	151
			fas2	12863	-	187858	191929	4071
			cyp1	12281	+	345828	346133	305
			orf1	12862	+	44025	44173	148
			atr1	12893	-	106441	108165	1724
NCYC1388	2	5	ahd1	15382	+	465554	465679	125
			fas2	15382	-	335911	337163	1252
			orf1	15378	-	22644	22728	84

			atr1	14823	+	99880	100231	351
			fhd1	14746	-	220946	221050	104
NCYC2480	2	5	ahd1	1424	+	4838	4986	148
			fas2	1711	-	56159	58396	2237
			orf1	1711	-	127491	127672	181
			atr1	1791	-	166685	167404	719
			fhd1	1789	+	32470	33012	542
NCYC3108	2	4	ahd1	2476	-	168650	168779	129
			fas2	2404	-	52831	54069	1238
			fhd1	2234	-	37388	37980	592
			atr1	2234	+	40052	40554	502
NCYC2889	2	4	ahd1	7282	+	39281	39636	355
			fas2	7310	-	23330	25583	2253
			fhd1	7393	+	14148	14922	774
			atr1	7310	-	98242	98793	551
CBS8763	2	4	ahd1	2862	-	164492	164620	128
			fas2	3070	+	266123	268378	2255
			orf1	3070	+	235577	235651	74
			atr1	3042	+	31705	32314	609
NCYC3776	2	4	ahd1	8711	-	34598	34727	129
			fas2	8719	+	209730	210968	1238
			fhd1	8599	+	21430	22022	592
			atr1	8599	-	18855	19357	502
NCYC2558	2	4	ahd1	22878	-	137697	137822	125
			fas2	22878	+	266121	267373	1252
			fhd1	23163	-	11666	11770	104
			atr1	22284	+	27550	27880	330
NCYC416	2	4	ahd1	1779	+	608755	608859	104
			fas2	1772	-	347202	349457	2255
			fhd1	1779	-	32530	32725	195
			atr1	1719	+	142189	142703	514
NCYC3369	2	4	ahd1	18612	+	401969	402094	125
			fas2	18612	-	272451	273703	1252
			fhd1	19097	-	25910	26014	104
			atr1	19014	+	27550	27880	330
NCYC2898	2	4	ahd1	2424	-	96231	96360	129
			fas2	2151	+	57296	58534	1238
			fhd1	2182	-	37251	37843	592
			atr1	2182	+	39915	40417	502
NCYC733	2	4	ahd1	2684	-	189812	189940	128
			fas2	2661	+	266206	268461	2255
			orf1	2661	+	235633	235707	74
			atr1	2566	+	31648	32257	609
NCYC1656	2	3	fas2	12139	+	17938	19171	1233

	orf1	12139	-	15180	15271	91
	atr1	11693	+	27002	27433	431

Table B.1: Full results table from the FindClusters search of the NCYC genomes. MaxCluster indicates the longest string of consecutive CBL genes in the relevant genome. NumGenes indicates the total number of CBL gene matches found. This will generally be larger than MaxCluster unless the entire gene cluster is present. The rest of the columns are the genomic coordinates of the gene matches.

C Appendix C

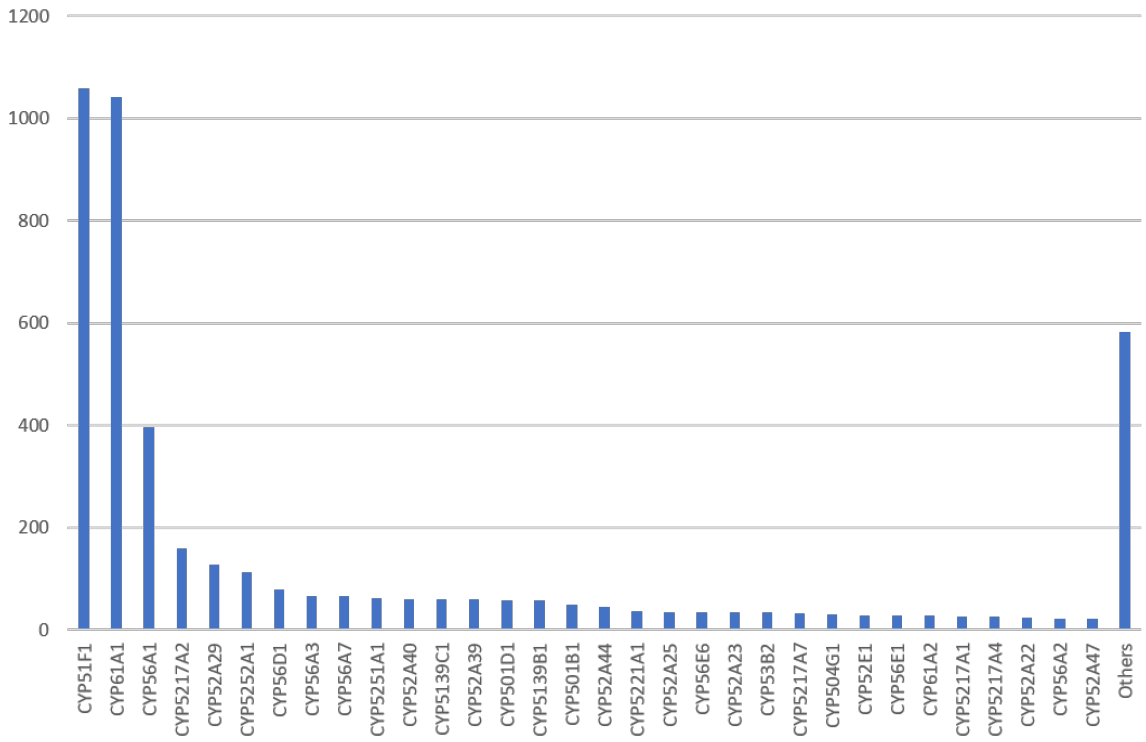

Figure C.1: Chart showing the number of putative CYPs classified as belonging to various known CYP families/ subfamilies. The top three here are CYP51, CYP56, and CYP61, which are the three found in *S. cerevisiae* (many sequenced strains belong to this species).

Strain	**Known CYPs Identified**
NCYC1	CYP52A45, CYP52A46, CYP52A47, CYP52A44, CYP5217A5, CYP52A43, CYP51F1
NCYC2	CYP505D6
NCYC4	CYP56E1, CYP61A2, CYP52C1, CYP51F1, CYP52A2, CYP52A1, CYP5217A1, CYP52A8, CYP52B1, CYP52A7, CYP52A6
NCYC6	CYP51F1, CYP5252A1, CYP61A1
NCYC8	CYP52A47, CYP52A46, CYP52A45, CYP52A44, CYP51F1, CYP5217A5, CYP52A43
NCYC9	CYP51F1, CYP52A45, CYP52A46, CYP52A47, CYP52A44, CYP52A43, CYP5217A5

NCYC10	CYP51F1, CYP51F1, CYP505D6, CYP52A44, CYP5217A5, CYP52A43, CYP52A45, CYP52A46, CYP52A47, CYP52A47, CYP52A46, CYP52A45, CYP52A45, CYP51F1
NCYC16	CYP61A1, CYP51F1, CYP56E6
NCYC17	CYP61A1, CYP501D1, CYP56E6, CYP51F1, CYP501D1
NCYC17A	CYP61A1, CYP501D1, CYP51F1, CYP56E6
NCYC17B	CYP56A10, CYP61A1, CYP5252A1, CYP51F1
NCYC18	CYP501D1, CYP56E6, CYP51F1, CYP51F1, CYP51F1, CYP61A1
NCYC20	CYP61A1, CYP61A1, CYP501D1, CYP51F1, CYP51F1, CYP5217A7, CYP5217A7, CYP5217A7, CYP56E6, CYP51F1, CYP61A1
NCYC21	CYP61A1, CYP51F1, CYP504G1
NCYC23	CYP56E6, CYP61A1, CYP51F1, CYP5217A7, CYP501D1
NCYC26	CYP56A10, CYP61A1, CYP51F1, CYP5252A1
NCYC31	CYP51F1, CYP61A1, CYP5252A1, CYP56A10
NCYC36	CYP5252A1, CYP56A10, CYP61A1, CYP51F1
NCYC39	CYP52A32, CYP52A13, CYP52A18, CYP56E4, CYP52A27, CYP52A23, CYP52A40, CYP52A33, CYP52A40, CYP52A39, CYP52A47, CYP52A23, CYP52A27, CYP61A2, CYP5217A4, CYP51F1
NCYC40	CYP52A39, CYP52A18, CYP52A14, CYP51F1, CYP52A23, CYP56E4, CYP52A33, CYP52A40, CYP52A39, CYP52A32, CYP61A2, CYP52A42, CYP52A23, CYP52A47, CYP5217A4, CYP52A13
NCYC43	CYP51F1, CYP61A1
NCYC44	CYP504G1, CYP61A1, CYP51F1
NCYC45	CYP51F1, CYP61A1
NCYC46	CYP56A1, CYP51F1, CYP61A1
NCYC47	CYP52H6, CYP51F1, CYP61A1, CYP56B2, CYP504A17, CYP51F1
NCYC49	CYP51F1, CYP51F1, CYP61A1, CYP504A17, CYP52H6, CYP56B3
NCYC52	CYP61A1, CYP504G1, CYP51F1
NCYC54	CYP61A1, CYP51F1, CYP51F1, CYP504G1
NCYC55	CYP504G1, CYP61A1, CYP51F1, CYP51F1
NCYC56	CYP5217A5, CYP5217A7, CYP52A43, CYP52A30, CYP52A39, CYP56E6, CYP52A53, CYP51F1, CYP61A1
NCYC57	CYP501D1, CYP5217A7, CYP51F1, CYP61A1, CYP56E6
NCYC58	CYP61A1, CYP51F1, CYP61A1, CYP5252A1, CYP56A10, CYP61A1
NCYC59	CYP505D6
NCYC60	CYP5065A3, CYP53B2, CYP5139C1, CYP5221A1, CYP5139C1, CYP5222A1, CYP51F1
NCYC63	CYP53B2, CYP5221A1, CYP61A1, CYP5139C1, CYP5139C1, CYP51F1
NCYC64	CYP51F1, CYP5221A1, CYP61A1, CYP5139C1, CYP5139C1, CYP53B2

NCYC65	CYP5139C1, CYP53B2, CYP5221A1, CYP5139C1, CYP51F1, CYP61A1
NCYC68	CYP5221A1, CYP53B2, CYP5139C1, CYP5139C1, CYP51F1, CYP61A1
NCYC70	CYP61A1, CYP56A1, CYP61A1, CYP51F1, CYP61A1
NCYC71	CYP52A47, CYP51F1, CYP5217A5, CYP52A46, CYP52A45, CYP52A46, CYP52A47, CYP52A44, CYP61A1, CYP52A47, CYP5217A5, CYP5217A5, CYP52A43
NCYC72	CYP56A1, CYP51F1, CYP61A1
NCYC73	CYP51F1, CYP56A3, CYP51F1, CYP61A1, CYP61A1
NCYC74	CYP61A1, CYP56A1, CYP51F1
NCYC75	CYP61A1, CYP56A1, CYP51F1
NCYC76	CYP51F1, CYP56A1, CYP61A1
NCYC77	CYP51F1, CYP51F1, CYP61A1, CYP61A1, CYP56A1
NCYC78	CYP51F1, CYP56A1, CYP61A1
NCYC79	CYP61A1, CYP56A1, CYP51F1
NCYC80	CYP61A1, CYP51F1, CYP61A1, CYP56A1
NCYC81	CYP61A1, CYP61A1, CYP56A1, CYP61A1, CYP61A1, CYP56A1, CYP51F1, CYP56A1
NCYC82	CYP56A1, CYP61A1, CYP51F1
NCYC83	CYP51F1, CYP51F1, CYP61A1, CYP56A1, CYP51F1
NCYC85	CYP61A1, CYP56A1, CYP51F1
NCYC86	CYP51F1, CYP51F1, CYP51F1, CYP56A1, CYP61A1
NCYC88	CYP61A1, CYP56A1, CYP51F1
NCYC89	CYP51F1, CYP51F1, CYP51F1, CYP56A1, CYP61A1
NCYC90	CYP51F1, CYP51F1, CYP51F1, CYP56A1, CYP61A1
NCYC92	CYP61A1, CYP51F1, CYP51F1, CYP56A1, CYP51F1
NCYC93	CYP61A1, CYP51F1, CYP56A1
NCYC94	CYP61A1, CYP61A1, CYP56A1, CYP56A1, CYP51F1
NCYC95	CYP61A1, CYP51F1, CYP51F1, CYP51F1, CYP56A1
NCYC96	CYP56A1, CYP61A1, CYP51F1
NCYC97	CYP51F1, CYP56A1, CYP61A1
NCYC98	CYP61A1, CYP51F1
NCYC100	CYP61A1, CYP56A7, CYP5252A1, CYP51F1
NCYC101	CYP504G1, CYP61A1, CYP51F1
NCYC102	CYP501D1, CYP51F1, CYP61A1, CYP56E1, CYP5217A7
NCYC103	CYP61A1, CYP51F1
NCYC105	CYP51F1, CYP61A1
NCYC106	CYP56E2, CYP51F1, CYP51F1, CYP56A1, CYP501D1, CYP56E6, CYP61A1, CYP51F1, CYP51F1, CYP51F1
NCYC111	CYP56A7, CYP61A1, CYP5252A1, CYP51F1
NCYC112	CYP51F1, CYP56A1, CYP51F1, CYP56A3, CYP61A1, CYP61A1
NCYC114	CYP56A3, CYP56A3, CYP61A1, CYP51F1
NCYC115	CYP51F1, CYP56A3, CYP61A1, CYP61A1, CYP51F1

NCYC116	CYP51F1, CYP56A1, CYP61A1
NCYC119	CYP56E6, CYP61A1, CYP51F1, CYP501D1, CYP51F1, CYP61A1
NCYC120	CYP51F1, CYP61A1
NCYC127	CYP501D1, CYP51F1, CYP61A1
NCYC128	CYP61A1, CYP61A1, CYP61A1, CYP51F1
NCYC129	CYP501D1, CYP56E6, CYP51F1, CYP61A1
NCYC130	CYP51F1, CYP61A1, CYP504G1, CYP61A1, CYP61A1, CYP61A1, CYP51F1
NCYC131	CYP51F1, CYP61A1, CYP51F1, CYP504G1, CYP61A1
NCYC132	CYP61A1, CYP51F1
NCYC133	CYP61A1, CYP51F1, CYP52A42
NCYC135	CYP51F1, CYP53B2, CYP5139C1, CYP5221A1, CYP5139C1, CYP61A1
NCYC137	CYP51F1, CYP61A1, CYP56A3, CYP61A1, CYP51F1
NCYC138	CYP5152A3, CYP5231A1, CYP53B2, CYP51F1, CYP61A1
NCYC140	CYP51F1, CYP56A3, CYP61A1
NCYC141	CYP52A29, CYP5217A2, CYP52A29, CYP51F1, CYP56D1, CYP51F1, CYP52A40, CYP61A1, CYP5251A1, CYP5217A2, CYP56A3, CYP5217A2, CYP61A1
NCYC142	CYP53B2, CYP51F1, CYP5221A1, CYP5139C1, CYP61A1, CYP5139C1
NCYC143	CYP51F1, CYP61A1, CYP61A1, CYP56A7, CYP5252A1
NCYC144	CYP504G1, CYP52A39, CYP5217A4, CYP52A41, CYP52A39, CYP61A1, CYP501C1, CYP51F1, CYP52A40, CYP52A42
NCYC145	CYP504G1, CYP5217A4, CYP52A41, CYP52A39, CYP501C1, CYP52A39, CYP61A1, CYP51F1, CYP52A42, CYP52A40
NCYC147	CYP5217A2, CYP5251A1, CYP56D1, CYP5217A7, CYP5217A2, CYP52A29, CYP501B1, CYP61A1, CYP51F1, CYP61A1, CYP56A4, CYP52A29, CYP52A40, CYP51F1
NCYC148	CYP51F1, CYP56A1, CYP52N1, CYP61A1
NCYC151	CYP5252A1, CYP61A1, CYP56A7, CYP51F1
NCYC152	CYP5252A1, CYP61A1, CYP56A7, CYP51F1
NCYC154	CYP5139C1, CYP5221A1
NCYC155	CYP5139C1, CYP5221A1, CYP5221A1
NCYC158	CYP51F1, CYP53B2, CYP5139C1, CYP5221A1, CYP61A1
NCYC159	CYP5221A1, CYP53B2, CYP5139C1, CYP51F1, CYP61A1, CYP5139C1
NCYC161	CYP61A1, CYP51F1, CYP56A3
NCYC162	CYP51F1, CYP53B2, CYP5065A3, CYP5221A1, CYP5222A1, CYP5139C1
NCYC163	CYP56A3, CYP61A1, CYP51F1, CYP61A1
NCYC166	CYP52A31, CYP56D1, CYP51F1, CYP51F1, CYP501B1, CYP501B1, CYP56D1, CYP56D1, CYP56D1, CYP61A1, CYP52A29, CYP501B1, CYP52A39, CYP52A44, CYP5251A1, CYP5217A2, CYP5217A2

NCYC168	CYP51F1, CYP61A1, CYP51F1, CYP56E6, CYP501D1, CYP51F1, CYP61A1
NCYC169	CYP51F1, CYP61A1
NCYC171	CYP61A1, CYP61A1, CYP61A1, CYP51F1
NCYC179	CYP56A7, CYP61A1, CYP5252A1, CYP51F1
NCYC188	CYP5252A1, CYP56A7, CYP51F1, CYP61A1
NCYC191	CYP51F1, CYP501D1, CYP5217A7, CYP61A1, CYP5217A7, CYP56E6
NCYC195	CYP53B2, CYP5221A1, CYP5139C1, CYP51F1, CYP5139C1, CYP61A1
NCYC232	CYP51F1, CYP61A1, CYP51F1, CYP56A1
NCYC234	CYP61A1, CYP51F1, CYP61A1, CYP56A1, CYP61A1
NCYC235	CYP61A1, CYP56A1, CYP51F1
NCYC236	CYP51F1, CYP56A1, CYP61A1
NCYC238	CYP51F1, CYP, CYP51F1, CYP, CYP5033A1, CYP53C2, CYP504A8, CYP, CYP61A1, CYP56A1, CYP61A1
NCYC239	CYP61A1, CYP61A1, CYP56A1, CYP51F1, CYP56A3
NCYC240	CYP61A1, CYP51F1, CYP56A1, CYP61A1
NCYC241	CYP61A1, CYP61A1, CYP56A1, CYP51F1
NCYC243	CYP56A7, CYP5252A1, CYP61A1, CYP51F1
NCYC244	CYP61A1, CYP5252A1, CYP56A7, CYP51F1, CYP51F1
NCYC324	CYP56A3, CYP61A1, CYP51F1
NCYC329	CYP51F1, CYP61A1
NCYC332	CYP61A1, CYP61A1, CYP51F1, CYP51F1, CYP51F1, CYP61A1
NCYC336	CYP61A1, CYP51F1
NCYC337	CYP61A1, CYP51F1
NCYC338	CYP61A1, CYP51F1, CYP51F1, CYP51F1
NCYC341	CYP56A3, CYP56A1, CYP51F1, CYP56A3, CYP61A1, CYP61A1, CYP51F1, CYP61A1, CYP51F1
NCYC343	CYP61A1, CYP56A3, CYP61A1, CYP61A1, CYP51F1, CYP56A3, CYP51F1, CYP51F1
NCYC347	CYP51F1, CYP61A1
NCYC350	CYP61A1, CYP56A5, CYP51F1
NCYC353	CYP56A5, CYP61A1, CYP51F1
NCYC356	CYP51F1, CYP56A1, CYP61A1
NCYC357	CYP51F1, CYP61A1, CYP56A1
NCYC358	CYP51F1, CYP61A1, CYP56A1
NCYC360	CYP56A1, CYP61A1, CYP61A1, CYP51F1
NCYC361	CYP51F1, CYP56A1, CYP61A1
NCYC363	CYP56A1, CYP51F1, CYP61A1
NCYC367	CYP56A1, CYP51F1, CYP61A1
NCYC368	CYP56A1, CYP51F1, CYP61A1

NCYC371	CYP52A29, CYP51F1, CYP52A40, CYP52A29, CYP52A40, CYP52A40, CYP52A40, CYP56D1, CYP56D1, CYP56D1, CYP56D1, CYP51F1, CYP51F1, CYP5217A2, CYP52A29, CYP61A1, CYP52A29, CYP501B1, CYP5217A7, CYP5251A1
NCYC372	CYP52A29, CYP56D1, CYP51F1, CYP501B1, CYP501B1, CYP61A1, CYP61A1, CYP52A29, CYP52A31, CYP61A1, CYP51F1, CYP5217A2, CYP5217A2, CYP5217A2, CYP5217A2, CYP52A29, CYP52A29, CYP5251A1, CYP52A40, CYP61A1
NCYC373	CYP5217A2, CYP52A29, CYP51F1, CYP5217A2, CYP52A29, CYP5217A2, CYP52A29, CYP501B1, CYP56D1, CYP52A29, CYP52A40, CYP56D1, CYP501B1, CYP51F1, CYP52A39, CYP52A39, CYP52A39, CYP52A39, CYP56D1, CYP56D1, CYP5217A2, CYP5217A2, CYP51F1, CYP5251A1, CYP51F1, CYP52A44, CYP5217A7, CYP5217A2
NCYC377	CYP51F1, CYP53B2, CYP61A1, CYP5139C1, CYP5221A1, CYP5139C1
NCYC385	CYP51F1, CYP56A1, CYP61A1
NCYC388	CYP56A5, CYP51F1, CYP61A1
NCYC392	CYP61A1, CYP51F1, CYP56A3
NCYC400	CYP51F1, CYP56A3, CYP61A1, CYP61A1, CYP51F1
NCYC401	CYP56A1, CYP61A1, CYP51F1
NCYC407	CYP51F1, CYP61A1, CYP52A44, CYP52A43, CYP5217A5, CYP52A45, CYP52A46, CYP52A47
NCYC408	CYP52A29, CYP52A29, CYP52A40, CYP51F1, CYP52A45, CYP61A1, CYP501B1, CYP52A40, CYP5217A2, CYP52A31, CYP5217A4, CYP51F1, CYP51F1, CYP61A1, CYP56D1, CYP56A3, CYP56D1
NCYC411	CYP5217A2, CYP51F1, CYP56D1, CYP61A1, CYP61A1, CYP501B1, CYP5217A2, CYP52A29, CYP56A3, CYP5251A1, CYP52A40, CYP51F1, CYP52A29
NCYC416	CYP5251A1, CYP5252A1, CYP5252A1, CYP5252A1, CYP61A1, CYP5252A1, CYP5252A1, CYP61A1, CYP56A7, CYP5251A1, CYP61A1, CYP5251A1, CYP51F1
NCYC417	CYP56A2, CYP61A1, CYP61A1, CYP51F1, CYP61A1
NCYC426	CYP56A7, CYP5252A1, CYP5252A1, CYP5252A1, CYP61A1, CYP51F1
NCYC430	CYP51F1, CYP56A1, CYP61A1
NCYC431	CYP61A1, CYP51F1, CYP56A1, CYP5252A1, CYP51F1, CYP61A1, CYP56A7
NCYC436	CYP5252A1, CYP5252A1, CYP5252A1, CYP5252A1, CYP501D1, CYP56E6
NCYC437	CYP5252A1, CYP5252A1, CYP5252A1, CYP5252A1, CYP56E6, CYP501D1

NCYC438	CYP5252A1, CYP5252A1, CYP5252A1, CYP5252A1, CYP56E6, CYP501D1
NCYC442	CYP5252A1, CYP501D1, CYP56E6
NCYC444	CYP5139B2, CYP505D6, CYP51F1, CYP5139B1., CYP5139B1, CYP504ANeosartorya
NCYC461	CYP504G1, CYP51F1, CYP61A1
NCYC463	CYP61A1, CYP56A3, CYP51F1
NCYC464	CYP61A1, CYP51F1, CYP56A1
NCYC472	CYP504A9, CYP5139B1, CYP53A8, CYP5139B1., CYP5139B1., CYP51F1, CYP5139B1., CYP53A4, CYP505D6, CYP61A1
NCYC478	CYP56A1, CYP56A1, CYP56A1, CYP51F1, CYP61A1
NCYC479	CYP61A1, CYP56A1, CYP51F1
NCYC482	CYP61A1, CYP56A1, CYP51F1
NCYC486	CYP51F1, CYP61A1
NCYC489	CYP51F1, CYP540B10, CYP51F1, CYP61A1, CYP, CYP5033A1, CYP61A1, CYP56A1
NCYC490	CYP51F1, CYP61A1, CYP61A1, CYP61A1, CYP56A1
NCYC491	CYP51F1, CYP56A1, CYP61A1
NCYC492	CYP501D1, CYP61A1, CYP61A1, CYP61A1, CYP51F1, CYP56E4, CYP56E6, CYP61A1
NCYC502	CYP5221A1, CYP5139C1
NCYC505	CYP61A1, CYP61A1, CYP51F1, CYP51F1, CYP56A1
NCYC523	CYP51F6, CYP61A1, CYP56A12, CYP51F1
NCYC524	CYP56A3, CYP61A1, CYP51F1
NCYC525	CYP5104B1, CYP53C2, CYP5033A1, CYP61A1, CYP51F1, CYP51F1, CYP56A1
NCYC538	CYP51F1, CYP5251A1, CYP56E6, CYP61A1, CYP56A7, CYP5252A1, CYP501D1, CYP51F1, CYP61A1
NCYC541	CYP51F1, CYP5139C1, CYP5221A1, CYP61A1, CYP51F1, CYP5139C1, CYP53B2
NCYC543	CYP56A10, CYP61A1, CYP51F1, CYP5252A1, CYP5252A1
NCYC546	CYP61A1, CYP51F1, CYP56A1
NCYC548	CYP56A7, CYP61A1, CYP51F1, CYP5252A1, CYP5251A1
NCYC551	CYP56A7, CYP61A1, CYP5252A1, CYP5251A1, CYP51F1
NCYC553	CYP61A1, CYP51F1
NCYC558	CYP5252A1, CYP56E6, CYP61A1, CYP501D1, CYP56A1, CYP51F1
NCYC559	CYP56A3, CYP51F1, CYP61A1
NCYC563	CYP51F1, CYP61A1, CYP61A1, CYP56A2, CYP61A1
NCYC566	CYP52A29, CYP52A29, CYP501D1, CYP51F1, CYP5217A2, CYP52A40, CYP51F1, CYP52A44, CYP61A1, CYP51F1, CYP56A3, CYP52A37, CYP5217A4, CYP5217A2, CYP52A29, CYP5217A4, CYP61A1
NCYC568	CYP61A1, CYP51F1, CYP56A3
NCYC570	CYP51F1, CYP56A7, CYP5251A1, CYP61A1, CYP5252A1

NCYC571	CYP56A7, CYP51F1, CYP5252A1, CYP5251A1, CYP61A1
NCYC573	CYP56A1, CYP51F1, CYP61A1
NCYC575	CYP51F1, CYP5252A1, CYP5251A1, CYP56A7, CYP61A1
NCYC576	CYP52A45, CYP52A45, CYP61A1, CYP52A47, CYP5217A5, CYP52A43, CYP52A44, CYP51F1, CYP52A47
NCYC580	CYP51F1, CYP56A1, CYP61A1
NCYC582	CYP52A37, CYP52A29, CYP5217A2, CYP56D1, CYP56D1, CYP52A29, CYP61A1, CYP5217A2, CYP51F1, CYP51F1, CYP5217A2, CYP52A44, CYP501D1, CYP501B1, CYP52A40, CYP5217A2, CYP52A29, CYP61A1, CYP5217A2, CYP52A29, CYP5217A2, CYP5217A2
NCYC585	CYP52A29, CYP52A29, CYP56D1, CYP61A1, CYP501B1, CYP5217A2, CYP5217A2, CYP501D1, CYP52A29, CYP56D1, CYP51F1, CYP51F1, CYP52A44, CYP5217A2, CYP5217A2, CYP51F1, CYP61A1, CYP52A44, CYP52A29, CYP5217A2, CYP5217A2, CYP5217A2
NCYC587	CYP51F1, CYP5252A1, CYP56A7, CYP61A1
NCYC591	CYP61A1, CYP51F1, CYP5139B1, CYP504G1, CYP51F1, CYP5139B2
NCYC597	CYP56E2, CYP52C3, CYP61A2, CYP501A1, CYP52A22, CYP52A23, CYP52A25, CYP52A25, CYP52A25, CYP5217A1, CYP52A21, CYP52A21, CYP52A21, CYP52C3, CYP52A25
NCYC601	CYP51F1, CYP52A35, CYP5217A3, CYP52A38, CYP52A34, CYP52A34, CYP52A33, CYP52C5, CYP52A37, CYP56E4, CYP52A34, CYP52A32, CYP52C6, CYP61A1, CYP501A4
NCYC608	CYP501B1, CYP52A37, CYP52A29, CYP51F1, CYP5217A2, CYP56D1, CYP52A29, CYP61A1, CYP52A44, CYP5217A2, CYP56D1, CYP5217A2, CYP52A29, CYP5217A2, CYP5217A2, CYP501D1, CYP52A29, CYP5217A2, CYP5217A2, CYP51F1, CYP56A3, CYP5217A2, CYP61A1, CYP5217A2
NCYC609	CYP56A1, CYP61A1, CYP51F1
NCYC610	CYP501A1, CYP61A2, CYP61A2, CYP52A21, CYP56E2, CYP52C3, CYP52A23, CYP52A25, CYP5217A1, CYP52A22
NCYC611	CYP52A44, CYP51F1, CYP5217A5, CYP52A43, CYP52A47, CYP52A46, CYP52A45
NCYC619	CYP51F1, CYP56A1, CYP61A1
NCYC620	CYP61A1, CYP56A1, CYP51F1
NCYC621	CYP61A1, CYP56A1, CYP51F1
NCYC622	CYP56A1, CYP61A1, CYP51F1
NCYC667	CYP51F1, CYP61A1, CYP56A1, CYP61A1, CYP61A1
NCYC671	CYP56A1, CYP61A1, CYP61A1, CYP51F1, CYP56A3
NCYC672	CYP61A1, CYP51F1, CYP56A1
NCYC675	CYP61A3, CYP51F1, CYP51F1, CYP61A1, CYP56A1

NCYC677	CYP56D1, CYP52A29, CYP52A29, CYP501B1, CYP61A1, CYP5217A2, CYP61A1, CYP51F1, CYP51F1, CYP52A29, CYP56A3, CYP5217A4, CYP52A44, CYP51F1, CYP51F1, CYP61A1, CYP5217A2
NCYC678	CYP51F1, CYP56A3, CYP52A29, CYP52A37, CYP51F1, CYP61A1, CYP5217A2, CYP51F1, CYP56A3, CYP61A1, CYP51F1, CYP61A1
NCYC684	CYP51F1, CYP61A1, CYP56A1
NCYC689	CYP5217A5, CYP52A47, CYP52A46, CYP52A45, CYP52A44, CYP5217A5, CYP52A43, CYP51F1
NCYC694	CYP, CYP504A8, CYP, CYP51F1, CYP53C2, CYP5033A1, CYP51F1, CYP56A1, CYP61A1
NCYC695	CYP56A1, CYP51F1, CYP61A1
NCYC696	CYP61A1, CYP51F1, CYP56A3
NCYC730	CYP56E6, CYP56E6, CYP61A1, CYP61A1, CYP61A1, CYP56E6, CYP61A1, CYP51F1
NCYC731	CYP61A1, CYP56A3, CYP51F1, CYP51F1, CYP56A1, CYP61A1
NCYC732	CYP51F1, CYP56A5, CYP61A1
NCYC733	CYP56A7, CYP51F1, CYP61A1, CYP5252A1
NCYC734	CYP56A10, CYP5252A1, CYP51F1, CYP61A1, CYP51F1, CYP61A1, CYP5252A1, CYP56A7
NCYC739	CYP56A1, CYP5252A1, CYP61A1, CYP51F1, CYP61A1, CYP56A7, CYP51F1, CYP5251A1
NCYC744	CYP56A7, CYP61A1, CYP51F1, CYP5252A1, CYP5252A1, CYP5252A1
NCYC745	CYP56D1, CYP61A1, CYP56D1, CYP52A31, CYP52A29, CYP51F1, CYP56D1, CYP501B1, CYP51F1, CYP61A1, CYP61A1, CYP5217A2, CYP52A44, CYP501B1, CYP501B1, CYP61A1, CYP5217A2, CYP56D1, CYP52A29, CYP52A29, CYP5217A2, CYP5217A2, CYP52A40
NCYC746	CYP52A29, CYP5217A2, CYP501B1, CYP51F1, CYP61A1, CYP56D1, CYP52A29, CYP5217A2, CYP5217A2, CYP52A44
NCYC747	CYP56D1, CYP51F1, CYP501B1, CYP5251A1, CYP5251A1, CYP51F1, CYP51F1, CYP61A1, CYP5217A2, CYP52A29, CYP52A29, CYP52A44, CYP5217A2, CYP501B1
NCYC752	CYP56A7, CYP51F1, CYP5252A1, CYP61A1, CYP5251A1
NCYC754	CYP56E6, CYP61A1, CYP501D1, CYP51F1, CYP501D1, CYP51F1
NCYC758	CYP53B2, CYP51F1, CYP5221A1, CYP5139C1, CYP61A1, CYP5139C1
NCYC759	CYP51F1, CYP53B2, CYP5221A1, CYP5139C1, CYP61A1, CYP5139C1
NCYC764	CYP51F1
NCYC765	CYP52N1
NCYC768	CYP56A5, CYP61A1, CYP51F1
NCYC776	CYP56A7, CYP51F1, CYP61A1, CYP5252A1

NCYC777	CYP56A2, CYP51F1, CYP61A1, CYP51F1
NCYC783	CYP51F1, CYP51F1, CYP52A29, CYP51F1, CYP5217A2, CYP61A1, CYP5217A2, CYP51F1, CYP501B1
NCYC784	CYP5215A1, CYP5139B1., CYP5139B1., CYP61A1, CYP51F1, CYP5216A1
NCYC794	CYP51F1, CYP51F1, CYP5217A2, CYP5217A2, CYP61A1, CYP51F1, CYP501B1, CYP501B1, CYP5217A2, CYP5217A2, CYP51F1, CYP52A29
NCYC796	CYP5139C1, CYP5221A1, CYP61A1, CYP501D1, CYP5139C1, CYP51F1, CYP61A1, CYP51F1, CYP53B2
NCYC807	CYP51F1, CYP501A1, CYP52A23, CYP61A1, CYP56A1, CYP61A2, CYP52A25
NCYC814	CYP61A1, CYP56A4, CYP51F1, CYP56A4, CYP61A1
NCYC816	CYP61A1, CYP61A1, CYP61A1, CYP51F1, CYP56A1
NCYC820	CYP5251A1, CYP52A29, CYP51F1, CYP61A1, CYP56D1, CYP5217A2, CYP61A1, CYP52A29, CYP52A40, CYP52A31, CYP56A3, CYP52A39, CYP51F1
NCYC826	CYP51F1, CYP61A1, CYP56A1
NCYC827	CYP5252A1, CYP51F1, CYP56A7, CYP61A1
NCYC844	CYP501D1, CYP51F1, CYP61A1
NCYC845	CYP61A1, CYP61A1, CYP61A1, CYP61A1, CYP61A1, CYP61A1
NCYC849	CYP51F1, CYP61A1, CYP504G1
NCYC851	CYP56A7, CYP51F1, CYP56A1, CYP51F1, CYP61A1, CYP5252A1, CYP5252A1, CYP5252A1, CYP5252A1, CYP61A1
NCYC854	CYP56E2, CYP52A22, CYP52A23, CYP501A1, CYP52A21, CYP52A25, CYP61A2, CYP52C3, CYP5217A1
NCYC872	CYP51F1, CYP61A1
NCYC894	CYP56D1, CYP5251A1, CYP52A29, CYP56D1, CYP52A44, CYP52A29, CYP5217A4, CYP5217A2, CYP51F1, CYP501B1, CYP52A31, CYP61A1
NCYC906	CYP61A1, CYP52A37, CYP52A35, CYP52C6, CYP52A32, CYP51F1, CYP56A7, CYP52A34, CYP5217A3, CYP51F1, CYP61A1, CYP5252A1, CYP56A7, CYP51F1, CYP51F1, CYP56A7, CYP52A38
NCYC911	CYP540B4, CYP5033A1, CYP504A8, CYP61A1, CYP53C2, CYP51F1
NCYC925	CYP52F8, CYP52F10, CYP52F1, CYP52S1, CYP52F9, CYP52F5, CYP504A17, CYP52F4, CYP52F7, CYP52F3, CYP52F11, CYP51F1, CYP548P1, CYP52F6, CYP5223A1, CYP52F6, CYP61A1, CYP52F2
NCYC929	CYP51F1, CYP61A1, CYP5252A1, CYP56A7, CYP5251A1
NCYC930	CYP5065A1, CYP61A1, CYP504B10, CYP53B2, CYP51F1, CYP53B2
NCYC931	CYP5221A1, CYP53B2, CYP5065A1, CYP504B10, CYP5231A1, CYP505A1, CYP51F1, CYP5231A1

NCYC935	CYP5217A7, CYP56E1, CYP52A44, CYP56E1, CYP56E1, CYP51F1, CYP61A1, CYP52A47
NCYC951	CYP56E6, CYP61A1, CYP501D1, CYP61A1, CYP61A1, CYP51F1, CYP5217A4, CYP51F1, CYP51F1, CYP56E6
NCYC956	CYP56A1, CYP51F1, CYP61A1
NCYC963	CYP56A1, CYP61A1, CYP51F1
NCYC970	CYP61A1, CYP51F1, CYP56A7, CYP5252A1
NCYC971	CYP51F1, CYP56A4, CYP61A1
NCYC974	CYP53B2, CYP5221A1, CYP5139C1, CYP5139C1, CYP51F1, CYP51F1, CYP51F1, CYP5139C1, CYP5221A1, CYP61A1, CYP5139C1, CYP61A1
NCYC975	CYP56A1, CYP61A1, CYP51F1, CYP61A1, CYP61A1
NCYC993	CYP61A1, CYP51F1
NCYC995	CYP56A1, CYP61A1, CYP51F1
NCYC996	CYP56A1, CYP51F1, CYP61A1
NCYC1001	CYP61A1, CYP56A1, CYP51F1
NCYC1004	CYP51F1, CYP61A1, CYP56A1
NCYC1006	CYP51F1, CYP51F1, CYP61A1, CYP56A1
NCYC1007	CYP51F1, CYP61A1, CYP56A1
NCYC1010	CYP61A1, CYP56A1, CYP51F1
NCYC1013	CYP51F1, CYP61A1, CYP56A1
NCYC1016	CYP56A1, CYP51F1, CYP61A1
NCYC1017	CYP61A1, CYP56A1, CYP51F1
NCYC1023	CYP61A1, CYP51F1, CYP56A1
NCYC1030	CYP61A1, CYP56A1, CYP51F1
NCYC1031	CYP51F1, CYP56A1, CYP501D1, CYP61A1, CYP56E6, CYP56E6, CYP51F1, CYP51F1, CYP61A1, CYP61A1, CYP501D1
NCYC1033	CYP61A1, CYP56A1, CYP51F1
NCYC1035	CYP56A1, CYP61A1, CYP51F1
NCYC1037	CYP51F1, CYP61A1, CYP56A1
NCYC1039	CYP61A1, CYP51F1, CYP61A1, CYP51F1, CYP56A1, CYP504G1
NCYC1044	CYP61A1, CYP56A1, CYP51F1, CYP61A1, CYP61A1
NCYC1046	CYP51F1, CYP61A1, CYP56A1
NCYC1049	CYP51F1, CYP56A1, CYP61A1
NCYC1052	CYP51F1, CYP56A1, CYP61A1
NCYC1053	CYP56A1, CYP61A1, CYP51F1
NCYC1054	CYP51F1, CYP56A1, CYP61A1
NCYC1055	CYP56A1, CYP61A1, CYP51F1, CYP61A1, CYP61A1
NCYC1060	CYP61A1, CYP51F1, CYP56A1
NCYC1063	CYP51F1, CYP56E6, CYP61A1
NCYC1064	CYP61A1, CYP61A1, CYP61A1, CYP51F1, CYP56A1
NCYC1066	CYP52A39, CYP504G1, CYP501C1, CYP5217A4, CYP52A39, CYP52A41, CYP61A1, CYP51F1, CYP52A42, CYP52A40
NCYC1069	CYP51F1, CYP51F1, CYP56A1, CYP61A1

NCYC1072	CYP61A1, CYP56A1, CYP51F1
NCYC1073	CYP51F1, CYP56A3, CYP61A1, CYP56A1, CYP61A1, CYP51F1
NCYC1079	CYP51F1, CYP61A1, CYP56A1
NCYC1082	CYP51F1, CYP61A1, CYP56A1, CYP61A1
NCYC1089	CYP61A1, CYP51F1, CYP56A1
NCYC1090	CYP61A1, CYP56A1, CYP51F1, CYP61A1, CYP61A1
NCYC1093	CYP51F1, CYP61A1, CYP61A1, CYP56A1
NCYC1097	CYP61A1, CYP56A1, CYP51F1
NCYC1102	CYP61A1, CYP56A1, CYP51F1
NCYC1103	CYP56E6, CYP61A1, CYP501D1, CYP51F1, CYP51F1, CYP56E6, CYP51F1, CYP56A1, CYP61A1, CYP61A1, CYP61A1
NCYC1106	CYP56A3, CYP61A1, CYP61A1, CYP56A1, CYP51F1, CYP51F1
NCYC1111	CYP56A1, CYP51F1, CYP61A1
NCYC1114	CYP61A1, CYP51F1, CYP56A1
NCYC1117	CYP61A1, CYP51F1, CYP56A1
NCYC1118	CYP51F1, CYP56A1, CYP61A1
NCYC1122	CYP61A1, CYP51F1, CYP56A1
NCYC1124	CYP51F1, CYP56A1, CYP61A1, CYP51F1
NCYC1129	CYP61A1, CYP56A1, CYP51F1
NCYC1132	CYP61A1, CYP56A1, CYP51F1
NCYC1134	CYP51F1, CYP51F1, CYP504G1, CYP61A1, CYP61A1, CYP56A1
NCYC1138	CYP52A39, CYP5217A4, CYP504G1, CYP61A1, CYP52A39, CYP501C1, CYP51F1, CYP51F1, CYP52A40, CYP52A42, CYP52A41, CYP61A1
NCYC1139	CYP51F1, CYP61A1, CYP504G1, CYP51F1, CYP56A1, CYP61A1
NCYC1141	CYP51F1, CYP504G1, CYP51F1, CYP56A1, CYP61A1, CYP61A1
NCYC1147	CYP56A1, CYP51F1, CYP61A1
NCYC1151	CYP56A1, CYP61A1, CYP51F1
NCYC1156	CYP51F1, CYP56A3, CYP61A1, CYP61A1, CYP56A1, CYP51F1, CYP51F1
NCYC1159	CYP61A1, CYP56A1, CYP51F1
NCYC1163	CYP51F1, CYP56A1, CYP61A1
NCYC1167	CYP51F1, CYP56A1, CYP61A1
NCYC1171	CYP51F1, CYP51F1, CYP51F1, CYP56A1, CYP61A1
NCYC1175	CYP61A1, CYP61A1, CYP51F1, CYP61A1, CYP56A1
NCYC1179	CYP61A1, CYP56A1, CYP51F1
NCYC1183	CYP56A1, CYP51F1, CYP61A1
NCYC1186	CYP51F1, CYP61A1, CYP56A1
NCYC1187	CYP51F1, CYP56A1, CYP56A1, CYP51F1, CYP61A1, CYP56A1, CYP51F1
NCYC1190	CYP56A1, CYP51F1, CYP61A1
NCYC1199	CYP56A1, CYP61A1, CYP51F1, CYP61A1
NCYC1203	CYP51F1, CYP56A1, CYP61A1
NCYC1210	CYP56A1, CYP51F1, CYP61A1, CYP61A1, CYP61A1, CYP61A1

NCYC1211	CYP51F1, CYP61A1, CYP56A1
NCYC1215	CYP56A1, CYP61A1, CYP61A1, CYP61A1, CYP51F1
NCYC1218	CYP56A1, CYP61A1, CYP51F1
NCYC1221	CYP51F1, CYP56A1, CYP61A1
NCYC1228	CYP51F1, CYP61A1, CYP51F1, CYP51F1, CYP56A1
NCYC1235	CYP61A1, CYP56A1, CYP61A1, CYP51F1, CYP61A1
NCYC1239	CYP51F1, CYP56A3, CYP61A1, CYP61A1, CYP51F1, CYP56A1
NCYC1240	CYP51F1, CYP61A1, CYP56A1
NCYC1243	CYP61A1, CYP61A1, CYP61A1, CYP56A1, CYP61A1, CYP51F1
NCYC1245	CYP61A1, CYP61A1, CYP61A1, CYP56A1, CYP51F1
NCYC1260	CYP51F1, CYP56A1, CYP61A1
NCYC1264	CYP56A1, CYP51F1, CYP61A1
NCYC1270	CYP61A1, CYP56A1, CYP51F1
NCYC1274	CYP56A1, CYP61A1, CYP51F1
NCYC1277	CYP51F1, CYP61A1
NCYC1280	CYP61A1, CYP56A1, CYP51F1
NCYC1283	CYP61A1, CYP51F1, CYP56A1, CYP61A1, CYP61A1, CYP61A1
NCYC1286	CYP51F1, CYP56A1, CYP61A1, CYP61A1
NCYC1289	CYP56A1, CYP61A1, CYP51F1
NCYC1292	CYP56A1, CYP51F1, CYP56A1, CYP61A1, CYP56A1, CYP56A1
NCYC1298	CYP61A1, CYP56A1, CYP51F1
NCYC1308	CYP61A1, CYP51F1, CYP56A1, CYP61A1
NCYC1311	CYP61A1, CYP51F1, CYP61A1, CYP61A1, CYP56A1
NCYC1314	CYP56A7, CYP61A1, CYP51F1, CYP5252A1
NCYC1315	CYP51F1, CYP56A1, CYP61A1
NCYC1318	CYP51F1, CYP5252A1, CYP56A7, CYP61A1
NCYC1321	CYP51F1, CYP56A1, CYP61A1
NCYC1337	CYP61A1, CYP51F1, CYP56A1
NCYC1339	CYP51F1, CYP56A1, CYP61A1
NCYC1363	CYP52C3, CYP501A1, CYP52A25, CYP56E2, CYP5217A1, CYP61A2, CYP52A21, CYP52A25, CYP52A25, CYP52A23, CYP52A22
NCYC1368	CYP56A7, CYP51F1, CYP5251A1, CYP5252A1, CYP61A1
NCYC1369	CYP51F1, CYP52A33, CYP61A2, CYP52A39, CYP52A13, CYP52A23, CYP52A40, CYP5217A4, CYP56E4, CYP52A40, CYP52A22, CYP52A27, CYP52A23, CYP52A47, CYP52A18, CYP52A27
NCYC1384	CYP51F1, CYP5033A1, CYP540A1, CYP53C2, CYP, CYP61A1
NCYC1388	CYP51F1
NCYC1389	CYP61A1, CYP504G1, CYP51F1
NCYC1393	CYP61A2, CYP5217A1, CYP52A1, CYP52A2, CYP52B1, CYP56E1, CYP51F1, CYP52A8, CYP52A6, CYP52A7, CYP52C1
NCYC1398	CYP51F1, CYP61A1
NCYC1400	CYP51F1, CYP51F1, CYP61A1, CYP61A1, CYP56A1, CYP61A1

NCYC1401 CYP51F1, CYP53B2, CYP5139C1, CYP5139C1, CYP61A1, CYP5221A1

NCYC1405 CYP51F1, CYP61A1, CYP504G1

NCYC1406 CYP61A1, CYP51F1, CYP56A1

NCYC1407 CYP51F1, CYP56A1, CYP61A1

NCYC1408 CYP56A1, CYP61A1, CYP51F1

NCYC1409 CYP56A1, CYP51F1, CYP61A1

NCYC1410 CYP61A1, CYP56A1, CYP51F1

NCYC1411 CYP56A1, CYP51F1, CYP61A1, CYP51F1, CYP61A1, CYP56A3

NCYC1412 CYP61A1, CYP51F1, CYP56A1, CYP56A3, CYP51F1, CYP61A1

NCYC1413 CYP61A1, CYP56A1, CYP51F1

NCYC1414 CYP56A1, CYP51F1, CYP61A1

NCYC1415 CYP61A1, CYP51F1, CYP56A1

NCYC1416 CYP51F1, CYP56A1, CYP61A1

NCYC1417 CYP51F1, CYP61A1, CYP56A2

NCYC1424 CYP61A1, CYP5252A1, CYP51F1, CYP56A7

NCYC1425 CYP61A1, CYP51F1, CYP5252A1, CYP51F1, CYP51F1, CYP51F1, CYP5252A1, CYP5252A1, CYP56A7

NCYC1426 CYP61A1, CYP5252A1, CYP56A7, CYP51F1

NCYC1427 CYP56A1, CYP51F1, CYP61A1

NCYC1429 CYP56A7, CYP56A7, CYP61A1, CYP51F1, CYP5252A1

NCYC1431 CYP61A1, CYP56A3, CYP51F1

NCYC1441 CYP61A1, CYP5252A1, CYP5252A1, CYP51F1, CYP56A7, CYP5252A1

NCYC1444 CYP61A1, CYP61A1, CYP61A1, CYP56A1, CYP51F1

NCYC1449 CYP61A1, CYP52N1, CYP52E1, CYP52N1, CYP51F1, CYP52E1, CYP52E1, CYP52N1, CYP52N1, CYP52N1, CYP52E1, CYP52E1, CYP52E1, CYP52M1

NCYC1462 CYP51F1, CYP61A1, CYP61A1, CYP5139B1, CYP504G1

NCYC1466 CYP501A1, CYP61A2, CYP52A25, CYP52A21, CYP52C3, CYP52A22, CYP52A23, CYP56E2, CYP5217A1

NCYC1467 CYP501A1, CYP56E2, CYP52A25, CYP52A21, CYP52C3, CYP61A2, CYP52A25, CYP52A23, CYP52A22, CYP52A25, CYP5217A1

NCYC1468 CYP56E2, CYP52A25, CYP61A2, CYP5217A1, CYP52A21, CYP52C3, CYP52A23, CYP52A22, CYP52A23, CYP52A22, CYP52A22, CYP52A23, CYP501A1, CYP51F1

NCYC1469 CYP501A1, CYP52A25, CYP52A22, CYP52A23, CYP5217A1, CYP61A2, CYP52A22, CYP52A23, CYP56E2, CYP52A23, CYP52A22, CYP52A21, CYP52C3

NCYC1470 CYP501A1, CYP52A21, CYP52A25, CYP52A22, CYP52A23, CYP5217A1, CYP5217A1, CYP52A25, CYP52A25, CYP52A25, CYP52C3, CYP56E2, CYP5217A1, CYP61A2

NCYC1471 CYP52A22, CYP52A23, CYP501A1, CYP52A21, CYP5217A1, CYP52A25, CYP52C3, CYP56E2, CYP61A2

NCYC1472	CYP501A1, CYP52A25, CYP61A2, CYP5217A1, CYP5217A1, CYP52A21, CYP56E2, CYP52C3, CYP5217A1, CYP52A23, CYP52A22
NCYC1473	CYP501A1, CYP61A2, CYP52A25, CYP52A25, CYP52C3, CYP52A25, CYP52A23, CYP52A22, CYP56E2, CYP5217A1, CYP52A21
NCYC1474	CYP56A4, CYP61A1, CYP61A1, CYP61A1, CYP51F1
NCYC1477	CYP56A4, CYP61A1, CYP51F1
NCYC1495	CYP56A1, CYP51F1, CYP61A1
NCYC1496	CYP51F1, CYP61A1, CYP56A1
NCYC1497	CYP56A1, CYP51F1, CYP61A1
NCYC1498	CYP56A2, CYP61A1, CYP51F1
NCYC1515	CYP61A1, CYP51F1
NCYC1521	CYP51F1, CYP56A1, CYP61A1
NCYC1529	CYP51F1, CYP56A1, CYP61A1
NCYC1536	CYP5216A3, CYP5215A1, CYP5139B2, CYP51F1
NCYC1548	CYP5252A1, CYP51F1, CYP56A7, CYP5251A1, CYP61A1
NCYC1553	CYP61A1, CYP56A1, CYP51F1, CYP61A1, CYP61A1
NCYC1554	CYP61A1, CYP56A1, CYP51F1, CYP61A1
NCYC1555	CYP501D1, CYP51F1, CYP61A1, CYP5217A7, CYP56E1
NCYC1556	CYP61A1, CYP501D1, CYP56E1, CYP5217A7, CYP51F1
NCYC1557	CYP501D1, CYP56E1, CYP61A1, CYP51F1, CYP5217A7
NCYC1558	CYP51F1, CYP61A1, CYP56E1, CYP5217A7, CYP501D1
NCYC1563	CYP501D1, CYP56E1, CYP5217A7, CYP51F1, CYP61A1
NCYC1572	CYP56A1, CYP61A1, CYP51F1
NCYC1573	CYP56A1, CYP56A1, CYP61A1, CYP51F1, CYP61A1
NCYC1591	CYP501D1, CYP56E1, CYP5217A7, CYP61A1, CYP51F1
NCYC1592	CYP61A1, CYP56A1, CYP56A1, CYP51F1, CYP61A1
NCYC1603	CYP61A1, CYP56A1, CYP51F1
NCYC1606	CYP61A1, CYP61A1, CYP51F1, CYP501D1, CYP56A1, CYP56E1, CYP5217A7, CYP51F1
NCYC1645	CYP51F1, CYP61A1, CYP5139C1, CYP51F1, CYP5221A1, CYP5139C1, CYP5139C1, CYP5221A1, CYP53B2, CYP5139C1, CYP53B2
NCYC1646	CYP501D1, CYP61A1, CYP51F1
NCYC1647	CYP5221A1, CYP5139C1, CYP53B2, CYP61A1, CYP51F1, CYP5139C1, CYP5139C1
NCYC1648	CYP5139C1, CYP5139C1, CYP53B2, CYP5221A1
NCYC1649	CYP501D1, CYP61A1, CYP51F1
NCYC1650	CYP5139C1, CYP5139C1, CYP53B2, CYP61A1, CYP5221A1, CYP51F1
NCYC1651	CYP501D1, CYP61A1, CYP51F1
NCYC1653	CYP5252A1, CYP56E6, CYP501D1
NCYC1656	CYP56A12

NCYC1659	CYP53B2, CYP5139C1, CYP51F1
NCYC1660	CYP5139C1, CYP5221A1, CYP5221A1, CYP51F1, CYP5139C1, CYP5139C1, CYP53B2
NCYC1673	CYP61A1, CYP56A1, CYP5217A7, CYP501D1, CYP56E1, CYP61A1, CYP51F1, CYP51F1
NCYC1681	CYP61A1, CYP51F1, CYP56A1, CYP56A1, CYP56A1
NCYC1765	CYP56A1, CYP61A1, CYP51F1
NCYC1766	CYP61A1, CYP51F1, CYP56A1
NCYC2258	CYP56A11, CYP51F1, CYP5252A1, CYP61A1
NCYC2265	CYP56A7, CYP61A1, CYP61A1, CYP51F1, CYP61A1, CYP5252A1
NCYC2307	CYP61A1, CYP61A1, CYP61A1, CYP56A4, CYP51F1
NCYC2321	CYP52A29, CYP5217A4, CYP52A31, CYP56D1, CYP5251A1, CYP5217A2, CYP56D1, CYP61A1, CYP52A29, CYP52A44, CYP501B1, CYP51F1
NCYC2322	CYP56D1, CYP56D1, CYP52A29, CYP52A31, CYP61A1, CYP51F1, CYP501B1, CYP5217A4, CYP5217A2, CYP52A40, CYP52A29, CYP5251A1
NCYC2396	CYP52A40, CYP52A30, CYP52A29, CYP5217A2, CYP61A1, CYP56D1, CYP501B1, CYP51F1
NCYC2397	CYP51F1, CYP56A1, CYP61A1
NCYC2400	CYP52E1, CYP51F1, CYP52N1, CYP61A1
NCYC2401	CYP61A1, CYP51F1, CYP56A1, CYP61A1
NCYC2402	CYP51F1, CYP56A1, CYP61A1
NCYC2403	CYP51F1, CYP61A1, CYP51F1, CYP56A12, CYP56A2
NCYC2423	CYP5217A1, CYP52A6, CYP61A2, CYP56E1, CYP52C1, CYP52A1, CYP52A2, CYP52B1, CYP52A8, CYP51F1, CYP52A7
NCYC2424	CYP56A1, CYP61A1, CYP52N1, CYP51F1
NCYC2431	CYP61A1, CYP51F1, CYP52E1, CYP52N1
NCYC2432	CYP56A1, CYP61A1, CYP51F1
NCYC2433	CYP61A1, CYP51F1, CYP56A11
NCYC2435	CYP52E1, CYP51F1, CYP61A1
NCYC2439	CYP5221A1, CYP51F1
NCYC2440	CYP5221A1, CYP51F1
NCYC2442	CYP51F1, CYP5139B1, CYP5216A2
NCYC2449	CYP61A1, CYP501D1, CYP51F1, CYP501D1
NCYC2450	CYP61A1, CYP61A1, CYP61A1, CYP51F1, CYP51F1, CYP56A3, CYP56A3, CYP51F1
NCYC2458	CYP52C1, CYP56E1, CYP5217A1, CYP52A6, CYP61A2, CYP52A7, CYP51F1, CYP52A2, CYP52A1, CYP52B1, CYP52A8
NCYC2471	CYP52A39, CYP504G1, CYP52A41, CYP5217A4, CYP501C1, CYP52A39, CYP61A1, CYP51F1, CYP52A42, CYP52A40
NCYC2473	CYP51F1, CYP56A3, CYP61A1
NCYC2474	CYP5139B1.

NCYC2479 CYP5139B1, CYP5139B1, CYP504ANeosartorya, CYP53A4, CYP51F1, CYP5139B2, CYP5139B1.

NCYC2480 CYP56D1, CYP52A40, CYP5217A2, CYP52A29, CYP61A1, CYP51F1, CYP52A31

NCYC2483 CYP61A1, CYP51F1, CYP56A3

NCYC2484 CYP61A1, CYP504ANeosartorya, CYP53A4, CYP5139B1., CYP5139B1, CYP5139B2, CYP505D6, CYP51F1, CYP505D8, CYP5139B1, CYP505D8

NCYC2486 CYP52A29, CYP5217A2, CYP61A1, CYP56D1, CYP5217A2, CYP5217A2, CYP52A29, CYP61A1, CYP5217A2, CYP52A37, CYP56D1, CYP52A44, CYP5217A2, CYP52A29, CYP5217A2, CYP501B1, CYP5217A2, CYP501D1, CYP52A29, CYP5217A2, CYP51F1, CYP52A44, CYP61A1, CYP5217A2, CYP51F1

NCYC2489 CYP5251A1, CYP61A1, CYP56A3, CYP51F1

NCYC2491 CYP52A29, CYP51F1, CYP51F1, CYP5217A4, CYP52A31, CYP61A1, CYP56D1, CYP51F1, CYP56D1, CYP52A40, CYP5217A2, CYP5217A2, CYP61A1, CYP5217A2, CYP52A29, CYP51F1, CYP52A44, CYP52A29, CYP52A29, CYP5217A2

NCYC2492 CYP5217A2, CYP5217A2, CYP5217A2, CYP5217A2, CYP5217A2, CYP5217A2, CYP52A29, CYP52A29, CYP52A29, CYP501B1, CYP501B1, CYP51F1, CYP51F1, CYP5217A2, CYP5217A2, CYP51F1, CYP51F1, CYP5217A2, CYP52A29, CYP56D1, CYP56D1, CYP56D1, CYP501B1, CYP5217A2, CYP61A1, CYP52A39

NCYC2499 CYP51F1, CYP52H6, CYP504A17, CYP51F1, CYP56B3, CYP61A1

NCYC2508 CYP5252A1, CYP61A1, CYP56A11, CYP51F1

NCYC2509 CYP505D6, CYP51F1, CYP5139B1., CYP5139B2, CYP504ANeosartorya, CYP53A4, CYP5139B1

NCYC2510 CYP504A18, CYP5152A2, CYP5139B1, CYP505A20, CYP5139B1, CYP51F1, CYP5139B1, CYP5139B1, CYP53A4, CYP5139B1, CYP61A1, CYP5139B1.

NCYC2513 CYP51F1, CYP61A1, CYP56A3

NCYC2515 CYP53A4, CYP504A8, CYP5139B1., CYP505D6, CYP5139B1, CYP505D6, CYP61A1, CYP505C2, CYP51F1

NCYC2516 CYP51F1, CYP61A1, CYP504G1

NCYC2517 CYP61A1, CYP51F1, CYP56A1

NCYC2521 CYP51F1, CYP61A1, CYP56D1, CYP52A29, CYP52A29, CYP51F1, CYP52A44, CYP56D1, CYP5217A2, CYP501B1, CYP52A29, CYP61A1, CYP5217A2, CYP61A1, CYP51F1, CYP56A3, CYP5217A4, CYP5217A7

NCYC2529 CYP52A29, CYP5217A4, CYP51F1, CYP501B1, CYP52A44, CYP5217A7, CYP52A29, CYP56D1, CYP5251A1, CYP61A1, CYP56D1, CYP51F1, CYP51F1, CYP51F1, CYP5217A2, CYP61A1, CYP5217A2

NCYC2542	CYP51F1, CYP56E6, CYP61A1, CYP56A3, CYP61A1, CYP61A1, CYP61A1, CYP51F1
NCYC2547	CYP56E2, CYP51F1, CYP51F1
NCYC2558	CYP51F1
NCYC2559	CYP56A7, CYP5252A1, CYP5251A1, CYP61A1, CYP51F1
NCYC2560	CYP56A3, CYP61A1, CYP51F1
NCYC2568	CYP56A3, CYP61A1, CYP51F1
NCYC2569	CYP51F1
NCYC2572	CYP52A45, CYP52A46, CYP52A47, CYP5217A5, CYP52A44, CYP52A43, CYP51F1
NCYC2577	CYP56A4, CYP61A1, CYP51F1
NCYC2578	CYP51F1, CYP61A1, CYP56A3
NCYC2579	CYP52A31, CYP61A1, CYP5217A2, CYP52A30, CYP51F1, CYP56D1
NCYC2580	CYP5217A2, CYP501B1, CYP61A1, CYP52A29, CYP5217A2, CYP56D1, CYP51F1, CYP5217A2, CYP501B1, CYP501B1, CYP51F1, CYP5217A2, CYP5251A1, CYP52A40, CYP52A29
NCYC2581	CYP61A1, CYP53B2, CYP51F1
NCYC2587	CYP51F1, CYP61A1, CYP56A1
NCYC2592	CYP56A1, CYP61A1, CYP61A1, CYP51F1, CYP61A1
NCYC2597	CYP51F1, CYP501D1, CYP56E1, CYP56E1, CYP51F1, CYP51F1, CYP61A1, CYP61A1, CYP51F1, CYP56E1, CYP56E1, CYP61A1
NCYC2599	CYP53B2, CYP5139C1, CYP5139C1, CYP53B2, CYP51F1, CYP5221A1
NCYC2600	CYP51F1, CYP56A2, CYP61A1
NCYC2602	CYP52H6, CYP504A17, CYP56B3, CYP51F1, CYP61A1, CYP51F1
NCYC2605	CYP53B2, CYP5139C1, CYP51F1, CYP63A3, CYP5221A1, CYP5221A1, CYP61A1
NCYC2620	CYP52N1
NCYC2623	CYP61A1, CYP51F1, CYP53C2
NCYC2628	CYP51F1, CYP61A1, CYP5139B1
NCYC2629	CYP61A1, CYP51F1, CYP56A3
NCYC2636	CYP51F1, CYP52E1, CYP52F11, CYP61A1, CYP504E2, CYP52A49, CYP52E1, CYP548P1, CYP52E1
NCYC2637	CYP52E1, CYP51F1, CYP52N1, CYP61A1, CYP52N1
NCYC2638	CYP53A8, CYP505D6, CYP5139B1, CYP5139B1, CYP5139B1, CYP61A1, CYP5139B1, CYP51F1, CYP504A12, CYP5139B1
NCYC2644	CYP51F1, CYP56A11, CYP61A1
NCYC2658	CYP61A1, CYP51F1
NCYC2665	CYP61A1, CYP53A8, CYP5139B1, CYP504A12, CYP5139B1, CYP51F1, CYP505D6, CYP5139B1, CYP5139B1., CYP5139B1, CYP55A1
NCYC2666	CYP5221A1, CYP5139C1, CYP53B2, CYP5139C1, CYP51F1, CYP61A1

NCYC2670 CYP501A3, CYP52A25, CYP52A26, CYP56E3, CYP61A1, CYP51F1, CYP52A27, CYP52A28, CYP52C4, CYP5217A1

NCYC2675 CYP51F1, CYP61A1, CYP501D1, CYP51F1, CYP51F1, CYP51F1, CYP61A1, CYP61A1, CYP56E1, CYP56E1, CYP56E1, CYP56E1

NCYC2677 CYP51F1

NCYC2683 CYP51F1, CYP51F1, CYP61A1, CYP61A1, CYP56A1

NCYC2688 CYP51F1, CYP56A1, CYP61A1

NCYC2693 CYP61A1, CYP56A4, CYP51F1

NCYC2695 CYP51F1, CYP61A1, CYP56A1

NCYC2698 CYP61A1, CYP51F1

NCYC2701 CYP51F1, CYP56A2, CYP61A1

NCYC2702 CYP51F1

NCYC2703 CYP61A1, CYP51F1

NCYC2712 CYP51F1, CYP5139B1, CYP5216A2, CYP51F1

NCYC2726 CYP52A12, CYP61A1, CYP56E4, CYP51F1, CYP52A13, CYP52A52, CYP52A23, CYP52A40, CYP52A39, CYP52A13, CYP52A23

NCYC2729 CYP56A2, CYP61A1, CYP51F1

NCYC2733 CYP51F1, CYP61A1, CYP56A1

NCYC2737 CYP61A1, CYP51F1, CYP56A1, CYP61A1, CYP61A1

NCYC2739 CYP61A1, CYP51F1

NCYC2741 CYP5217A2, CYP51F1, CYP56A3, CYP52A29, CYP61A1, CYP56D1, CYP5251A1, CYP61A1, CYP56A3, CYP51F1, CYP52A29, CYP52A31, CYP51F1, CYP61A1, CYP61A1, CYP5217A2, CYP61A1, CYP52A40, CYP52A40, CYP51F1, CYP56D1, CYP52A40, CYP5251A1, CYP52A29, CYP52A29, CYP5251A1, CYP5251A1, CYP52A29, CYP52A29, CYP61A1, CYP51F1, CYP56D1, CYP56D1, CYP52A40, CYP51F1

NCYC2742 CYP56A7, CYP5252A1, CYP51F1, CYP5251A1, CYP61A1

NCYC2745 CYP52E1, CYP52M1, CYP52M1, CYP52E1, CYP61A1, CYP52M1, CYP51F1, CYP52E1, CYP52A39

NCYC2746 CYP501D1, CYP51F1, CYP52A40, CYP5217A4, CYP52A39, CYP61A1

NCYC2748 CYP51F1, CYP53A4, CYP505D6, CYP61A1, CYP505C2, CYP5139B1, CYP5139B1, CYP505D6

NCYC2753 CYP52A29, CYP5217A2, CYP51F1, CYP51F1, CYP5217A2, CYP51F1, CYP51F1, CYP501B1, CYP61A1

NCYC2754 CYP56A12, CYP51F1

NCYC2775 CYP5252A1, CYP51F1, CYP61A1, CYP56A7

NCYC2776 CYP56A1, CYP51F1, CYP61A1

NCYC2777 CYP51F1, CYP56A1, CYP61A1

NCYC2778 CYP51F1, CYP56A1, CYP61A1

NCYC2779 CYP51F1, CYP61A1, CYP56A1

NCYC2780 CYP56A1, CYP51F1, CYP61A1

NCYC2786	CYP56E1, CYP52A43, CYP52A30, CYP52A39, CYP52A43, CYP61A1, CYP51F1, CYP5217A7, CYP501D1, CYP52A44
NCYC2789	CYP51F1, CYP61A1, CYP56A2
NCYC2790	CYP56A7, CYP61A1, CYP5252A1, CYP5251A1, CYP51F1
NCYC2791	CYP61A1, CYP56A7, CYP56A7, CYP56A7, CYP51F1, CYP5252A1
NCYC2797	CYP61A1, CYP51F1, CYP56A7, CYP5251A1, CYP5252A1
NCYC2798	CYP51F1, CYP56A1, CYP61A1
NCYC2808	CYP5033A1, CYP61A1, CYP, CYP, CYP53C2, CYP504A8, CYP51F1, CYP, CYP56A1, CYP61A1, CYP51F1
NCYC2809	CYP61A1, CYP56A3, CYP51F1, CYP56A1, CYP51F1, CYP61A1
NCYC2827	CYP56A2, CYP61A1, CYP51F1
NCYC2833	CYP51F1
NCYC2853	CYP51F1, CYP61A1
NCYC2855	CYP61A1, CYP51F1, CYP56A1
NCYC2866	CYP51F1, CYP52A23, CYP56E4, CYP52A13, CYP52A39, CYP52A39, CYP61A1, CYP52A12, CYP52A39, CYP52A14, CYP52A40, CYP52A23
NCYC2875	CYP56A11, CYP56A11, CYP61A1, CYP61A1, CYP61A1, CYP5252A1, CYP51F1, CYP56A11
NCYC2878	CYP56A4, CYP61A1, CYP51F1
NCYC2885	CYP61A1, CYP5217A2, CYP5217A7, CYP56A3, CYP5251A1, CYP61A1, CYP51F1, CYP5217A2, CYP56D1, CYP51F1, CYP5251A1, CYP52A40, CYP52A29, CYP51F1, CYP56A3, CYP61A1
NCYC2886	CYP56A7, CYP51F1, CYP5252A1, CYP61A1, CYP5252A1, CYP5252A1
NCYC2887	CYP51F1, CYP61A1, CYP5252A1, CYP5252A1, CYP5252A1, CYP56A7
NCYC2888	CYP51F1, CYP61A1, CYP56A2
NCYC2889	CYP61A1, CYP56A3, CYP51F1
NCYC2890	CYP61A1, CYP56A2, CYP51F1
NCYC2897	CYP51F1, CYP61A1, CYP56A5
NCYC2898	CYP61A1, CYP51F1, CYP56A4
NCYC2899	CYP52E1, CYP52N1, CYP61A1, CYP51F1
NCYC2900	CYP52M1, CYP56A1, CYP61A1, CYP51F1, CYP51F1, CYP52N1, CYP61A1
NCYC2903	CYP505H1, CYP53A4, CYP61A1, CYP505A20, CYP, CYP505D8, CYP5139B1, CYP5139B1., CYP51F1, CYP504A9
NCYC2907	CYP61A1, CYP5252A1, CYP5252A1, CYP5252A1, CYP56A7, CYP51F1
NCYC2908	CYP61A1, CYP52N1, CYP51F1, CYP52N1, CYP52M1, CYP61A1, CYP52E1, CYP56A1, CYP51F1, CYP52E1, CYP52E1, CYP52E1, CYP52N1, CYP52E1

NCYC2913	CYP52A39, CYP51F1, CYP52M1, CYP52E1, CYP52M1, CYP52M1, CYP61A1, CYP52E1, CYP52E1
NCYC2925	CYP51F1, CYP51F1, CYP56A1, CYP61A1, CYP51F1, CYP61A1, CYP61A1, CYP56A1, CYP56A1, CYP51F1
NCYC2927	CYP51F1, CYP61A1, CYP56A1, CYP61A1, CYP51F1, CYP61A1
NCYC2931	CYP61A1, CYP56A7, CYP5252A1, CYP51F1
NCYC2932	CYP61A1, CYP51F1, CYP56A1, CYP61A1, CYP56A1, CYP56A1
NCYC2933	CYP56A1, CYP56A1, CYP51F1, CYP61A1
NCYC2934	CYP51F1, CYP56A1, CYP61A1
NCYC2935	CYP51F1, CYP51F1, CYP61A1, CYP56A1, CYP56A1
NCYC2945	CYP56A1, CYP51F1, CYP61A1
NCYC2947	CYP61A1, CYP51F1, CYP56A1
NCYC2948	CYP51F1, CYP61A1, CYP56A1
NCYC2956	CYP51F1, CYP56A7, CYP5251A1, CYP61A1, CYP5252A1
NCYC2963	CYP5139B1, CYP51F1
NCYC2965	CYP5217A4, CYP52A39, CYP52A44, CYP61A1, CYP501D1, CYP51F1
NCYC2966	CYP61A1, CYP53A4, CYP5139B1, CYP505D6, CYP505D6, CYP504ANeosartorya
NCYC2967	CYP56A1, CYP61A1, CYP51F1
NCYC2974	CYP56A1, CYP51F1, CYP61A1
NCYC2976	CYP5252A1, CYP61A1, CYP56A10, CYP51F1
NCYC2980	CYP51F1, CYP56A7, CYP5251A1, CYP5252A1, CYP61A1
NCYC2981	CYP5252A1, CYP61A1, CYP56A7, CYP51F1, CYP5251A1
NCYC2990	CYP61A1, CYP51F1, CYP504G1
NCYC2991	CYP51F1, CYP56A2, CYP61A1
NCYC2995	CYP51F1, CYP56A2, CYP61A1
NCYC2999	CYP56A1, CYP61A1, CYP51F1
NCYC3000	CYP56A1, CYP61A1, CYP51F1
NCYC3001	CYP61A1, CYP56A1, CYP51F1
NCYC3008	CYP51F1, CYP61A1, CYP53C2, CYP505A20, CYP5139B1, CYP62C2
NCYC3013	CYP51F1
NCYC3020	CYP51F1, CYP61A1, CYP56A1
NCYC3021	CYP52A33, CYP51F1, CYP56A1, CYP61A1
NCYC3022	CYP51F1, CYP56A1, CYP61A1
NCYC3024	CYP56D1, CYP5217A7, CYP52A29, CYP52A40, CYP5251A1, CYP52A40, CYP52A29, CYP52A29, CYP5217A2, CYP52A29, CYP52A29, CYP52A29, CYP52A29, CYP5217A2, CYP5217A2, CYP5217A2, CYP5217A2, CYP5251A1, CYP5251A1, CYP61A1, CYP61A1, CYP52A40, CYP52A40, CYP51F1, CYP52A29, CYP61A1, CYP61A1, CYP56D1, CYP5251A1, CYP501B1
NCYC3025	CYP51F1, CYP56A1, CYP61A1
NCYC3026	CYP51F1, CYP56A1, CYP56A1, CYP61A1, CYP56A1

NCYC3027	CYP51F1, CYP52A30, CYP52A44, CYP5217A7, CYP52A39, CYP52A43, CYP61A1, CYP56E1, CYP52A43
NCYC3028	CYP51F1, CYP61A1, CYP56A1
NCYC3029	CYP51F1
NCYC3030	CYP51F1, CYP61A1, CYP56A1
NCYC3031	CYP61A1, CYP56A1, CYP51F1
NCYC3032	CYP61A1, CYP56A1, CYP51F1
NCYC3033	CYP51F1, CYP56A1, CYP51F1, CYP51F1, CYP61A1
NCYC3034	CYP51F1, CYP61A1, CYP56A5
NCYC3035	CYP56A1, CYP61A1, CYP61A1, CYP61A1, CYP51F1
NCYC3036	CYP56A1, CYP61A1, CYP61A1, CYP61A1, CYP51F1
NCYC3037	CYP61A1, CYP56A1, CYP51F1
NCYC3038	CYP51F1, CYP61A1, CYP56A1
NCYC3039	CYP56A1, CYP51F1, CYP61A1
NCYC3041	CYP51F1, CYP56A7, CYP5252A1, CYP61A1
NCYC3047	CYP52A31, CYP52A39, CYP5251A1, CYP5217A2, CYP501B1, CYP52A40, CYP56D1, CYP56D1, CYP56D1, CYP52A40, CYP61A1, CYP61A1, CYP61A1, CYP5217A2, CYP5217A2, CYP51F1, CYP51F1, CYP5217A2, CYP5217A2, CYP5217A2, CYP52A40, CYP52A29
NCYC3048	CYP61A1, CYP56A1, CYP51F1
NCYC3051	CYP61A1, CYP51F1, CYP56A1
NCYC3052	CYP51F1, CYP51F1, CYP61A1, CYP51F1, CYP56A1
NCYC3053	CYP61A1, CYP61A1, CYP56A4, CYP61A1, CYP51F1
NCYC3076	CYP56A1, CYP51F1, CYP51F1, CYP51F1, CYP61A1
NCYC3077	CYP56A1, CYP51F1, CYP61A1
NCYC3078	CYP61A1, CYP51F1, CYP56A1
NCYC3080	CYP51F1, CYP56A1, CYP61A1
NCYC3086	CYP51F1, CYP53A4, CYP5139B1., CYP504A8, CYP5139B1, CYP, CYP, CYP5139B1
NCYC3090	CYP51F1, CYP61A1, CYP56A1
NCYC3091	CYP61A1, CYP61A1, CYP56A1, CYP51F1
NCYC3092	CYP51F1, CYP56A1, CYP61A1, CYP61A3, CYP51F1
NCYC3096	CYP52A29, CYP5217A2, CYP5251A1, CYP56D1, CYP61A1, CYP52A29, CYP52A29, CYP52A29, CYP51F1, CYP5217A2, CYP52A29, CYP501B1, CYP52A40
NCYC3104	CYP61A1, CYP5217A2, CYP52A31, CYP52A39, CYP501B1, CYP52A44, CYP51F1, CYP52A29, CYP52A39, CYP52A29, CYP52A30, CYP56D1
NCYC3108	CYP51F1, CYP61A1, CYP56A4
NCYC3109	CYP505A5, CYP61A1
NCYC3114	CYP51F1, CYP61A1, CYP56A1

NCYC3115	CYP52C3, CYP52A25, CYP61A2, CYP52A25, CYP52A22, CYP52A23, CYP5217A1, CYP501A1, CYP52A25, CYP52A21, CYP56E2
NCYC3121	CYP56A1, CYP51F1, CYP51F1, CYP51F1, CYP61A1
NCYC3122	CYP56A1, CYP61A1, CYP51F1
NCYC3123	CYP61A1, CYP56A1, CYP51F1
NCYC3124	CYP61A1, CYP56A1, CYP56A1, CYP56A1, CYP51F1
NCYC3125	CYP51F1, CYP61A1, CYP56A1
NCYC3126	CYP56A1, CYP51F1, CYP61A1
NCYC3127	CYP61A1, CYP51F1, CYP56A1, CYP61A1, CYP61A1, CYP61A1
NCYC3129	CYP52A29, CYP52A44, CYP51F1, CYP52A29, CYP501B1, CYP5217A2, CYP5217A2, CYP52A29, CYP51F1, CYP501B1, CYP51F1, CYP51F1, CYP61A1, CYP61A1, CYP5251A1, CYP56D1, CYP52A39, CYP5251A1, CYP61A1, CYP56D1, CYP56D1, CYP52A39, CYP501B1, CYP5251A1, CYP5217A2, CYP5217A2, CYP52A44
NCYC3133	CYP56A5, CYP61A1, CYP51F1
NCYC3134	CYP51F1, CYP56A1, CYP61A1
NCYC3138	CYP52F10, CYP51F1, CYP61A1, CYP51F1, CYP504A17, CYP56B2
NCYC3141	CYP52A29, CYP56D1, CYP52A40, CYP5217A2, CYP51F1, CYP56A3, CYP51F1, CYP61A1, CYP5217A4
NCYC3146	CYP51F1, CYP61A1, CYP56A1
NCYC3239	CYP52A29, CYP5217A2, CYP52A29, CYP51F1, CYP5217A2, CYP5217A4, CYP52A44, CYP5217A2, CYP61A1, CYP52A29, CYP5217A2, CYP61A1, CYP52A29, CYP51F1, CYP56A3, CYP51F1, CYP61A1, CYP56D1, CYP52A37
NCYC3242	CYP61A1, CYP51F1
NCYC3252	CYP51F1, CYP51F1, CYP53A4, CYP5139B1, CYP504A9, CYP505A20, CYP61A1, CYP505D6, CYP5139B1., CYP5139B1
NCYC3254	CYP61A1, CYP505A5, CYP52E1, CYP61A1, CYP52N1, CYP51F1
NCYC3256	CYP61A1, CYP51F1
NCYC3264	CYP61A1, CYP51F1, CYP56A1, CYP51F1, CYP61A1, CYP56A2
NCYC3265	CYP56A1, CYP61A1, CYP51F1
NCYC3266	CYP51F1, CYP61A1, CYP56A1
NCYC3267	CYP, CYP51F1, CYP, CYP5033A1, CYP61A1, CYP53C2, CYP504A8, CYP
NCYC3272	CYP504A17, CYP51F1, CYP51F1, CYP56B2, CYP61A1, CYP52A38
NCYC3298	CYP51F1, CYP61A1, CYP56A2
NCYC3302	CYP56A1, CYP51F1, CYP61A1
NCYC3303	CYP51F1, CYP56A5, CYP61A1
NCYC3306	CYP51F1, CYP51F1, CYP61A1, CYP56A3, CYP56A1, CYP61A1
NCYC3307	CYP51F1, CYP61A1, CYP56A1
NCYC3309	CYP51F1, CYP61A1
NCYC3311	CYP61A1, CYP51F1, CYP56A1

NCYC3313	CYP51F1, CYP61A1, CYP56A1
NCYC3314	CYP61A1, CYP56A1, CYP51F1
NCYC3315	CYP61A1, CYP56A1, CYP51F1
NCYC3318	CYP61A1, CYP56A1, CYP51F1
NCYC3319	CYP51F1, CYP61A1, CYP56A1
NCYC3324	CYP51F1, CYP56A1, CYP61A1
NCYC3325	CYP56A1, CYP61A1, CYP51F1
NCYC3326	CYP56A1, CYP61A1, CYP61A1, CYP61A1, CYP51F1
NCYC3331	CYP51F1, CYP61A1, CYP56A1, CYP61A1
NCYC3332	CYP52A44, CYP61A1, CYP52A47, CYP5217A5, CYP51F1, CYP52A46, CYP61A1, CYP56A1, CYP52A43, CYP51F1
NCYC3333	CYP61A1, CYP56A1, CYP51F1
NCYC3334	CYP56A1, CYP51F1, CYP61A1
NCYC3338	CYP51F1, CYP51F1, CYP56A1, CYP61A1
NCYC3339	CYP51F1, CYP56A1, CYP61A1
NCYC3340	CYP51F1, CYP61A1, CYP61A1, CYP61A1, CYP61A1, CYP56A1
NCYC3341	CYP51F1, CYP56A1, CYP61A1
NCYC3342	CYP51F1, CYP56A1, CYP61A1
NCYC3343	CYP51F1, CYP56A1, CYP61A1
NCYC3344	CYP5252A1, CYP51F1, CYP56A7, CYP61A1
NCYC3345	CYP61A1, CYP5252A1, CYP51F1, CYP56A10
NCYC3354	CYP56A4, CYP61A1, CYP51F1
NCYC3358	CYP56A4, CYP61A1, CYP61A1, CYP61A1, CYP51F1
NCYC3369	CYP51F1
NCYC3373	CYP52A39, CYP52A39, CYP5217A7, CYP5217A7, CYP52A39, CYP52A39, CYP52A39, CYP52A42, CYP56E4, CYP51F1, CYP52A39, CYP52A39, CYP52A40, CYP52A40, CYP51F1, CYP51F1, CYP56E4, CYP56E4, CYP56E4, CYP52A39, CYP52A39, CYP52A39, CYP52A39, CYP61A2, CYP51F1, CYP5217A7, CYP501D2, CYP52A42, CYP52A42, CYP52A42, CYP52A39, CYP52A40, CYP52A39, CYP52A42
NCYC3378	CYP51F1, CYP56A1, CYP61A1
NCYC3379	CYP61A1, CYP56A2, CYP51F1
NCYC3392	CYP56A5, CYP61A1, CYP51F1
NCYC3393	CYP61A1, CYP51F1
NCYC3396	CYP5252A1, CYP61A1, CYP51F1, CYP56A7
NCYC3397	CYP51F1, CYP56A5, CYP61A1
NCYC3398	CYP52A31, CYP51F1, CYP5251A1, CYP51F1, CYP5217A2, CYP5217A7, CYP5217A2, CYP52A29, CYP51F1

NCYC3400	CYP52A31, CYP52A30, CYP61A1, CYP51F1, CYP52A40, CYP52A40, CYP52A40, CYP52A40, CYP61A1, CYP5217A2, CYP52A39, CYP51F1, CYP5251A1, CYP5217A2, CYP5217A2, CYP5217A2, CYP56D1, CYP5217A2, CYP5217A2, CYP5251A1, CYP5251A1, CYP5251A1, CYP52A39, CYP52A39, CYP61A1, CYP501B1, CYP52A29
NCYC3402	CYP51F1, CYP52A47, CYP52A46, CYP5217A5, CYP52A44, CYP52A43
NCYC3403	CYP56A1, CYP61A1, CYP51F1, CYP51F1, CYP61A1, CYP61A1, CYP51F1
NCYC3406	CYP61A1, CYP61A1, CYP61A1, CYP51F1, CYP56A1
NCYC3407	CYP51F1, CYP56A1, CYP61A1
NCYC3410	CYP51F1
NCYC3414	CYP56A1, CYP51F1, CYP61A1
NCYC3423	CYP51F1, CYP61A1, CYP504G1
NCYC3431	CYP, CYP51F1, CYP53C2, CYP, CYP61A1, CYP504A3, CYP5033A1
NCYC3445	CYP61A1, CYP51F1, CYP56A1
NCYC3447	CYP61A1, CYP51F1, CYP56A1
NCYC3448	CYP51F1, CYP61A1, CYP56A1
NCYC3449	CYP56A1, CYP61A1, CYP51F1
NCYC3451	CYP56A1, CYP61A1, CYP51F1
NCYC3453	CYP56A1, CYP56A1, CYP56A1, CYP61A1, CYP51F1
NCYC3454	CYP61A1, CYP51F1, CYP61A1, CYP56A1, CYP61A1
NCYC3455	CYP51F1, CYP56A1, CYP61A1
NCYC3456	CYP51F1, CYP61A1, CYP56A1
NCYC3457	CYP51F1, CYP61A1, CYP56A1
NCYC3458	CYP51F1, CYP56A1, CYP61A1
NCYC3460	CYP51F1, CYP61A1, CYP56A1
NCYC3461	CYP56A1, CYP61A1, CYP51F1
NCYC3462	CYP51F1, CYP61A1, CYP56A1
NCYC3464	CYP61A1, CYP56A1, CYP51F1
NCYC3465	CYP56A1, CYP61A1, CYP51F1
NCYC3466	CYP51F1, CYP56A1, CYP56A7, CYP51F1, CYP61A1, CYP61A1, CYP5252A1, CYP5252A1, CYP5252A1
NCYC3467	CYP56A1, CYP51F1, CYP61A1
NCYC3469	CYP61A1, CYP56A1, CYP51F1
NCYC3470	CYP51F1, CYP61A1, CYP56A1
NCYC3471	CYP51F1, CYP61A1, CYP56A1
NCYC3472	CYP56A1, CYP61A1, CYP51F1
NCYC3486	CYP61A1, CYP56A1, CYP51F1
NCYC3487	CYP61A1, CYP56A1, CYP61A1, CYP61A1, CYP51F1
NCYC3491	CYP51F1, CYP56A1, CYP51F1, CYP51F1, CYP61A1

NCYC3492	CYP52A38, CYP52A37, CYP52A35, CYP52A32, CYP56E4, CYP52A34, CYP5217A3, CYP56A1, CYP501A4, CYP61A1, CYP52C6, CYP52C5, CYP51F1, CYP61A1, CYP51F1
NCYC3493	CYP51F1, CYP56A1, CYP61A1
NCYC3497	CYP56A1, CYP61A1, CYP61A1, CYP51F1
NCYC3498	CYP61A1, CYP56A1, CYP51F1
NCYC3499	CYP56A1, CYP61A1, CYP56A1, CYP51F1
NCYC3500	CYP52A33, CYP51F1, CYP61A1, CYP56A1
NCYC3501	CYP61A1, CYP51F1, CYP56A5
NCYC3502	CYP504A9
NCYC3506	CYP5217A2, CYP51F1, CYP5217A4, CYP52A44, CYP51F1, CYP56A3, CYP61A1
NCYC3510	CYP61A1, CYP56A1, CYP51F1
NCYC3511	CYP51F1, CYP61A1, CYP56A1, CYP61A1, CYP61A1, CYP61A1, CYP61A1
NCYC3512	CYP52A35, CYP52C6, CYP52A33, CYP52A33, CYP501A4, CYP52A34, CYP52C5, CYP52A32, CYP56A1, CYP61A1, CYP51F1, CYP61A1, CYP5217A3, CYP51F1, CYP52A38, CYP52A37, CYP56E4
NCYC3513	CYP51F1, CYP61A1, CYP56A1
NCYC3514	CYP56A1, CYP51F1, CYP61A1
NCYC3515	CYP51F1, CYP56A1, CYP61A1
NCYC3516	CYP51F1, CYP61A1, CYP56A1
NCYC3519	CYP61A1, CYP56A5, CYP51F1
NCYC3520	CYP51F1, CYP56E2, CYP52C3, CYP52A21, CYP52A25, CYP52A25, CYP61A2, CYP5217A1, CYP52A25, CYP501A1, CYP52A23, CYP52A22
NCYC3521	CYP56A1, CYP61A1, CYP51F1
NCYC3522	CYP61A1, CYP56A1, CYP51F1
NCYC3523	CYP56A1, CYP61A1, CYP61A1, CYP61A1, CYP51F1, CYP61A1
NCYC3526	CYP5139B1, CYP51F1, CYP5216A3, CYP505D6, CYP61A1
NCYC3527	CYP52C3, CYP501A1, CYP5217A1, CYP52A21, CYP56E2, CYP61A2, CYP52A25, CYP52A22, CYP52A23
NCYC3528	CYP61A1, CYP61A1, CYP51F1, CYP56A1
NCYC3529	CYP51F1, CYP56A1, CYP51F1, CYP61A1, CYP56A1, CYP56A1
NCYC3530	CYP51F1, CYP51F1, CYP51F1
NCYC3537	CYP61A1, CYP56A5, CYP51F1
NCYC3539	CYP5139B1, CYP5216A3, CYP505D6, CYP51F1, CYP61A1, CYP51F1, CYP51F1, CYP51F1
NCYC3546	CYP61A1, CYP56A1, CYP51F1, CYP61A1
NCYC3549	CYP61A1, CYP56A1, CYP61A1, CYP51F1, CYP61A1
NCYC3552	CYP61A1, CYP61A1, CYP61A1, CYP56A1, CYP51F1
NCYC3557	CYP56A1, CYP61A1, CYP51F1

NCYC3562	CYP56A1, CYP56A7, CYP5251A1, CYP5252A1, CYP61A1, CYP51F1
NCYC3612	CYP51F1, CYP61A1, CYP56A1
NCYC3630	CYP61A1, CYP56A1, CYP51F1
NCYC3662	CYP56A2, CYP61A1, CYP51F1
NCYC3716	CYP56A4, CYP61A1, CYP51F1, CYP53C2, CYP61A1
NCYC3719	CYP61A1, CYP52A29, CYP5217A2, CYP56D1, CYP5217A2, CYP501B1, CYP5217A2, CYP52A40, CYP52A29, CYP52A39, CYP52A40, CYP52A40, CYP5217A2, CYP5217A2, CYP52A29, CYP52A29, CYP5251A1, CYP5251A1, CYP61A1, CYP61A1, CYP51F1, CYP51F1, CYP52A40, CYP52A39, CYP52A39, CYP61A1, CYP5251A1, CYP5217A7, CYP52A39, CYP56D1
NCYC3724	CYP56A1, CYP61A1, CYP51F1
NCYC3730	CYP61A1, CYP5139B1, CYP53C2, CYP51F1
NCYC3740	CYP51F1, CYP52A42, CYP52A39, CYP5217A4, CYP61A1, CYP52A41, CYP51F1, CYP52A39, CYP52A40
NCYC3745	CYP61A1, CYP51F1
NCYC3751	CYP51F1, CYP61A1
NCYC3752	CYP61A1, CYP504G1, CYP504G1, CYP61A1, CYP61A1, CYP51F1
NCYC3759	CYP52A16, CYP51F1, CYP504A19, CYP51F1, CYP52A44, CYP56C2, CYP61A1
NCYC3776	CYP51F1, CYP56A4, CYP61A1
NCYC3777	CYP51F1, CYP504A17, CYP61A1, CYP52H6, CYP56B3, CYP51F1
NCYC3778	CYP501A1, CYP56E2, CYP61A2, CYP52C3, CYP52A21, CYP52A25, CYP52A23, CYP52A22, CYP5217A1
NCYC3779	CYP61A1, CYP501A1, CYP61A2, CYP52C3, CYP52A22, CYP52A23, CYP51F1, CYP52A23, CYP52A22, CYP52A23, CYP52A22, CYP51F1, CYP5217A1, CYP56E2, CYP52A21, CYP52A25
NCYC3785	CYP52N1, CYP52E1, CYP61A1, CYP51F1
NCYC3788	CYP5252A1, CYP51F1, CYP51F1, CYP61A1, CYP56A10
NCYC3792	CYP51F1, CYP5217A2, CYP5217A2, CYP52A44, CYP51F1, CYP51F1, CYP5217A2, CYP52A29, CYP52A29, CYP52A29, CYP501D1, CYP52A29, CYP52A29, CYP56D1, CYP501B1, CYP61A1, CYP5217A2, CYP61A1, CYP61A1, CYP52A29, CYP56D1, CYP52A40, CYP5217A2, CYP5217A4
NCYC3815	CYP52A43, CYP52A44, CYP52A43, CYP5217A5, CYP61A1, CYP51F1, CYP52A47, CYP52A46, CYP52A45, CYP51F1
NCYC3853	CYP51F1, CYP61A1, CYP56A4
NCYC3870	CYP56D1, CYP501B1, CYP5217A2, CYP56D1, CYP61A1, CYP52A29, CYP51F1, CYP51F1, CYP5217A2, CYP52A29, CYP501D1, CYP5217A2, CYP51F1, CYP61A1, CYP52A44, CYP52A44, CYP5217A2, CYP5217A2, CYP5217A2, CYP52A29
NCYC3877	CYP56A3, CYP51F1, CYP61A1

NCYC3879	CYP51F1, CYP56E6, CYP501D1, CYP51F1, CYP51F1, CYP61A1
NCYC3961	CYP61A1, CYP56A2, CYP51F1
NCYC3963	CYP61A1, CYP51F1, CYP56A2, CYP51F1, CYP61A1
NCYC3964	CYP52A29, CYP52A29, CYP5217A2, CYP5217A2, CYP52A44, CYP61A1, CYP56D1, CYP51F1, CYP5217A2
NCYC3966	CYP51F1, CYP56A4, CYP61A1
NCYC3968	CYP61A1, CYP51F1, CYP56A1
NCYC3983	CYP51F1, CYP62C2, CYP5139B1, CYP53C2, CYP61A1, CYP505D8
NCYC3997	CYP56A1, CYP51F1, CYP51F1, CYP61A1, CYP51F1
NCYC4000	CYP56A4
NCYC4001	CYP61A1, CYP51F1
NCYC4002	CYP51F1, CYP61A1
NCYC4015	CYP61A1, CYP51F1
NCYC4017	CYP51F1, CYP61A1
NCYC4020	CYP51F1, CYP56A3, CYP61A1
NCYC4034	CYP61A1, CYP51F1
NCYC4045	CYP51F1, CYP61A1, CYP56A3, CYP61A1, CYP51F1, CYP56A1
NCYC4063	CYP61A1, CYP51F1, CYP56A1
NCYC4068	CYP51F1, CYP56A1, CYP61A1
NCYC4081	CYP61A1, CYP61A1, CYP61A1, CYP56A1, CYP61A1, CYP61A1, CYP51F1
CBS106	CYP51F1, CYP61A1
CBS109	CYP51F1, CYP56A6
CBS270	CYP61A1, CYP61A1, CYP51F1, CYP504G1, CYP51F1
CBS285	CYP61A1, CYP56A1, CYP51F1
CBS2186	CYP61A1, CYP56A1, CYP56A2, CYP51F1, CYP61A1, CYP51F1
CBS2592	CYP61A1, CYP501D1, CYP51F1, CYP51F1, CYP51F1, CYP56E6
CBS4332	CYP51F1, CYP530A5
CBS4417	CYP61A1, CYP51F1, CYP56A5, CYP56A12, CYP61A1, CYP51F1
CBS4438	CYP61A1, CYP5252A1, CYP51F1, CYP5251A1
CBS6284	CYP61A1, CYP56A1, CYP56A1, CYP51F1
CBS7729	CYP51F1, CYP56E1, CYP5217A7, CYP501D1, CYP61A1
CBS8199	CYP51F1, CYP61A1
CBS8665	CYP56A3, CYP51F1, CYP61A1
CBS8762	CYP56A12, CYP61A1, CYP51F1
CBS8763	CYP51F1, CYP61A1, CYP56A7, CYP5252A1
CBS8765	CYP61A1, CYP51F1, CYP56A5
CBS8778	CYP56A7, CYP51F1, CYP61A1
CBS9961	CYP51F1, CYP61A1, CYP504G1
JCM16988	CYP51F1, CYP61A1, CYP51F1
NRRLY7792	CYP51F1, CYP61A1, CYP61A1, CYP51F1

Table C.1: Table showing the identity of the classifiable putative CYPs found in the NCYC collection, per strain.

D Appendix D

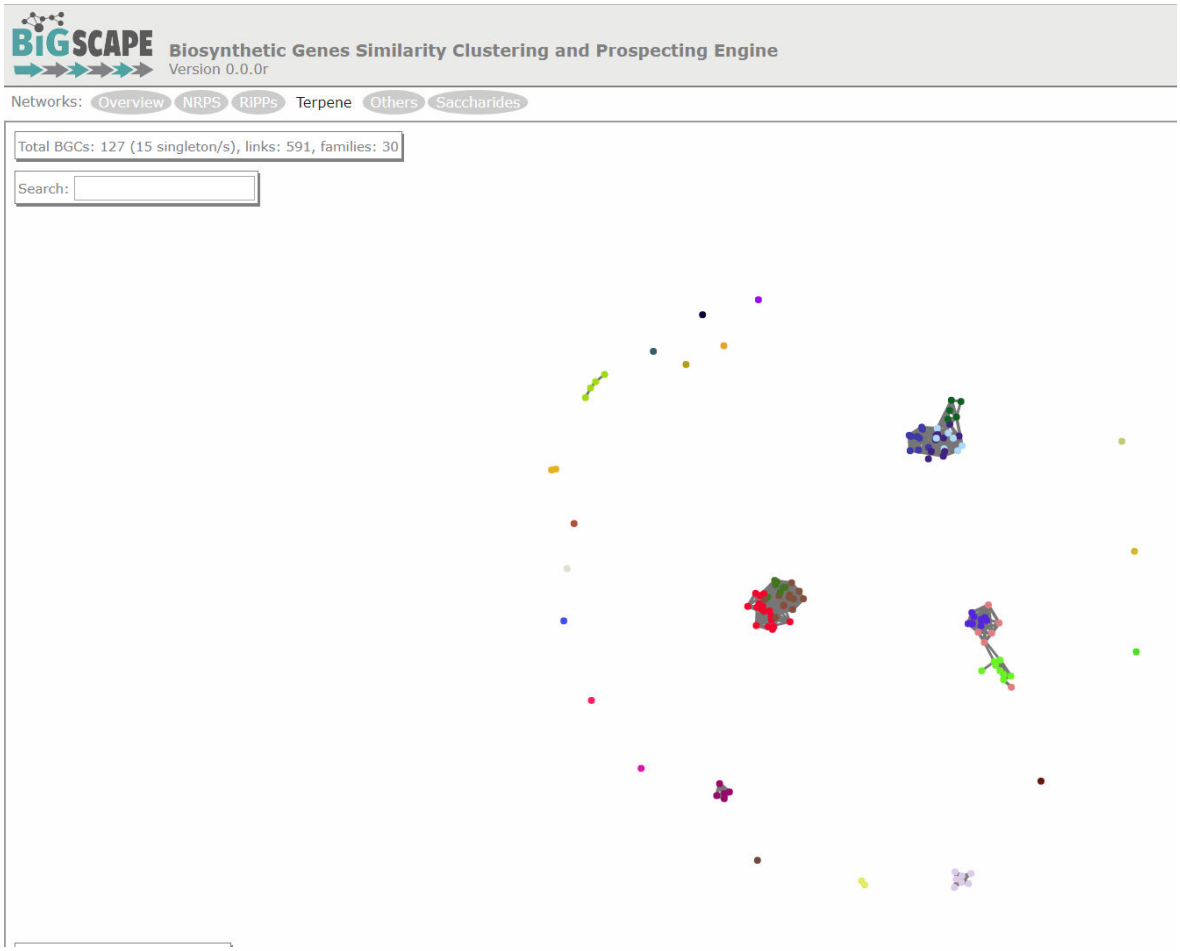

Figure D.1: Output of BiG-SCAPE. Network diagram showing groupings of 127 terpenoid gene clusters (30 families, 15 singletons) found in *Rhodotorula* genomes by antiSMASH. Interactive HTML output available at `https://github.com/chrispyatt/PhData`.

Strain	**Species name**	**BUSCO score (%)**	**No. predicted gene clusters**	**No. gene clusters containing CYP**
NCYC59	Rhodotorula glutinis var. glutinis	3.5	54	3
NCYC60	Rhodotorula glutinis var. glutinis	89.3	9	0
NCYC62	Rhodotorula minuta var. minuta	58.2	0	0
NCYC63	Rhodotorula mucilaginosa	73.8	9	0
NCYC64	Rhodotorula mucilaginosa	74.6	9	0

NCYC65	Rhodotorula mucilaginosa	92.2	14	2
NCYC68	Rhodotorula mucilaginosa	84.8	11	0
NCYC135	Rhodotorula mucilaginosa	90.8	15	2
NCYC138	Rhodotorula aurantiaca	91.3	18	0
NCYC142	Rhodotorula mucilaginosa	72.3	4	0
NCYC154	Rhodotorula glutinis var. glutinis	33.6	1	0
NCYC155	Rhodotorula glutinis var. glutinis	26	2	0
NCYC158	Rhodotorula mucilaginosa	47.3	4	0
NCYC159	Rhodotorula mucilaginosa	56.8	5	0
NCYC162	Rhodotorula glutinis var. glutinis	76.7	3	0
NCYC195	Rhodotorula mucilaginosa	65.7	6	0
NCYC377	Rhodotorula glutinis var. glutinis	81.1	11	0
NCYC502	Rhodotorula graminis	28.3	4	0
NCYC539	Rhodotorula minuta var. minuta	0	0	0
NCYC541	Rhodotorula minuta var. minuta	80.2	8	0
NCYC758	Rhodotorula mucilaginosa	86	11	0
NCYC759	Rhodotorula mucilaginosa	76.3	8	0
NCYC796	Rhodotorula mucilaginosa	92.4	21	2
NCYC797	Rhodotorula mucilaginosa	0	0	0
NCYC844	Rhodotorula minuta var. minuta	69.6	9	1
NCYC845	Rhodotorula minuta var. minuta	58.6	11	0
NCYC930	Rhodotorula minuta var. minuta	91.9	29	3
NCYC931	Rhodotorula minuta var. minuta	92	39	6
NCYC974	Rhodotorula glutinis var. glutinis	90.9	20	0
NCYC1401	Rhodotorula graminis	75.1	7	0
NCYC1645	Rhodotorula mucilaginosa	86.9	12	2
NCYC1646	Rhodotorula mucilaginosa	69.8	7	2
NCYC1647	Rhodotorula mucilaginosa	75.5	3	0
NCYC1648	Rhodotorula mucilaginosa	45.1	4	1
NCYC1649	Rhodotorula mucilaginosa	69.8	9	1
NCYC1650	Rhodotorula mucilaginosa	81.6	7	1
NCYC1651	Rhodotorula mucilaginosa	69.7	9	1
NCYC1660	Rhodotorula mucilaginosa	59.9	9	0
NCYC2439	Rhodotorula glutinis	28.6	1	0
NCYC2440	Rhodotorula glutinis	36.5	2	0
NCYC2581	Rhodotorula minuta var. minuta	90.6	17	1
NCYC2599	Rhodotorula mucilaginosa	58.5	6	1
NCYC2605	Rhodotorula vanillica	97.3	18	1
NCYC2666	Rhodotorula glutinis var. glutinis	77.1	6	1
NCYC2752	Rhodotorula cresolica	NA	NA	NA
NCYC2864	Rhodotorula mucilaginosa	NA	NA	NA
NCYC2873	Rhodotorula creatinovora	NA	NA	NA
NCYC2972	Rhodotorula graminis	NA	NA	NA
NCYC3056	Rhodotorula sp. nov.	NA	NA	NA

NCYC3057	Rhodotorula mucilaginosa	NA	NA	NA
NCYC3072	Rhodotorula laryngis	NA	NA	NA
NCYC3120	Rhodotorula phylloplana	NA	NA	NA
NCYC3401	Rhodotorula graminis	NA	NA	NA
NCYC3411	Rhodotorula mucilaginosa	NA	NA	NA
NCYC3444	Rhodotorula dairenensis	NA	NA	NA
NCYC3504	Rhodotorula mucilaginosa	NA	NA	NA
NCYC3536	Rhodotorula mucilaginosa	49.4	1	0
NCYC3721	Rhodotorula slooffiae	91.3	36	9
NCYC3722	Rhodotorula graminis	93	7	0
NCYC3725	Rhodotorula dairenensis	62.7	4	0
NCYC3735	Rhodotorula mucilaginosa	57.1	4	0
NCYC3772	Rhodotorula mucilaginosa	90.8	14	1
NCYC3775	Rhodotorula mucilaginosa	0	0	0
NCYC3816	Rhodotorula mucilaginosa	95.9	18	1
NCYC3817	Rhodotorula mucilaginosa	95.5	17	1
NCYC3820	Rhodotorula mucilaginosa	93.7	13	1
NCYC3821	Rhodotorula mucilaginosa	95	14	1
NCYC3832	Rhodotorula sp. nov.	68.7	4	0
NCYC3833	Rhodotorula sp. nov.	92.2	26	6
NCYC3834	Rhodotorula laryngis	89.4	24	4
NCYC3835	Rhodotorula sp. nov.	70.3	5	0
NCYC3836	Rhodotorula laryngis	91.8	26	3
NCYC3837	Rhodotorula laryngis	93	32	5
NCYC3838	Rhodotorula laryngis	47.7	2	0
NCYC3867	Rhodotorula mucilaginosa	35.2	0	0
NCYC3872	Rhodotorula mucilaginosa	48.7	2	0

Table D.1: Summary statistics, including BUSCO (genome assembly completeness) results, for the *Rhodotorula* genomes investigated in Chapter 6. See `https://github.com/chrispyatt/PhData` for HTML output files for each strain.

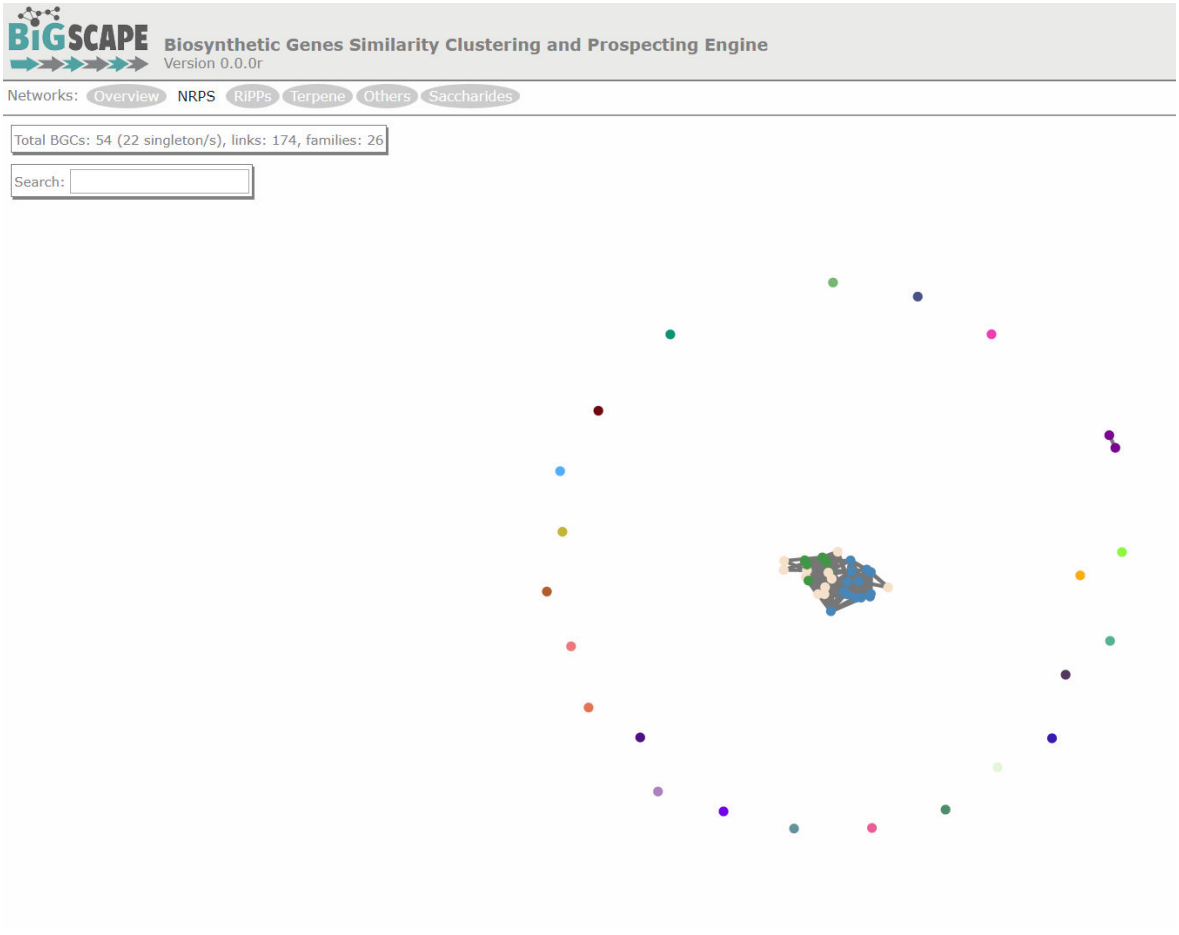

Figure D.2: Output of BiG-SCAPE. Network diagram showing groupings of 54 NRPS gene clusters found (26 families, 22 singletons) in *Rhodotorula* genomes by antiSMASH. Interactive HTML output available at `https://github.com/chrispyatt/PhData`.

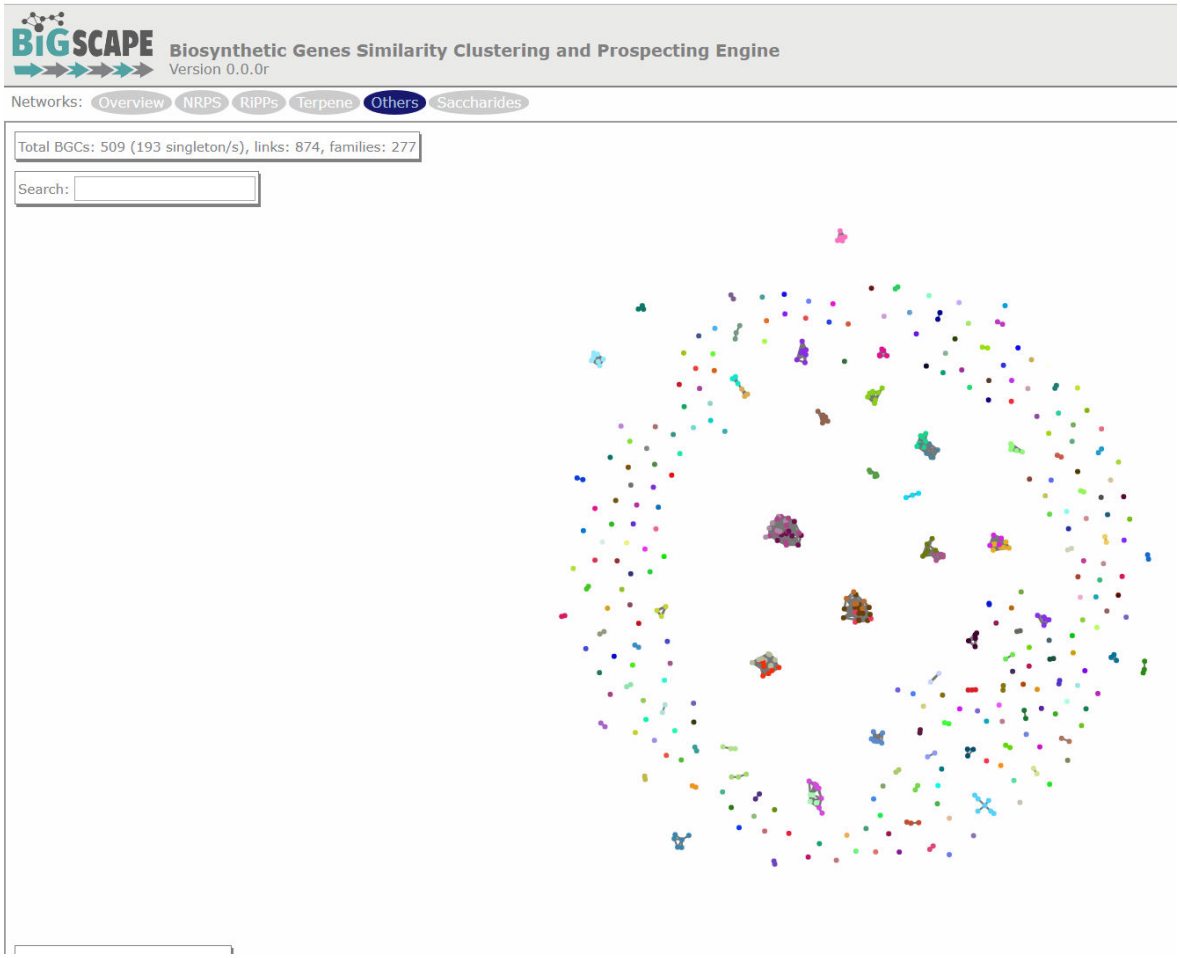

Figure D.3: Output of BiG-SCAPE. Network diagram showing groupings of 509 'others' gene clusters (277 families, 193 singletons; many 'clusterfinder' predictions) found in *Rhodotorula* genomes by antiSMASH. Interactive HTML output available at `https://github.com/chrispyatt/PhData`.