Computing a Yeast Tree of Life

Ann-Marie Keane

A thesis submitted to the University of East Anglia for the degree of Doctor of Philosophy

University of East Anglia
Quadram Institute Bioscience
December 2020

Abstract

This study aimed to compare the results of distinct state-of-the-art phylogenetic tree-building methodologies for a key yeast NGS dataset, with the ultimate goal of establishing a yeast tree of life. Draft genome assemblies of seventy-five species from the Saccharomyces complex, a well-studied group of species of academic and industrial importance, first underwent a stringent quality control process, along with a dataset from an outgroup species. This process uncovered a vast amount of genomic information. New, good quality genome assemblies were introduced for six Saccharomyces complex species and for four strains.

Key genomic differences were found in a quality-controlled subset of this dataset including varying genome sizes (8-29Mbp), coding genome proportions (54-77%) and number of genes (4,131-11,243). The total GC content was also found to vary significantly across the dataset, ranging from 31.7% in a *Tetrapisispora blattae* strain to 52% in a representative of *Torulaspora globosa*. The core genome of forty Saccharomyces complex species was also identified in this study and it was found that 591 genes with $\geq$50% amino-acid sequence identity were present across all strains.

Phylogenetic trees were then built from the full 76 species dataset, comprising Maximum Likelihood approaches for a seven-region Multi-Locus Sequence Typing and 1,711 BUSCO gene datasets along with three variations of a recently developed NGS alignment-free approach - Feature Frequency Profiles (FFP). The resulting trees were then compared, with all trees found to be different, though with the BUSCO and FFP 20-letter amino acid trees highly superior to the other approaches. Despite the success of the FFP 20-letter amino acid approach for the Saccharomyces complex dataset, simulation studies confirmed a sequence length bias with the FFP two-letter RY alphabet and a GC bias with the FFP four-letter DNA alphabet approaches. In an effort to overcome the biases within the current FFP approach, a new software tool, *jellyphy*, was developed. Further development of tools such as this will undoubtedly lead to new methods capable of accurate phylogenetic estimation from yeast NGS datasets.

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.

First and foremost, I would like to thank my amazing primary supervisor Jo Dicks for all the help, advice and time given over the past four (and a bit) years. I felt very lucky to have such a nice supervisor, which made the experience much easier. I would also like to thank my supervisory team, Katharina Huber, Steve James and Ian Roberts (in my first year), for their invaluable advice and their insights into Phylogenetics and Yeast. I'm very thankful to all the NCYC lab members, in particular Steve James and Adrian Turner, for their help and advice in the lab particularly while my wet-lab skills were still a bit rusty. Thanks also to my lovely colleagues Chris Pyatt and Prithika Sritharan for all the coding help and the well-needed coffee breaks.

I'm very grateful to my family for all the encouragement they have given me throughout my life. Thanks for always making me believe in myself. A special thanks to Colleen for all the great adventures and for that bit of home that I sometimes really needed. Thanks to all the lovely friends I've made along the way, in particular Dan, Magda, Bartek, Edel, Dimitra and Erika. All the fun and laughs we had together definitely kept me sane throughout the four years. Finally, I'd like to thank Ruud for the huge amount of support he has given me over the last year and a half. It would have been much more difficult to complete this PhD without him.

Chapter 1

Introduction

Numerous scientific studies, from evolutionary analyses to industrially-focused trait improvement programs, are enabled by having an accurate tree of life that describes the relationships between organisms. The estimation of such a tree of life depends on the availability of rigorous analytical methods and good quality datasets. Ideally, such methods should be able to take advantage of new genomic datasets that are becoming ever more prevalent. This study is focused on the goal of achieving a tree of life for yeasts. It describes work to compare different phylogenetic tree-building approaches with use of a large new dataset of sequenced yeast genomes, as well as taking a closer look at the genomic differences between such datasets. Here, we begin by touching upon the diversity of yeast species and the current computational methods for showing their phylogenetic relationships.

1.1 Yeast

Yeasts are defined as predominantly unicellular organisms of the kingdom of Fungi, found within the phyla of Ascomycota and Basidiomycota (see Figure 1.1). At present there are around 1,500 recognised species distributed between the two phyla. Although yeasts are predominantly unicellular organisms, they evolved from multicellular ancestors, with some species able to develop multicellular characteristics by forming strings of connected bud-

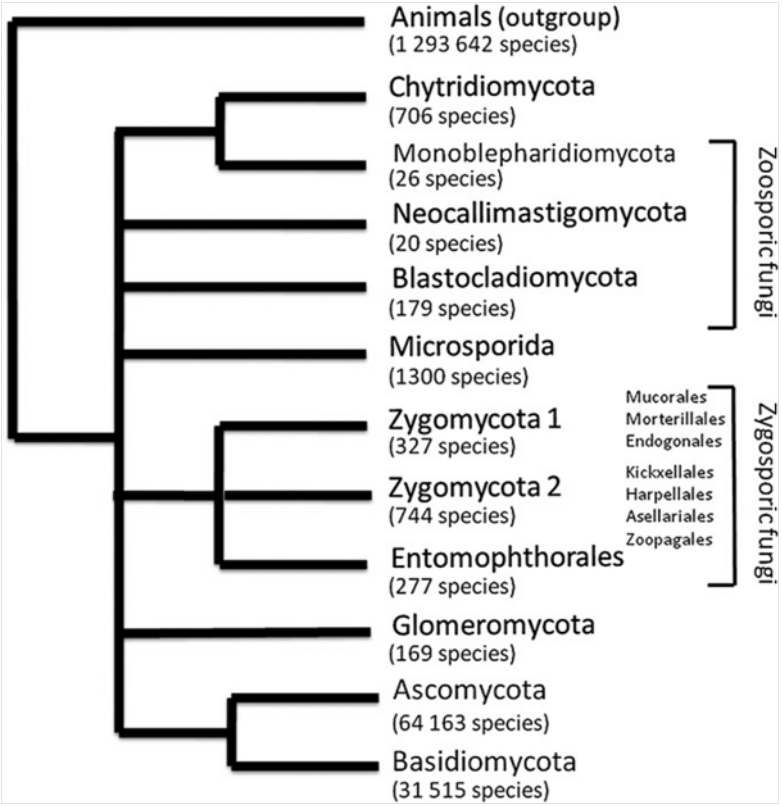


Figure 1.1: Fungal phyla and approximate number of species in each group, shown with the Animal kingdom outgroup. Yeasts are found in the Ascomycota and Basidiomycota at the bottom of the figure. Taken from Blackwell (2011).

ding cells known as pseudohyphae (Knop (2011)). Most yeasts reproduce asexually by mitosis and many do so by the asymmetric division process known as budding. Budding yeasts ('true yeasts') are classified in the order Saccharomycetales within the Ascomycota phylum.

Yeasts are an ancient group of microorganisms and are believed to have originated around 500 million years ago (Lücking et al. (2009)). Extant yeast species use a plethora of organic compounds as a source of energy. Yeast species can be obligate aerobes or facultative anaerobes and survive under a variety of temperatures and conditions. As a result, yeast can be found in a multitude of environments from the human gut (Huffnagle and Noverr (2013)) to deep-sea environments (Kutty and Philip (2008)).

The subphylum Saccharomycotina of the Ascomycota phylum comprises most of the ascomycete yeasts which includes the well-known baker's yeast *Saccharomyces cerevisiae* as well as opportunistic human pathogens such as *Candida albicans* (Thompson et al. (2011)) and *Eremothecium gossypii*, an agriculturally important plant pathogen (Wendland and Walther (2005)). *Saccharomyces cerevisiae* in particular has long been exploited for its capacity to convert sugars to ethanol and desirable flavour compounds (Michel and McGovern (1992)). Brewers, winemakers and bakers have been fermenting alcohol well before Louis Pasteur demonstrated that yeast were responsible for this process (Pasteur (1857)). Recent phylogenetic research into the domestication and divergence of *Saccharomyces cerevisiae* beer yeasts has shown the domestication of some yeast strains before 1857 (Gallone et al. (2016)). The oldest known vessel for alcohol storage was found in China in 2005 and has been dated at around 7,000 years old (McGovern et al. (2004)).

Several species of the Saccharomycotina subphylum are of economic importance and have been of great benefit to human society and quality of life for a long time. In industry, yeasts have been used to produce more than just beer and bread but also insulin, vaccines, food products, food supplements, ethanol for the biofuel industry and to generate electricity in microbial fuel cells. Yeasts have been acknowledged as amongst the most important organisms in biotechnology for some time. They have small compact genomes and thirty one percent of *Saccharomyces cerevisiae* genes have homologs in the human genome which has led to them being used in genetic studies. Academic yeast research is a very fast-moving field as, despite several Noble prizes involving yeast research in recent years, much is still to be learned from these extremely diverse organisms and their genomes.

Understanding the relationships between species of yeast can be highly important, for example in enabling the leveraging of information across closely related organisms and visualising evolutionary patterns of trait nov-

elty amongst species and strains. Evolutionary analyses and the generation of family trees of yeast species for academic and industrial reasons have been a focus of study for some time. Early methods of understanding relationships between yeast species were phenotype-based or single-gene based (usually a conserved ribosomal gene). In 2003, Kurtzman and Robnett improved upon this by conducting a multi-gene approach to building a phylogenetic tree of the 'Saccharomyces complex' species. It must be noted that this term, used for convenience in this thesis, is not in common use today as it infers a large degree of similarity between this set of species to *Saccharomyces cerevisiae* although this group of genera are quite divergent from the model yeast. The approach used in the study resolved 75 species into 14 clades (Figure 1.2, Kurtzman and Robnett (2003)). This same approach led to the circumscription of species and the proposal of new genera (Kurtzman (2003)). In 2013, Kurtzman and Robnett once again took a multi-gene approach to tree building in Ascomycota using Maximum Likelihood techniques (Kurtzman and Robnett (2013)). The results again showed the limited congruence between a system of classification based on phenotype and a system based on DNA sequence.

More recently, this multi-gene approach has been expanded to include 1,233 protein-coding genes from 86 yeast genomes to reconstruct the backbone of the Saccharomycotina yeast phylogeny (Shen et al. (2016)). In 2017, Choi and Kim (Choi and Kim (2017)) reconstructed the phylogeny of 244 fungal species (including a number of Saccharomycotina species) from whole proteome sequences by use of an alignment-free approach called Feature Frequency Profiles (FFP) (Sims et al. (2009a)). A phylogenetic tree of the largest set of budding yeast species to date was constructed by Shen *et al.*, in 2018 which took a concatenation-based Maximum Likelihood approach using 2,408 amino acid orthologous groups (Shen et al. (2018)). Three hundred and thirty two species from 12 major clades of the Saccharomycotina subphylum were included in this genus-level phylogeny (See Figure 1.3).

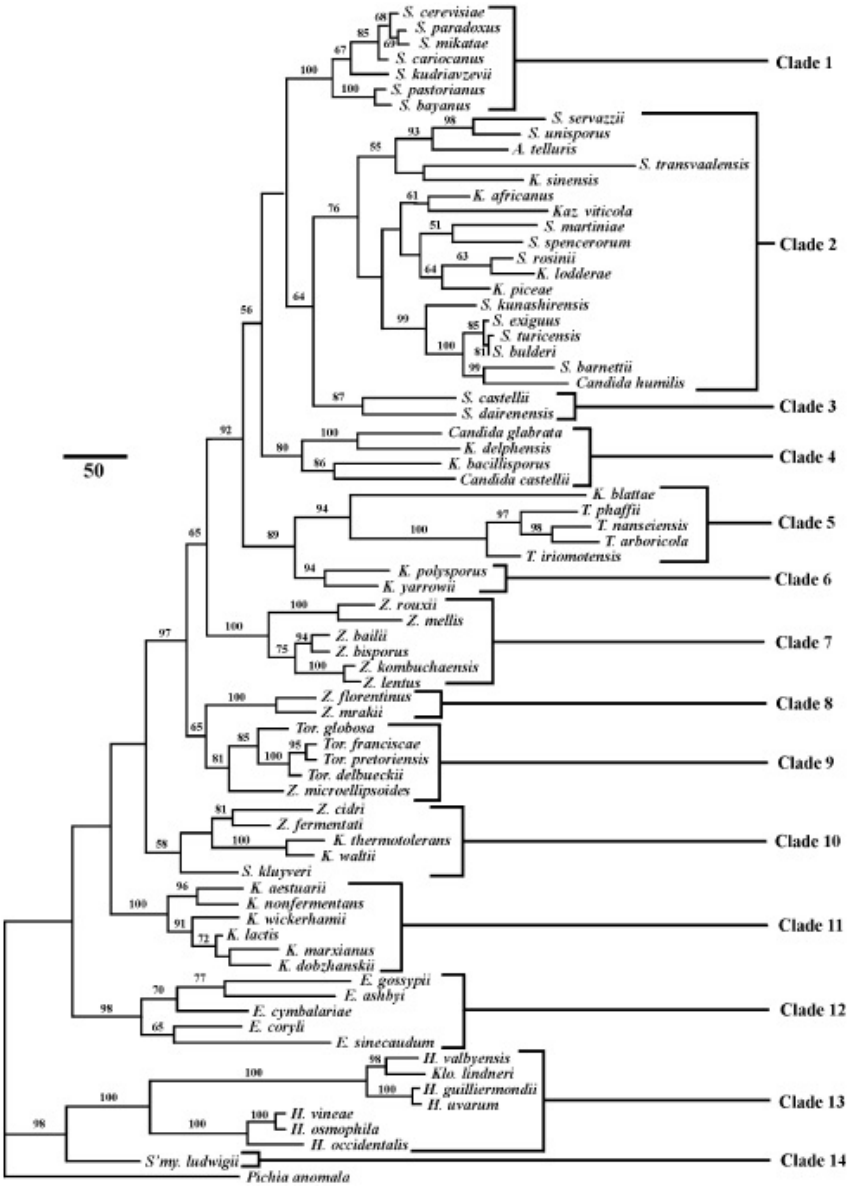


Figure 1.2: Kurtzman and Robnett Multi-locus Sequence Typing tree of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (formerly *Pichia anomala*; Kurtzman and Robnett (2003)).

There are number of yeast culture collections around the world, the National Collection of Yeast Cultures (NCYC) is one of the largest and contains some 4,000 strains from around 530 different species. The collection originally consisted of brewing strains but now also contains genetically-defined yeast (used in many applications including cancer research), yeast associated with food spoilage and yeast of medical and industrial importance. Phylogenetic analysis of the NCYC strains has been highly useful for species identification and historically, such analysis of selected species from the collection has been based upon the use of ribosomal RNA sequences (Stratford et al. (2002), West et al. (2014)). More detailed analyses involving whole genome datasets could help in learning more about the extensive biodiversity of the NCYC collection, alongside other yeast datasets from around the world.

1.2 Introduction to Phylogenetic Analysis

Phylogenetic analysis, or phylogenetics, is the study of evolutionary relationships among biological entities such as species, individuals or genes. The development of an evolutionary tree helps us to think more clearly about the differences between species and allows us to analyse them in a statistical sense. Phylogenetic inferences generally involve finding *homologous* characters - characters in different organisms that are similar because they were inherited from a common ancestor that also had that character - and comparing them using tree reconstruction methods (Delsuc and Brinkmann (2005)).

The first known evolutionary tree was drawn by Charles Darwin in a notebook in 1937 and later illustrated in his famous book On The Origin of Species in 1859 (Figure 1.4, Darwin (1859)). This simple tree was soon after elaborated on by the German scientist Ernst Haeckel, who coined the term *phylogeny* and developed trees that looked more like those we seek to esti-

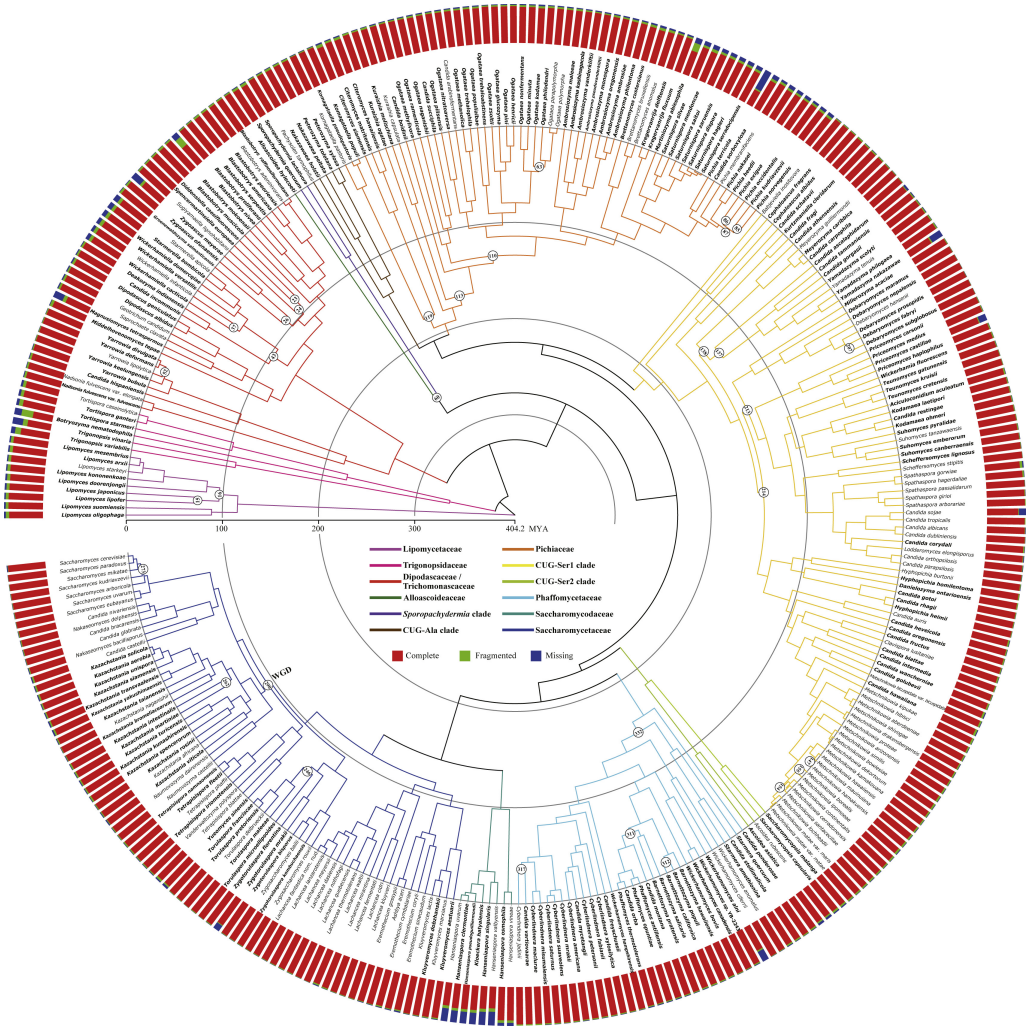


Figure 1.3: Time-calibrated phylogeny of the Budding Yeast subphylum, taken from Shen et al. (2018). Topology reconstructed from concatenation-based Maximum Likelihood analysis of 2,408 amino acid orthologous groups. Bar plots next to species indicate genomic quality assessed by a set of 1,759 genes selected through use of the BUSCO software (Red: Single copy; Orange: Duplicated; Green: Fragmented; Blue: Missing).

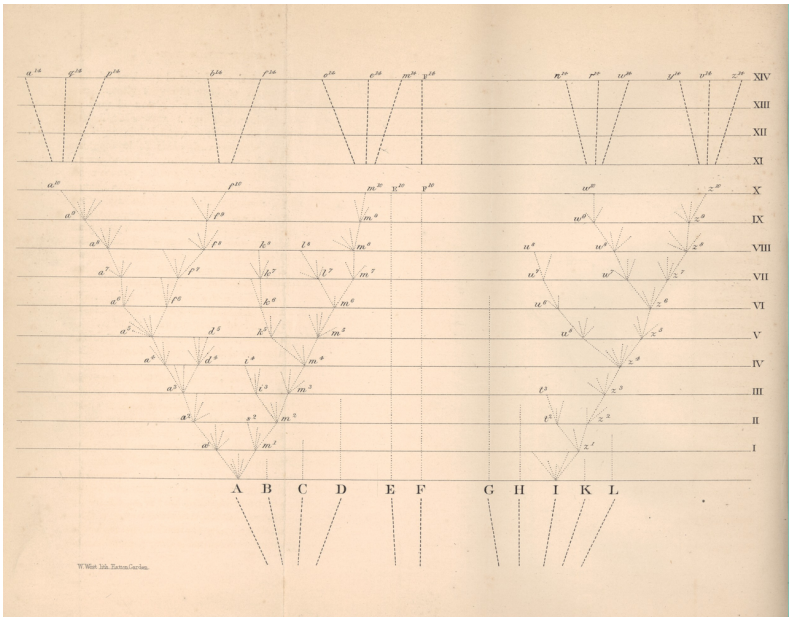


Figure 1.4: The first published illustration in 1859 of a phylogenetic tree, from Charles Darwin's famous book 'On The Origin of Species'.

mate today (see Figure 1.5). As the homologous characters used for phylogenetic inference moved from phenotypic to molecular, with the rise of DNA sequencing, our understanding of the relationships between species began to change. The work done by Carl Woese and collaborators in 1990, which used DNA sequences to construct a tree of life, enabled the definition of three domains of life: Bacteria, Archaea, and Eukaryotes (Woese et al. (1990)). More recently, a tree built from the genomic data of over 1,000 species has led to questions regarding the validity of the three-domain topology, with eukaryotes potentially relocated into the archael domain (Hug et al. (2016)). Whilst there is yet to be a definitive tree of life consisting of all lifeforms, great progress has been been made in quite a short period of time.

Phylogenetic trees can be used in different ways. They can help us to understand how genes, genomes and species have evolved. We not only learn about how sequences have evolved in the past but can predict how they may change in the future. As mentioned previously, phylogenetics based

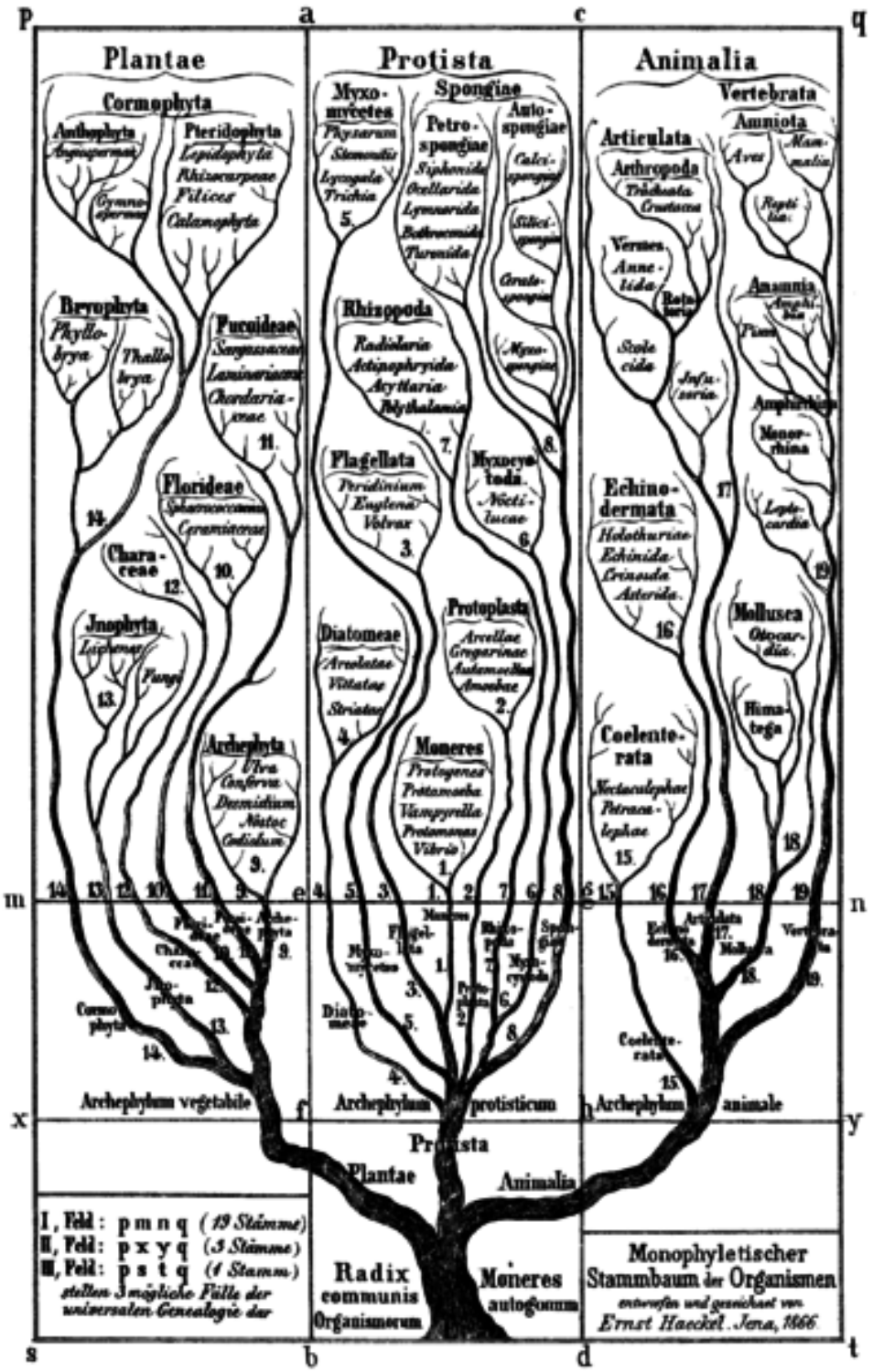


Figure 1.5: Ernst Haeckel's Tree of Life in *Generelle Morphologie der Organismen*, published in 1866. This is the first depiction of a comprehensive Tree of Life, showing the kingdoms of Plantae, Protista and Animalia.

on DNA sequence data can provide us with more accurate information regarding classification of species than traditional phenotype-based methods (Kurtzman and Robnett (2003)). They can also help to inform conservation policy and forensics as well as being an important area of research within the fields of bioinformatics and computing through the development of new algorithms.

In recent years we have also seen the application of phylogenetics in infectious disease surveillance and control. Molecular sequencing technology and phylogenetic approaches can be used to learn more about a new pathogen outbreak. This includes finding out about which species the pathogen is related to and subsequently the likely source of transmission. This can lead to a new recommendation for public health policy or new disease control measures. It can also be used in tracking of disease spread over time and space during an outbreak, such as was done during the 2015 Ebola outbreak in West Africa (Carroll et al. (2015), Holmes et al. (2016)) and more recently in 2020 for the Coronavirus outbreak (Shereen et al. (2020), Hamilton et al. (2020)).

A phylogenetic tree is comprised of nodes and branches (See Figure 1.6). A node represents a taxonomic unit which can be either an existing species or strain, or an ancestor. A branch defines the relationship between the taxa in terms of descent and ancestry. The topology of a tree refers to the branching patterns of the tree and branch length can be used to represent the number of changes that have occurred along the branch. Where a tree is rooted, the root represents the common ancestor of all taxa. A tree will also have a distance scale which represents the number of differences between organisms or sequences. The term clade refers to a group of two or more taxa or DNA sequences that includes both their common ancestor and all of their descendants.

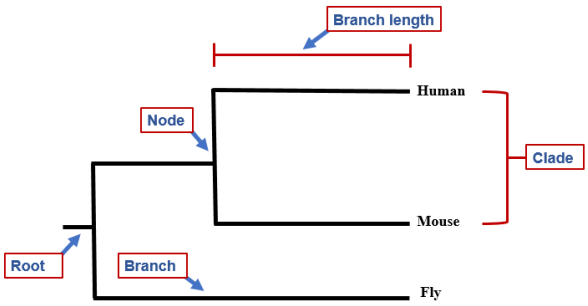


Figure 1.6: Basic structure of a phylogenetic tree, showing the key concepts of taxon-representing nodes connected by branches that can be scaled relative to the rate of evolutionary change. A tree may be unrooted or be rooted by various algorithmic means. Clades represent taxonomically-meaningful groups of taxa. In this example, human and mouse are more closely related to one another than either is to fly, with the depicted clade representing the common taxonomic class of Mammalia.

1.2.1 Rooted and unrooted trees

Trees constructed by computational means can be rooted or unrooted depending on the input data and the algorithm used. A tree can be rooted by using a species distantly related to the species of interest called an outgroup. It can also be rooted with the use of a molecular clock that defines that changes that happen on tree branches arise at a constant rate across all branches. This assumption can be unrealistic in cases where the species are more distantly related and the circumstances under which they evolved may have changed considerably throughout the tree (see below).

1.2.2 Characters and defining homology

Reconstruction of a phylogenetic tree involves the identification of homologous characters that are shared between different organisms, and the inference of phylogenetic trees from the comparison of these characters using reconstruction methods. Originally morphological or ultrastructural char-

acters were used to distinguish different groups but such features are limited in microorganisms. With access to DNA sequence since the 1980's, the number of homologous characters has greatly increased. RNA and amino acid sequences can also be used. Some genes are commonly used as reference markers, for example the ITS region of the ribosomal DNA gene in yeast is used as a type of DNA barcode (Schoch et al. (2012)). The reason for such choices is because of the considerable degree of conservation in these regions across organisms, although it must be noted that information from a single gene can often be insufficient to obtain firm statistical support for a node of a phylogeny. Using multiple genes or whole genomes can help alleviate this problem by expanding the number of characters and reducing biases caused by genes that evolve differently from the species as a whole.

In sequence-based phylogenetic reconstruction, we typically align DNA or amino acid sequences of homologous genes from different organisms, such that many of the characters match across most of the sequences (see Figure 1.7). For some datasets this can be both computationally difficult and time consuming as deletions and insertions in the sequence make it difficult to decide where to put gaps. Software is available to construct these *multiple sequence alignments* (e.g. MUSCLE (Edgar (2004))) but with large numbers of sequences with much variation the result may not seem optimal upon human inspection and thus additional time will be required to manually adjust the alignment.

1.2.3 Phylogenetic Methods

There are two main approaches for phylogenetic inference, distance-based approaches and character-based approaches. Distance-based approaches first convert a character matrix into a distance matrix that represents the evolutionary distances between all pairs of species. The phylogenetic tree is then inferred from this distance matrix using algorithms such as Neighbor-Joining (Saitou and Nei (1987)) and UPGMA (Unweighted Pair Group

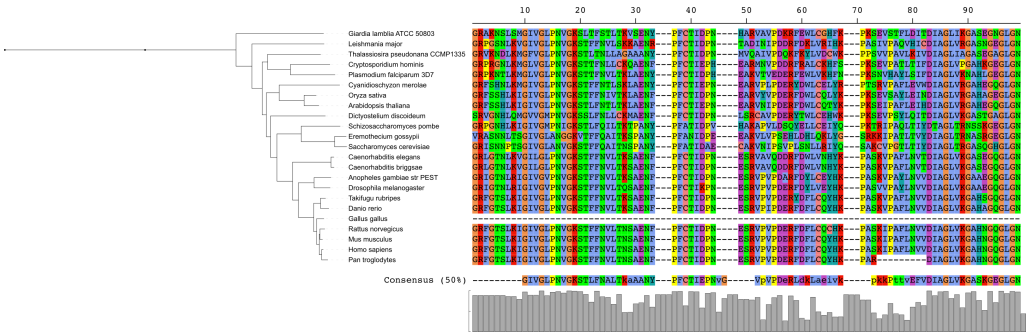


Figure 1.7: Example of a Multiple Sequence Alignment of homologous amino acid sequences from 23 species across the tree of life, and a phylogenetic tree derived from it, taken from the iTOL website (https:// itol.embl.de/ help.cgi).

Method with Arithmetic mean) (Sokal and Michener (1958)). Character-based approaches involve the simultaneous estimation of the tree topology and other evolutionary parameters. These methods include Maximum Parsimony (Sober (1983)), originally developed as a computationally tractable method at a time when computing resources were limited. This method selects the tree that requires the minimum number of character changes along all branches of the tree to explain the observed data. With today's better computational resources the character-based method of Maximum Likelihood (Felsenstein (1981)), rooted in statistical theory, has become more widely used. This method is based on a function that calculates the probability that the given tree could have been produced from the observed data. This function allows the incorporation of the processes of character evolution (e.g. nucleotide substitution) into the model. A more recent approach to phylogenetic reconstruction is Bayesian inference (Huelsenbeck et al. (2001)). Bayesian methods derive the distribution of trees according to their posterior probability, using Bayes' Theorem to combine the likelihood function with prior probabilities on trees and other tree parameters. Unlike the Maximum Likelihood approach, which optimizes model parameters by finding the highest peak in the parameter space, Bayesian approaches integrate model parameters across all possible values.

1.2.4 Models of evolution

Within a Maximum Likelihood or Bayesian analysis, a model of sequence evolution (DNA, RNA or amino acid) is required to generate a phylogenetic tree from a multiple sequence alignment, as the rate at which nucleotides are substituted in a given sequence must be taken into account to get the best possible tree estimate. The theory of a molecular clock was proposed in 1965 (Zuckerkandl and Pauling (1965)) and received backing by Motoo Kimura in 1987 (Kimura (1987)). This approach considered the rate of molecular evolution to be approximately constant over time in all lineages. This meant that times of divergence between genes, proteins or lineages could be dated by measuring the number of differences between sequences. A stochastic evolutionary model for DNA sequences was proposed by Jukes and Cantor in 1969 (Jukes and Cantor (1969)) followed by more complex models such as the HKY (Hasegawa Kishino Yano) model (Hasegawa et al. (1985)) and the GTR (General Time Reversible) model (Tavaré (1986)). These models take into account the unequal rates of nucleotide change in sequences leading, in many cases, to a better estimate of distances between two DNA sequences. A range of evolutionary models also exist for RNA and amino acid sequences.

1.2.5 Evaluating tree support

Bootstrapping and jackknifing are statistical techniques for assessing the accuracy of almost any statistical estimate. Bootstrap and jackknife tests on phylogenies started with the work of Mueller and Ayala (Mueller and Ayala (1982)), Felsenstein (Felsenstein (1985)) and Penny and Hendy (Penny and Hendy (1985), Penny and Hendy (1986)). The bootstrap involves inferring variability in an unknown distribution from which a dataset was drawn by resampling from the data with replacement and the creation of a new dataset of the same size (see Figure 1.8). This new dataset is termed a pseudoreplicate and statistical support involves evaluating whether the original dataset is similar in a defined way to a set of pseudoreplicates. In

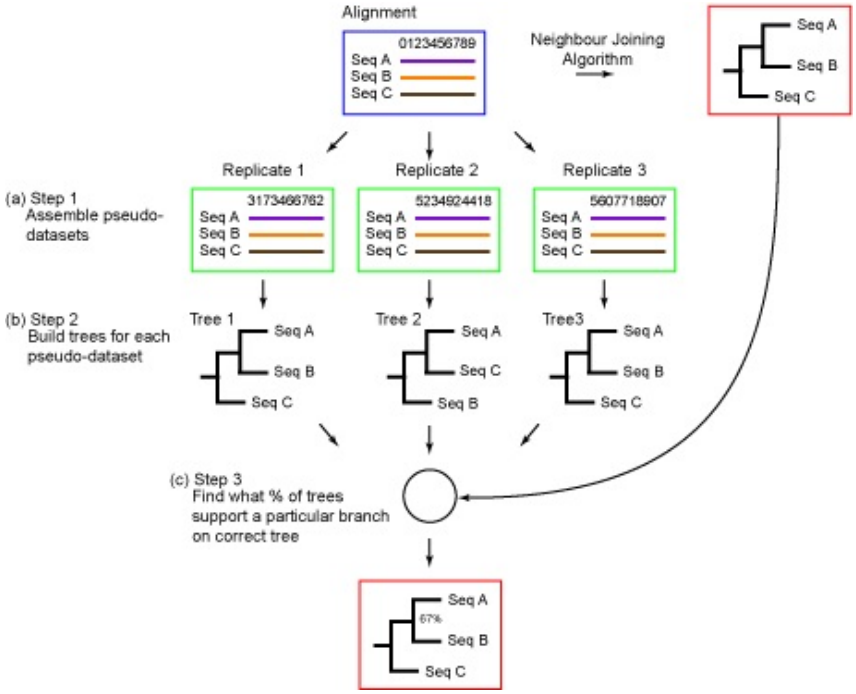


Figure 1.8: Bootstrap analysis. A Neighbor Joining tree (red box) is inferred from the input alignment (blue box). Columns within the input alignment are then randomly sampled (green boxes) and a tree inferred for each sample. This sampling-inference process is repeated – usually 1,000 times. Branching patterns within the original tree are compared to those within the trees derived from the random samples (circle) (Figure taken from Baldauf (2003)). Here, two of the three trees built from pseudoreplicate datasets match the original tree.

phylogenetic analysis, this similarity is measured in the similarity of the phylogenetic trees derived from the pseudoreplicate and original datasets. The bootstrapping process randomly samples and replaces columns from a species character matrix (species are the rows, characters are the columns, for example from a multiple sequence alignment), whilst jackknifing does so without replacement. A phylogeny is then constructed from each pseudoreplicate and a consensus tree can be built to summarise the number of nodes that are shared among the set of trees.

When taking the Bayesian inference approach to phylogeny reconstruction, a posterior distribution of highly probable trees are generated given the

data and evolutionary model. The statistical support at a node in this case reflects the probability that a clade exists given the data and evolutionary model. Whilst using Bayesian inference takes an evolutionary model into account and so may be more biologically realistic than bootstrapping, both measures have their weaknesses. Smaller and larger clades tend to attract larger support than a mid-sized clade as a result of the number of taxa they contain (Prevosti and Chemisquy (2010)). Also, bootstrap support can provide high estimates of node support as a result of noise in the data rather than the true existence of a clade (Phillips et al. (2004)).

1.2.6 Tree comparison metrics and dataset generation

There is quite a variety of approaches to phylogenetic tree building which can make choosing the most appropriate method for one's question and data challenging. There is also no way to measure whether a particular tree is accurate or not unless the true relationships among taxa are known from another source of evidence. One method of assessing the most appropriate tree building method to use would be the simulation of datasets which allows the researcher to know the true or expected topology a method should produce in a certain evolutionary scenario (Hall (2005)). There are also different tree comparison metrics which look at similarities in topology alone, branch lengths or both.

The Robinson-Foulds distance metric (Robinson and Foulds (1981)) was designed as a way to measure the distance between unrooted trees. This method counts the number of branch partitions (or *splits*) that occur in one tree but not in the other, scoring 1 for each non-matched partition (Figure 1.9). The metric can either be unweighted or weighted, where the weighted version takes branch lengths into account. The Mantel test (Mantel (1967)) tests the correlation between two distance matrices. It is non-parametric and computes the significance of the correlation through permutations of the rows and columns of one of the input distance matrices. The test statis-

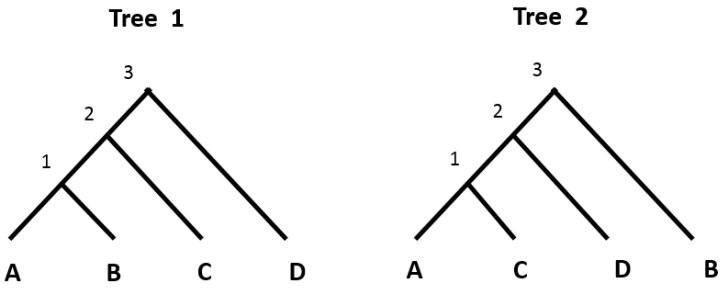


Figure 1.9: Robinson-Foulds distance (Figure taken from Mantel (1967)). This method counts the number of branch partitions that occur in one tree but not in the other, scoring 1 for each non-matched partition. Tree 1 contains the splits AB|CD (obtained by bisecting the branch between nodes 1 and 2) and ABC|D (between nodes 2 and 3) that are not seen in Tree 2. Conversely, Tree 2 contains the unique splits AC|BD and ACD|B.

tic is the Pearson product-moment correlation coefficient r. The coefficient r falls in the range of -1 to +1, where being close to -1 indicates a strong negative correlation and +1 indicates a strong positive correlation. A value of $r = 0$ indicates no correlation. Another more recently developed metric is the Kendall-Colijn metric which measures the distance between rooted trees (Jombart et al. (2015), Figure 1.10). The metric records the distance between the most recent common ancestor (MRCA) of a pair of tips and the root, in each tree. The metric can either be unweighted or weighted.

Tree comparison methods were evaluated by Kuhner and Yamato in their 2015 study (Kuhner and Yamato (2015)). Here they evaluated the performance of nine tree distance measures on two tasks; distinguishing trees separated by lesser versus greater numbers of recombination events and distinguishing trees inferred with lower versus higher quality data. When comparing trees of similar topology, measures that make use of branch lengths were found to be superior (the Robinson-Foulds metric performed best in

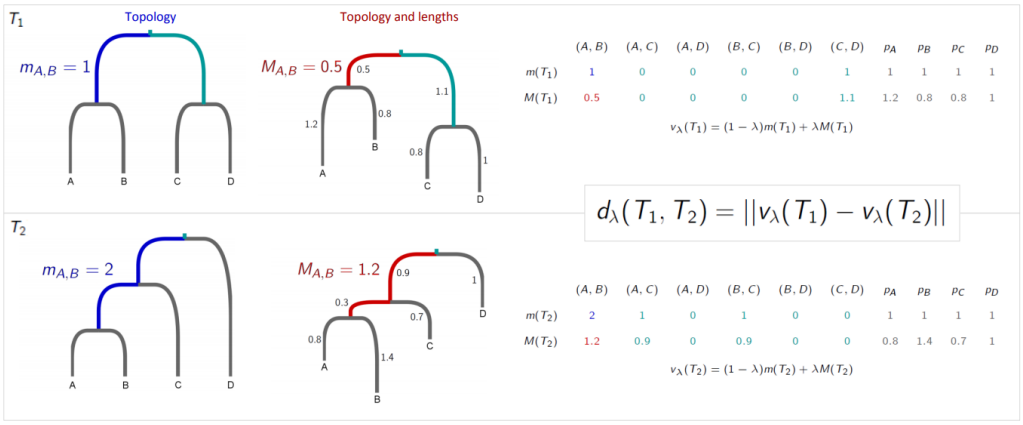


Figure 1.10: Kendall-Colijn metric (Figure taken from Jombart et al. (2015)). A tree is characterized by the vectors m and M, which are calculated as shown. These are used to calculate the distance between the trees for any $\lambda \in [0, 1]$. Here, d0(T1, T2) = 2 and d1(T1, T2) = 1.96.

this study). For dissimilar trees the topology-only measures are superior (the Align metric of Nye et al. (2006) proved optimal in this case). The authors concluded that the best metric depends on whether branch-length information is of interest.

1.3 Phylogenetics in the Next Generation Sequencing era

Past methods used for phylogenetic analysis have had their limitations when affected by confounding biological phenomena such as horizontal gene transfer, hybridisation and missing data although these can often be combated by increasing taxon sampling, the number of genes analysed or whole genome analysis. Many phylogenetic tree building approaches were developed in a time where we only had access to sequences of relatively short length, like a single gene. In the space of a few years there has been a drastic reduction in sequencing prices. This has allowed us to sequence whole genomes at scale. As a result large whole genome datasets have become available for analysis. With this though comes challenges in tree building and new approaches are

being explored.

Next-generation sequencing (NGS) has resulted in much larger datasets which can make the alignment step in tree building extremely difficult. As a result, new alignment-free approaches have started to be developed which essentially compare 'word' frequencies in sequences. These methods are much more efficient as aligning two sequences takes time proportional to their total sequence length whilst word frequencies, which most alignment-free methods use, can be calculated in linear time (Vinga and Almeida (2003)). Comparisons of the different approaches were made by Vinga and Almeida in 2003 and since then many more improved approaches have been developed which use word (or k-mer) frequency in alignment and assembly (Flicek and Birney (2009)). In 2013, Roychowdury *et al.*, generated accurate trees with an assembly-free method using short sequence reads (Roychowdhury et al. (2013)). In the same year Yin and Jin developed an assembly-free method that created micro-alignments (Yi and Jin (2013)). More recently, the use of spaced-words instead of contiguous words to estimate distances for trees has been proposed (Leimeister et al. (2014)).

These alignment-free methods of sequence comparison are becoming increasingly popular for genome analysis and phylogeny reconstruction as they avoid the various difficulties of alignment-based approaches. One particular method, FFP (Feature Frequency Profiles), has become widely used in the last decade and will be discussed next. (Sims et al. (2009a)).

1.4 Feature Frequency Profiles

The Feature Frequency Profiles (FFP) approach compares k-mer frequencies between whole genomes and is particularly well suited to analysing large whole-genome datasets as a result of its efficiency and the ability of this method to capture phylogenetic signal(s) across entire genomes (Sims et al. (2009a)). This alignment-free method can and has been used with

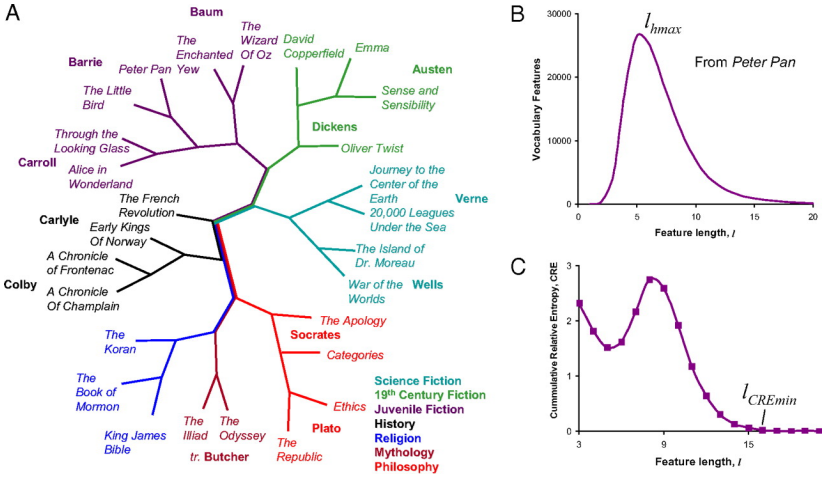


Figure 1.11: The Feature Frequency Profile (FFP) approach is used to A) create a tree of books of literature based on word frequencies, showing that books of a common genre tend to group together. Graphs in B) and C) indicate that the optimal word length for discrimination of these books is $l = 9$. Figure taken from Sims et al. (2009a).

viral, bacterial, fungal and mammalian sequences in the past (Wu et al. (2009), Sims and Kim (2011), Choi and Kim (2017),Choi and Kim (2020)).

FFP calculates the frequencies of features (e.g. DNA strings such as 'AATT') of a suitable length in one genome and compares them to those frequencies observed in other genomes, using either a 4-letter DNA alphabet (ACGT), a 2-letter RY alphabet (where R is Purines - A or G - and Y is Pyrimidines - C or T) or a 20-letter amino acid alphabet. Then the relative frequency of each feature is calculated and a distance matrix created using the Jensen-Shannon distance (Lin (1991)). The distance matrix can then be used to create a phylogenetic tree, either using early versions of the FFP program (versions up to 3.19) or using third-party software such as PHYLIP's neighbor program (Felsenstein (1989)) or BIONJ (Gascuel (1997)). Figure 1.11 illustrates the use of the FFP software to show relationships - in terms of common words - between books of literature.

1.5 Sequencing Technology

In 2001 the full sequence of the human genome was released to the public. This was the result of 15 years of work by scientists across the world and 3 billion dollars in funding. Since then the price and time required for genome sequencing has been reduced significantly (Wetterstrand (2016)). Today a high quality draft of the human genome costs around $1,500 and the whole exome can be sequenced for less than $1,000 (Wetterstrand (2016)). There has been a sequencing technology explosion since the Human Genome Project ended that has allowed a multitude of questions about the genome to be asked and answered at unprecedented speed and resolution.

Sequencing of DNA has evolved from 2D chromatography in the 1970's to the Sanger chain termination method which drove the Human Genome Project and most recently the advances in next generation sequencing technologies (Metzker (2010), Glenn (2011)). The "first generation" instruments which employed the automated capillary electrophoresis based sequencing were considered high throughput at the time but by 2005 the Genome Analyzer by Illumina took sequencing runs from 84Kb/run to 1GB/run (Illumina (2016)). This was a fundamentally different approach which used short reads and was done in a massively parallel fashion. This technology revolutionised sequencing capabilities and launched the "next-generation" in genomic science. Other technologies include SMRT sequencing (single molecule real-time) which produces longer reads with an average read length of around 10,000bp, modified base detection, high accuracy from unamplified molecules and can even detect methylation (Roberts et al. (2013)). Most recently, the extremely long lengths of Nanopore sequencing (Branton et al. (2010)) are revolutionising the field. Once the cost and accuracy of these long reads achieve levels similar to those of short-read sequencing, genome sequencing and assembly will undoubtedly gain a further foothold in modern biological research.

The data output from NGS has more than doubled every year since its inception (Wetterstrand (2016)). By 2014, the rate of data production had climbed to 1.8 terabases of data in a single sequencing run. With this massive increase in data output at reduced time and cost, the ability to sequence the whole genomes of many related organisms allows large-scale comparative and evolutionary studies to be performed that were unimaginable a few years ago. Within the yeast community, the genomes of thousands of strains have now been sequenced, though as many of these were in population genomics studies, the number of sequenced species is still relatively low. Recently, the NCYC sequenced ∼1,000 strains, approximately a quarter of the collection, from ∼200 of its species. This project aims to take full advantage of these revolutionary technologies and massive datasets to reveal the genetic interrelationships of extant yeast species in the most precise detail possible.

1.6 Aims and Objectives

The ultimate goal of the project is to contribute to the computation of a yeast tree of life from whole genome sequences. However, several data and technology gaps stand in the way of achieving this goal. It was noticed that the NCYC sequenced genomes encompass almost all of the 75 species (and outgroup) included in the seminal study of Saccharomyces complex species shown in Figure 5.3. Consequently, an achievable aim of this project is to use this new dataset as a basis to compare state-of-the-art phylogenetic methodologies. No comparisons of phylogenetic methods have yet encapsulated the broad range of techniques on offer to resources such as the NCYC, for which no whole collection tree yet exists. Also, current choices may not be optimal for large NGS datasets and potentially result in errors in downstream analyses that rely on a high quality tree (e.g. convergent trait evolution). Two additional aims of the project are to explore the genomic diversity within the dataset and to assess the composition of and phylogenetic signal within the core genome of a subset of the dataset.

The objectives of the project are to:

- undertake stringent quality control of the 76 species draft genome assemblies for use in the project.
- compare the results of five different phylogenetic tree-building approaches on the full dataset.
- investigate a hypothesised GC bias in the FFP four-letter DNA alphabet approach through a simulation study.
- begin an effort to improve the current FFP approach by writing a new piece of alignment-free software.
- assess the genomic similarities and differences within the dataset.
- test different approaches of core genome identification.
- compare phylogenetic trees built from yeast core genomes of varying levels of sequence identity.

1.7 Summary of Thesis

In Chapter 2, generation of this new NGS dataset for the Saccharomyces complex yeasts is discussed and measures to ensure that its quality is sufficient for further analysis are examined. Chapter 3 discusses concepts such as yeast core, pan and accessory genomes and looks at the composition of, and phylogenetic signal within, a range of core genome estimated from a subset of the Saccharomyces complex dataset. Chapter 4 takes a first look at some global differences between the genome assemblies, examining statistics such as gene count and GC content. In Chapter 5, phylogenetic trees are generated from the new NGS datasets using five different methodologies, comparing them to one another and to the tree in Figure 5.3 using two computational measures (the Robinson-Foulds and Kendall-Colijn metrics). Chapter 6 examines the GC content within the Saccharomyces

complex dataset and investigates a hypothesised GC bias with a simulation study. Finally, Chapter 7 discusses the project as a whole and looks at the next steps that could be taken in pursuit of a yeast tree of life.

Chapter 2

Quality Control for a classic Saccharomyces complex dataset

2.1 Introduction

Laboratory quality control or analytical quality control, refers to the processes and procedures designed to ensure that the results of laboratory analysis are consistent, comparable, accurate and within specified limits of precision. The high quality of biological data is crucial for achieving an accurate outcome to a research question. With whole genome sequencing (WGS) data, it is imperative to confirm the identity of a given genome sequence when making biological inferences about a specific strain. In phylogenetic studies, an incorrectly identified strain or a contaminated sequence can lead to an inaccurate conclusion being drawn regarding evolutionary relationships. A whole genomic sequence should also be of good quality so as to cover the full genome and thus give sufficient phylogenetic signal. Data quality becomes even more important when making inferences regarding gene counts and annotating a novel genome. Errors in an early publication are then more likely to snowball through the scientific literature.

There are a number of steps prior to obtaining a whole genome sequence where issues may arise that could lead to a poor quality, misclassified or contaminated sequence. Within the laboratory, there can be issues with the original sample which is grown up prior to DNA extraction. In some cases, particularly if working with old collections, like the NCYC, strains could have been historically misclassified as a result of outdated classification approaches or human error. A quality control step of sequencing a DNA barcode such as the ITS, 26S or 18S rDNA regions could be carried out here prior to sending the whole genome for sequencing. Although, as DNA sequencing has become more affordable this step may not be done. During the strain growing process and DNA extraction, although done aseptically, there is potential for human contamination which includes the microbes that live on us. The subsequent library preparation and DNA sequencing steps introduce further risk. These steps are usually done with more than one strain on a plate which could result in sample mix up or contamination from other strains.

If indeed all of these steps go according to plan and the strain is as expected, quality control of the sequence must still be undertaken to assess if there is sufficient genome coverage to make a reasonably accurate contig or scaffold assembly, should these be required. A wide range of software tools are now available for pre-processing of sequence read datasets and their subsequent analysis. Here, outlines of the main third-party software tools used in this project are given.

2.1.1 Sequence read pre-processing

Different software tools are available for the quality control of raw sequencing data. Programs such as FastQC (Andrews et al. (2010)) are commonly used to assess a range of sequence quality statistics such as read depth and base quality measures. Sequencing reads can also be pre-processed prior to genome assembly with tools such as Trimmomatic.

Trimmomatic Bolger et al. (2014). This software includes a variety of processing steps for read trimming and quality filtering, including the removal of read substrings where base quality is low. Trimmomatic also uses two approaches to detect technical sequences within the reads. The first mode searches for an approximate match between each sequence read and user-supplied technical sequences such as sequence adapters and polymerase chain reaction (PCR) primers, or fragments thereof. The second mode is specifically aimed at detecting the common 'adapter read-through' scenario, whereby the sequenced DNA fragment is shorter than the read length, which results in adapter contamination at the ends of reads.

2.1.2 Read mapping

BWA Li and Durbin (2009), Li and Durbin (2010) and Li (2013). The Burrows-Wheeler Alignment tool (BWA) is a read alignment package that is based on backward search with Burrows-Wheeler Transform (BWT), to efficiently align short or long sequencing reads against a large reference sequence, allowing mismatches and gaps. Read mappers that use the Burrows-Wheeler transform, such as BWA, are very fast but tend to be less sensitive than the best hash-based mappers.

Stampy Lunter and Goodson (2011). This is a package for the mapping of short reads onto a reference genome and which uses a hybrid mapping algorithm and a detailed statistical model to achieve both speed and sensitivity, particularly when reads include sequence variation. To achieve good sensitivity Stampy uses a hash table, representing the location of selected k-mers in the reference genome. To increase speed, the hybrid mode can be selected which uses BWA to map the majority of reads that have a close representative in the reference.

2.1.3 Genome assembly

ABySS Simpson et al. (2009). ABySS (Assembly By Short Sequencing) is a parallelized sequence assembler for reference-based and *de novo* genome assembly. In the first stage of this approach, all possible substrings of length k (k-mers) are generated from the sequence reads. The k-mer data set is then processed to remove read errors and initial contigs are built. In the second stage, mate-pair information - should this be available - is used to extend contigs by resolving ambiguities in contig overlaps. The primary innovation in ABySS is a distributed representation of a de Bruijn graph, which allows parallel computation of the assembly algorithm across a network of commodity computers.

SPAdes Bankevich et al. (2012). This approach was originally designed for single- and multi-cell bacterial data sets, although it can also be used for the assembly of fungal and other small genomes. SPAdes uses k-mers for building an initial de Bruijn graph and performs graph-theoretical operations which are based on graph structure, coverage and sequence lengths. This approach can be used for short, long and hybrid assemblies.

2.1.4 Genome assembly quality

There are a number of ways to assess genome assembly quality, contamination or completeness, three of which are described below.

N50. The number of contigs and a statistic known as the N50 can indicate the quality of a genome assembly. The N50 score is the sequence length of the shortest contig at 50% of the total assembly length. While a widely-used statistic, if a set of N50 values are derived from assemblies of significantly different lengths they are usually not informative, even if for the same genome. An alternative statistic is the NG50 statistic, which is calculated in a similar way to N50, except that here 50% of the known or estimated genome size must be of the NG50 length or longer. This would

depend on having prior knowledge, or a good estimate, of the genome size.

Jellyfish Marçais and Kingsford (2011). This software looks at various k-mer counts in sequence data such as draft genome assemblies and has the potential to identify contamination in a sequencing dataset. Jellyfish is an efficient k-mer counting tool which can count the number of features which occur within a genome only once ('Unique'), the number of features not counting multiplicity ('Distinct'), the total number of features including multiplicity ('Total'), the maximum number of occurrences of a feature ('Max count') and the frequency of features of length k (e.g. f100 = features of length k which appear at a frequency of 100 or more).

BUSCO Simão et al. (2015). One quality control approach which measures both genome assembly quality and annotation completeness is the identification of Benchmark Universal Single Copy Orthologous genes (BUSCO genes), based on expectations of gene content within the assembly informed by its implied evolutionary history. Protein coding genes that make up the BUSCO datasets are defined as evolving under "single-copy control" (Waterhouse et al. (2011)), and are selected from OrthoDB (Kriventseva et al. (2019)) orthologous groups that contain genes present as single-copy orthologs in at least 90% of the species included in the group. This establishes an evolutionarily informed expectation that these genes should be found as single-copy orthologs in any newly sequenced genome or gene set from that group. Therefore, if there are many BUSCO genes from the appropriate clade that cannot be identified in a genome assembly or annotated gene set, it is possible that the sequencing and/or assembly and/or annotation approaches have failed to fully capture the complete expected gene content (Seppey et al. (2019)).

2.1.5 Species identification

Whilst factors such as high read coverage and a good genome assembly are optimal for many downstream analyses, they may not be necessary for a phylogenetic study. A mediocre assembly may still be sufficient to give the phylogenetic signal needed to infer the correct evolutionary relationships during tree-building. More important in this case is that the correct species has been confirmed for a given sequencing dataset. Inclusion of a misidentified strain is a major problem in phylogenetic tree estimation. Species identification can be confirmed through the genome assembly step, if the sequence reads map very closely to a selected reference genome. Furthermore, the two approaches described below can be used where the species' identity is unknown.

Kraken Wood and Salzberg (2014). Kraken is a system for assigning taxonomic labels to short DNA sequences, usually obtained through metagenomic studies. This approach utilizes exact alignments of k-mers and a novel classification algorithm. If the expected species' genome is in the database then the query genome's sequence reads will map to it. It also has the potential to show contamination and even hybrid samples as the proportion of reads mapping to different species or genera in a database can be shown. One highly illustrative example of contamination detection with Kraken, as well as a cautionary tale with regards to using publicly available genomes, was shown in a 2014 study (Merchant et al. (2014)). Here, cow and sheep DNA was found in the *Neisseria gonorrhoeae* TCDC-NG08107 genome that had been submitted to GenBank as complete. Kraken comes with a standard database which consists of bacterial, archaeal, and viral RefSeq genomes, along with the human genome and a collection of known vectors. Additional databases including all RefSeq fungal genomes can also be downloaded as of 2020. There is also an option to build a custom database if required.

BLAST Altschul et al. (1990). Although thousands of genomes are now publicly available, there are a number which are misclassified or contaminated and even more which have yet to be sequenced. Where this is the case a BLAST approach to species identification can be used. This is a relatively simple and efficient test of species identification which finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of each match. When taking this approach to species identification, one must decide which or how many conserved genes to test and the identity threshold or e-value at which to accept or reject identity. Different conserved genes, usually rDNA genes are assessed. The 26S, 18S and ITS regions of the rDNA are commonly tested for yeast species, of which a large number have been made publicly available within the GenBank database. In 2016, yeast ITS and LSU (large subunit or D1/D2 domain of the 26S gene) sequences from 4,730 yeast strains (of 1,351 yeast species) from the CBS collection were sequenced and submitted to GenBank (Vu et al. (2016)). The taxonomic threshold to discriminate the Ascomycete CBS yeast species was found to be 98.31% using ITS barcodes and 99.41% when using LSU barcodes. The authors found that only 6 and 9.5% of CBS yeast species could not be distinguished by ITS and LSU, respectively. Among them, 3% were indistinguishable by both loci.

To conclude, quality control of sequence data is essential. There are many ways in which one can receive sequencing data which is misclassified, contaminated or of too poor a quality to give an accurate representation of a genome. As a result, there are a range of different approaches and software available for just this issue. Setting high thresholds for sequence quality and species identification is very important. Unfortunately, even with a stringent quality control step minor contamination can still occur. Depending on how the sequence is to be used (e.g. for a phylogenetic tree comparison project) this may be acceptable, whilst genome annotation would not.

This project aimed to compare the accuracy of different phylogenetic methods for a common dataset. To ensure fairness of findings and inferences made, a high quality dataset was required. The aim of the quality control step was therefore to prevent the construction of an incorrect phylogenetic tree and thus a misinterpretation of the relationships between species. The identity and sequence quality of a dataset consisting of 75 Saccharomyces complex strains and an outgroup were assessed by different methods. Assembly quality was assessed using metrics such as the N50 score, number of contigs and k-mer distribution (using Jellyfish, Marçais and Kingsford (2011)). Species identity was confirmed with BLAST (blast+ 2.2.30, Altschul et al. (1990)) and in some cases a custom Kraken database. Genome sizes, gene counts and BUSCO gene counts were also assessed to aid in contamination detection or poor genome assembly. These steps resulted in the removal of a number of misclassified or likely contaminated genome sequences.

2.2 Data and methods

2.2.1 Data selection

Whole genome sequence reads from one strain of each of the 75 Saccharomyces complex species seen in (Kurtzman and Robnett (2003)) plus those from the outgroup species, were obtained. Of these 76 species representatives (See Table 2.1), 63 datasets were within the NCYC laboratory (See Section 2.2.2 for more details). Fifty datasets were derived from NCYC strains, 12 were from CBS strains and 1 was from a DBVPG strain. The remaining fourteen datasets were obtained from either the NCBI GenBank (9 strains) or RefSeq (5 strains) databases. All strains selected were type strains where possible and all underwent stringent quality control, including the publicly available genomes.

The species identity of twenty-five of these datasets were confirmed using both BLAST and the Kraken QC approach and the remaining fifty-one were confirmed using the BLAST QC approach alone (See Section 2.2.5). The N50 score, genome size and BUSCO gene counts were also taken into consideration when confirming species identity. These combined approaches aided in determining whether contamination or a sequencing or assembling issue may have occurred during dataset generation. Several strain datasets were rejected during this iterative process, resulting in a final 76 strain set.

2.2.2 DNA extraction

Of the 63 NCYC-derived datasets, 48 were achieved within a BBSRC-funded whole genome sequencing project. The remainder were achieved within this project, following the same DNA extraction protocol used in the earlier project. All strains were grown in YM media at 25°C for approximately three days. DNA was extracted with the Epicentre MasterPure Yeast DNA purification kit. 1mL was taken from each of the 15 samples and placed in labelled 1.5mL Eppendorf tubes. The cells were pelleted in 14K rpm for 5 minutes. The supernatant was discarded and pellets were frozen at -20°C overnight. 100μL of Zymolyase (10mg/mL) was added to defrosted samples. The samples were incubated for 30 minutes at 37°C then centrifuged for 5 minutes at 14K rpm. The supernatant was discarded and 300μL of Cell Lysis Solution and 5μL of RNAse A were added. Cells were resuspended by gentle vortexing and incubated at 65°C for 15 minutes. Samples were then cooled on ice for 5 minutes. 150μL of MPC Protein Precipitation Reagent was added followed by vortexing. Cell debris was pelleted by centrifugation at 14K rpm for 5 minutes. Supernatant was transfered to a clean Eppendorf tube and 500μL of ice cold isopropanol was added. Samples were centifuged at 14K rpm for 10 minutes to get DNA pellets. Supernatant was discarded and 500μL of ice cold ethanol was added. Samples were centrifuged at 14K

for 5 minutes. Supernatant was removed and tubes were left open for 5 minutes to allow any residual ethanol to evaporate. 35μL of TE buffer was added and the samples were left in the fridge overnight for the DNA to dilute. A Qubit fluorometer was used for DNA concentration estimation, to ensure a sufficient quantity was present for whole genome sequencing.

2.2.3 Sequence read generation and pre-processing

All strains were sequenced using Illumina HiSeq or NextSeq short-read sequencers. The raw sequencing reads of the 62 datasets produced within QIB (the 48 NCYC sequencing project datasets plus the 14 datasets generated here) were pre-processed by Dr Jo Dicks. For the 48 NCYC sequencing project datasets, 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 strain). Draft genome assemblies were subsequently estimated from the paired trimmed reads using ABySS v1.9.0 (Simpson et al. (2009)), with the -k = 80 option. For a small number of datasets with poor assembly statistics, genome assemblies were also generated using the SPAdes software (Bankevich et al. (2012)) v3.13.1, with k = 30,40,50,60,70 and 80 to see if improvements could be achieved. For the 15 datasets generated here, a similar data processing regime was used, differing only in the use of BBTools v38.47 (Bushnell B. – sourceforge.net/projects/bbmap/) with the clumpify option prior to Trimmomatic in order to remove PCR duplicates.

Datasets for each of these strains were made available to this project in three formats: a) raw sequence paired-end reads (FASTQ), b) trimmed paired-end reads (FASTQ) following quality and adapter trimming, and c) sequence contigs (FASTA) following assembly of trimmed reads.

2.2.4 Sequence quality

The N50 scores and the number of contigs for each assembled genome were assessed (See Table A.1 of the Appendix for all assembly results). The average N50 of the final dataset was 217,552 with the highest being 1,398,029 (ATCC58844, publicly available genome) and the lowest being 2,328 (CBS106.43). The average number of contigs/scaffolds/chromosomes was 3,075, with the lowest being 7 chromosomes (again ATCC58844) and the highest count being 28,870 for NCYC2450. Although some strains had lower than expected scores, this was taken into account along with other quality control measures such as k-mer distribution, BUSCO gene count and BLAST species identity score, for the final decision on inclusion in this dataset.

Jellyfish k-mer counting was carried out on the contigs, looking for the number of k-mers of length 14. The k-mer length of 14 was determined through the use of the FFP phylogenetic tree-building method (Sims et al. (2009a)) and the Robinson-Foulds metric (Robinson and Foulds (1981)), explained in full detail in Chapter 5. For all analyses 10 threads, a hash size of 100M and the counting on both strands option was selected. The average counts for this dataset were: Unique k-mers: 8,343,675; Distinct k-mers: 10,133,967; Total k-mers: 13,440,181; Max count for a k-mer: 3,868. Strains particularly far above or below these counts were noted (See Table A.1 of the Appendix for all assembly results).

2.2.5 Species identification

The BLAST quality control step was undertaken with BLAST version 2.2.30 (Altschul et al. (1990)) (blastn). This involved downloading the gene sequence of the relevant 26S (or 28S where this was available) and 18S rDNA genes from GenBank (see Accession numbers in Table A.3 of the Appendix) to use as query sequences. BLAST databases were created from the contig

files of all strains, which were then queried. Strains were kept if the identity score was above 95% for both regions. All BLAST results can be found in Table 2.1.

Kraken was used to further confirm the identity of 25 datasets. A custom database was built with the publicly available whole genomes of 30 Saccharomyces complex species and two outgroup species (as of early 2018) with the Kraken v1.0 software (Wood and Salzberg (2014)). Eight other publicly available genomes were originally selected for the database but failed to pass the BLAST quality control step (i.e. failed to match their expected species rDNA regions). Each strains' trimmed paired read dataset was used to test identity with the appropriate reference sequence. Strains with identity above 70% were kept, with the exception of hybrids (See Table A.2 of the Appendix for Accession numbers).

2.2.6 BUSCO assessment

BUSCO v.3 (Simão et al. (2015)) was used as both a QC tool and a means of assessing differences between the genomes, including gene duplications. The Saccharomycetales (odb9) lineage was used to generate the gene set from the 76 species datasets, with contigs used as input. This resulted in a total of 1,711 core genes across these species. Results were recorded as 'Complete', 'single-copy complete', 'duplicated complete', 'fragmented' or 'missing'.

2.3 Results

The identities of 75 of the 76 species dataset were confirmed with the BLAST step. The results of this process are shown in Table 2.1. For the 26S or 28S region, 71 of the 76 species had scores of 99% or 100% sequence identity. For the 18S region, 70 of the 76 had 99% or 100% sequence identity. All strains had above 95% sequence identity to their respective rDNA re-

gions with the exception of CBS8763. This strain, which was expected to be *Tetrapisipora nansiensis*, had 91% sequence identity to the relevant 26S query sequence and 98% to the 18S gene. This strain was re-sequenced but the newer dataset appeared to be contaminated. The first assembly was maintained in the dataset although it seems possible that this dataset does not in fact derive from a *Tetrapisispora nansiensis* strain. The strain would have been re-sequenced again if time had permitted.

The BUSCO analysis (shown in Table 2.2) found 64 out of 76 strains had more than 90% complete BUSCO genes (n = 1,711) present in their genome assemblies. Irregardless of their low BUSCO gene counts, the remaining 12 datasets were kept for the phylogenetic comparison study as they passed the BLAST QC and had relatively average k-mer distributions and genome sizes. The BUSCO results of a subset of these genomes were investigated further in Chapter 4.

Clade	**Strain ID**	**Species name**	**26S**	**18S**
1	CR85	*Saccharomyces kudriavzevii*^{N}	100	100
1	NCYC2578	*Saccharomyces bayanus*^{B,T}	100	100
1	NCYC2888	*Saccharomyces mikatae*^{B,T}	100	100
1	NCYC2890	*Saccharomyces cariocanus*^{B,T}	100	100
1	NCYC3662	*Saccharomyces paradoxus*^{B}	99	100
1	NCYC392	*Saccharomyces pastorianus*^{B,T}	100	100
1	NCYC78	*Saccharomyces cerevisiae*^{B}	100	100
2	DBVPG7206	*Kazachstania turicensis*^{P}	99*	99*
2	NCYC1417	*Kazachstania lodderae*^{B,T}	99	99
2	NCYC2449	*Kazachstania telluris*^{P,T}	96	95
2	NCYC2450	*Candida humilis*^{P}	100	99
2	NCYC2483	*Kazachstania piceae*^{B}	99	99
2	NCYC2560	*Kazachstania sinensis*^{B}	99	99
2	NCYC2693	*Kazachstania servazzii*^{B}	99	100
2	NCYC2701	*Kazachstania viticola*^{B,T}	100	100
2	NCYC2702	*Kazachstania kunashirensis*^{B,T}	100	100
2	NCYC2703	*Kazachstania martiniae*^{B,T}	100	100
2	NCYC2729	*Kazachstania africana*^{B}	100	99

Clade	Strain ID	Species name	26S	18S
2	NCYC2827	*Kazachstania rosinii*^{P}	99	100
2	NCYC2878	*Kazachstania barnettii*^{B,T}	100	100
2	NCYC2991	*Kazachstania spencerorum*^{B,T}	100	99
2	NCYC3853	*Kazachstania bulderi*^{B}	100	100
2	NCYC814	*Kazachstania exigua*^{P,T}	100	100
2	NRRLY1556	*Kazachstania unispora*^{N,T}	99	100
2	NRRLY17245	*Kazachstania transvaalensis*^{N,T}	100	100
3	CBS421	*Naumovozyma dairenensis*^{N,T}	100	100
3	NCYC2898	*Naumovozyma castellii*^{B,T}	100	100
4	CBS4332	*Candida castellii*^{B,T}	100	99
4	CBS7729	*Nakaseomyces bacillisporus*^{B,T}	100	99
4	NCYC388	*Candida glabrata*^{B}	100	100
4	NCYC768	*Nakaseomyes delphensis*^{B,T}	100	99
5	CBS4417	*Tetrapisispora phaffii*^{B,T}	99	100
5	CBS6284	*Tetrapisispora blattae*^{N,T}	100	100
5	CBS8762	*Tetrapisispora arboricola*^{B,T}	100	96
5	CBS8763	*Tetrapisispora nanseiensis*^{B,T}	91	98
5	NRRLY27309	*Tetrapisispora iriomotensis*^{N,T}	100	99
6	NCYC2754	*Vanderwaltozyma yarrowii*^{P,T}	100	100
6	NCYC523	*Vanderwaltozyma polyspora*^{B}	99	100
7	NCYC1495	*Zygosaccharomyces bisporus*^{B,T}	100	99
7	NCYC2403	*Zygosaccharomyces mellis*^{P,T}	100	100
7	NCYC2789	*Zygosaccharomyces lentus*^{P,T}	99	99
7	NCYC3000	*Zygosaccharomyces kombuchaensis*^{B}	100	100
7	NCYC568	*Zygosaccharomyces rouxii*^{B,T}	100	99
7	NCYC573	*Zygosaccharomyces bailii*^{B}	100	99
8	NCYC2489	*Zygotorulaspora mrakii*^{B,T}	100	99
8	NCYC2513	*Zygotorulaspora florentinus*^{B,T}	100	99
9	NCYC4020	*Torulaspora delbrueckii*^{B}	100	100
9	NCYC524	*Torulaspora pretoriensis*^{B}	100	99
9	NCYC820	*Torulaspora globosa*^{P,T}	100	99
9	NRRLY1549	*Torulaspora microellipsoides*^{N,T}	100	99
9	NRRLY17532	*Torulaspora franciscae*^{N}	100	99
10	CBS6340	*Lachancea thermotolerans*^{N,T}	99	99
10	NCYC2508	*Lachancea fermentati*^{B,T}	100	99
10	NCYC2644	*Lachancea waltii*^{B,T}	100	99
10	NCYC2875	*Lachancea cidri*^{B,T}	100	99
10	NCYC543	*Lachancea kluyveri*^{B,T}	100	100

Clade	Strain ID	Species name	26S	18S
11	CBS4438	*Kluyveromyces aestuarii*^{B,T}	100	99
11	CBS8778	*Kluyveromyces nonfermentans*^{B,T}	99	98
11	NCYC2559	*Kluyveromyces dobzhanskii*^{B}	100	98
11	NCYC2791	*Kluyveromyces marxianus*^{B,T}	99	99
11	NCYC416	*Kluyveromyces lactis*^{B,T}	100	99
11	UCD54210	*Kluyveromyces wickerhamii*^{N,T}	100	99
12	ATCC58844	*Eremothecium sinecaudum*^{N,T}	100	100
12	CBS106.43	*Eremothecium ashbyi*^{P}	96	97
12	CBS109.51	*Eremothecium gossypii*^{B}	100	99
12	DBVPG7215	*Eremothecium cymbalariae*	99	100
12	NCYC1563	*Eremothecium coryli*^{P}T	100	100
13	AWRI3580	*Hanseniaspora uvarum*^{N}	100	100
13	CBS2592	*Hanseniaspora occidentalis*^{P,T}	100	99
13	CBS285	*Hanseniaspora lindneri*^{B,T}	95	100
13	NCYC31	*Hanseniaspora osmophila*^{B,T}	100	100
13	NCYC36	*Hanseniaspora vineae*^{B}	99	99
13	NCYC4006	*Hanseniaspora valbyensis*^{P}	100	100
13	UTAD222	*Hanseniaspora guilliermondii*^{N}	100	99
14	NCYC3345	*Saccharomycodes ludwigii*^{B}	99	99
Outgroup	NCYC18	*Wickerhamomyces anomalus*^{B}	100	99

Table 2.1: Results of BLAST querying of species-specific 26S and 18S rDNA gene sequences for 75 Saccharomyces complex species and outgroup. Superscripts: B - strains sequenced within a BBSRC-funded project to the NCYC; P - strains sequenced within this project; N - NCBI-sourced genomes; and T - Type strains.

Clade	Strain ID	Species name	Complete	Single-copy	Duplicated	Fragmented	Missing
1	CR85	*Saccharomyces kudriavzevii*	98.0%	97.4%	0.6%	0.7%	1.3%
1	NCYC2578	*Saccharomyces bayanus*	93.0%	81.9%	11.1%	3.2%	3.8%
1	NCYC2888	*Saccharomyces mikatae*	98.2%	97.7%	0.5%	0.8%	1.0%
1	NCYC2890	*Saccharomyces cariocanus*	98.1%	97.5%	0.6%	0.7%	1.2%
1	NCYC3662	*Saccharomyces paradoxus*	97.7%	97.1%	0.6%	0.8%	1.5%
1	NCYC392	*Saccharomyces pastorianus*	98.1%	86.6%	11.5%	0.9%	1.0%
1	NCYC78	*Saccharomyces cerevisiae*	98.3%	97.7%	0.6%	0.7%	1.0%
2	DBVPG7206	*Kazachstania turicensis*	86.1%	85.4%	0.7%	7.7%	6.2%
2	NCYC1417	*Kazachstania lodderae*	97.9%	96.8%	1.1%	0.8%	1.3%
2	NCYC2449	*Kazachstania telluris*	97.9%	59.2%	38.7%	1.0%	1.1%
2	NCYC2450	*Candida humilis*	79.1%	47.5%	31.6%	10.5%	10.4%
2	NCYC2483	*Kazachstania piceae*	97.4%	96.5%	0.9%	1.2%	1.4%
2	NCYC2560	*Kazachstania sinensis*	97.0%	96.2%	0.8%	1.2%	1.8%
2	NCYC2693	*Kazachstania servazzii*	95.7%	74.8%	20.9%	1.8%	2.5%
2	NCYC2701	*Kazachstania viticola*	96.9%	66.6%	30.3%	0.6%	2.5%
2	NCYC2702	*Kazachstania kunashirensis*	96.8%	95.4%	1.4%	1.2%	2.0%
2	NCYC2703	*Kazachstania martiniae*	97.4%	96.5%	0.9%	1.0%	1.6%
2	NCYC2729	*Kazachstania africana*	98.6%	98.1%	0.5%	0.4%	1.0%
2	NCYC2827	*Kazachstania rosinii*	97.3%	96.4%	0.9%	1.2%	1.5%
2	NCYC2878	*Kazachstania barnettii*	96.6%	95.5%	1.1%	1.4%	2.0%
2	NCYC2991	*Kazachstania spencerorum*	96.9%	95.8%	1.1%	1.3%	1.8%
2	NCYC3853	*Kazachstania bulderi*	87.4%	70.0%	17.4%	7.9%	4.7%

Clade	Strain ID	Species name	Complete	Single-copy	Duplicated	Fragmented	Missing
2	NCYC814	*Kazachstania exigua*	97.2%	24.6%	72.6%	1.1%	1.7%
2	NRRLY1556	*Kazachstania unispora*	95.8%	95.1%	0.7%	1.3%	2.9%
2	NRRLY17245	*Kazachstania transvaalensis*	91.8%	91.4%	0.4%	1.8%	6.4%
3	CBS421	*Naumovozyma dairenensis*	97.8%	97.1%	0.7%	1.0%	1.2%
3	NCYC2898	*Naumovozyma castellii*	98.5%	96.3%	2.2%	0.6%	0.9%
4	CBS4332	*Candida castellii*	90.7%	90.2%	0.5%	3.0%	6.3%
4	CBS7729	*Nakaseomyces bacillisporus*	95.3%	94.6%	0.7%	1.3%	3.4%
4	NCYC388	*Candida glabrata*	97.9%	97.4%	0.5%	0.8%	1.3%
4	NCYC768	*Nakaseomyes delphensis*	96.2%	95.4%	0.8%	1.8%	2.0%
5	CBS4417	*Tetrapisispora phaffii*	99.1%	22.0%	77.1%	0.5%	0.4%
5	CBS6284	*Tetrapisispora blattae*	96.2%	94.7%	1.5%	1.2%	2.6%
5	CBS8762	*Tetrapisispora arboricola*	98.0%	96.7%	1.3%	0.7%	1.3%
5	CBS8763	*Tetrapisispora nanseiensis*	96.7%	96.5%	0.2%	1.1%	2.2%
5	NRRLY27309	*Tetrapisispora iriomotensis*	97.6%	96.5%	1.1%	0.8%	1.6%
6	NCYC2754	*Vanderwaltozyma yarrowii*	96.7%	94.0%	2.7%	1.1%	2.2%
6	NCYC523	*Vanderwaltozyma polyspora*	98.4%	95.9%	2.5%	0.7%	0.9%
7	NCYC1495	*Zygosaccharomyces bisporus*	97.9%	97.7%	0.2%	0.9%	1.2%
7	NCYC2403	*Zygosaccharomyces mellis*	98.3%	98.1%	0.2%	0.7%	1.0%
7	NCYC2789	*Zygosaccharomyces lentus*	97.7%	97.3%	0.04%	1.1%	1.2%
7	NCYC3000	*Zygosaccharomyces kombuchaensis*	98.1%	97.7%	0.4%	1.0%	0.9%
7	NCYC568	*Zygosaccharomyces rouxii*	98.7%	98.5%	0.2%	0.4%	0.9%
7	NCYC573	*Zygosaccharomyces bailii*	97.7%	97.3%	0.4%	1.1%	1.2%

Clade	Strain ID	Species name	Complete	Single-copy	Duplicated	Fragmented	Missing
8	NCYC2489	*Zygotorulaspora mrakii*	97.2%	97.1%	0.1%	0.8%	2.0%
8	NCYC2513	*Zygotorulaspora florentinus*	97.9%	97.7%	0.2%	0.8%	1.3%
9	NCYC4020	*Torulaspora delbrueckii*	98.5%	98.3%	0.2%	0.5%	1.0%
9	NCYC524	*Torulaspora pretoriensis*	98.3%	98.2%	0.1%	0.5%	1.2%
9	NCYC820	*Torulaspora globosa*	98.0%	97.9%	0.1%	0.6%	1.4%
9	NRRLY1549	*Torulaspora microellipsoides*	98.2%	97.7%	0.5%	0.5%	1.3%
9	NRRLY17532	*Torulaspora franciscae*	98.2%	97.6%	0.6%	0.3%	1.5%
10	CBS6340	*Lachancea thermotolerans*	98.0%	97.7%	0.3%	0.9%	1.1%
10	NCYC2508	*Lachancea fermentati*	98.3%	98.1%	0.2%	0.6%	1.1%
10	NCYC2644	*Lachancea waltii*	97.4%	97.2%	0.2%	1.1%	1.5%
10	NCYC2875	*Lachancea cidri*	97.8%	48.4%	49.4%	0.9%	1.3%
10	NCYC543	*Lachancea kluyveri*	98.6%	98.5%	0.1%	0.5%	0.9%
11	CBS4438	*Kluyveromyces aestuarii*	96.6%	96.4%	0.2%	1.3%	2.1%
11	CBS8778	*Kluyveromyces nonfermentans*	95.3%	95.0%	0.3%	1.8%	2.9%
11	NCYC2559	*Kluyveromyces dobzhanskii*	97.0%	96.8%	0.2%	1.0%	2.0%
11	NCYC2791	*Kluyveromyces marxianus*	93.4%	64.6%	28.8%	4.0%	2.6%
11	NCYC416	*Kluyveromyces lactis*	97.7%	94.3%	3.4%	0.6%	1.7%
11	UCD54210	*Kluyveromyces wickerhamii*	92.5%	92.3%	0.2%	3.0%	4.5%
12	ATCC58844	*Eremothecium sinecaudum*	94.5%	94.3%	0.2%	1.3%	4.2%
12	CBS106.43	*Eremothecium ashbyi*	19.5%	19.5%	0.0%	22.7%	57.8%
12	CBS109.51	*Eremothecium gossypii*	96.4%	96.1%	0.3%	1.2%	2.4%
12	DBVPG7215	*Eremothecium cymbalariae*	96.9%	96.8%	0.1%	0.8%	2.3%

Clade	Strain ID	Species name	Complete	Single-copy	Duplicated	Fragmented	Missing
12	NCYC1563	*Eremothecium coryli*	95.0%	94.8%	0.2%	1.9%	3.1%
13	AWRI3580	*Hanseniaspora uvarum*	51.2%	51.0%	0.2%	2.6%	46.2%
13	CBS2592	*Hanseniaspora occidentalis*	84.1%	83.9%	0.2%	2.9%	13.0%
13	CBS285	*Hanseniaspora lindneri*	54.8%	54.4%	0.4%	4.7%	40.5%
13	NCYC31	*Hanseniaspora osmophila*	84.7%	84.6%	0.1%	3.4%	11.9%
13	NCYC36	*Hanseniaspora vineae*	84.0%	82.3%	1.7%	3.1%	12.9%
13	NCYC4006	*Hanseniaspora valbyensis*	45.1%	44.8%	0.3%	8.3%	46.6%
13	UTAD222	*Hanseniaspora guilliermondii*	52.7%	52.4%	0.3%	3.9%	43.4%
14	NCYC3345	*Saccharomycodes ludwigii*	85.3%	84.5%	0.8%	7.1%	7.6%
Outgroup	NCYC18	*Wickerhamomyces anomalus*	93.2%	43.0%	50.2%	2.7%	4.1%

Table 2.2: BUSCO gene count information of 75 Saccharomyces complex species and outgroup

2.4 Discussion

The quality control steps used were highly necessary for this project and shed light on a crucial but sometimes overlooked part of working with biological data. Many in-house sequenced and publicly available genome assemblies were, upon investigation, found to be misclassified or extensively contaminated. This extended quality control analysis identified 75 of the 76 genomes as the expected species, with uncertainty surrounding the final species dataset.

With regard to sequence quality, some genomes were found to have assembled poorly which, although this can affect downstream inferences such as identifying genes, were accepted for further analysis as the phylogenetic signal was likely to be strong enough for this study. However, this would only be acceptable were signal-weakening factors such as contamination to be ruled out first.

Whilst poor assembly quality can be the result of low read coverage or contamination, it can also be the use of an inappropriate assembly parameter (e.g. k-mer size) for a specific genome. Testing different assembly software can also be the solution to a more accurate assembly. In this project the ABySS assembly software was used for all genomes with an additional assembly with the SPAdes software for some of the more recently sequenced genomes which failed to assemble well with ABySS. SPAdes did not always increase the assembly quality, so testing different approaches to find the best option for one's data is the optimal choice.

Assessing the assembly quality involved checking the number of contigs, the N50 statistic and the k-mer distribution. A number of genomes had lower than expected N50 scores and high contig counts. After passing other methods of QC, such as having relatively normal k-mer distributions, reasonable genome sizes and having identified as the correct species with BLAST, these genomes were determined to be sufficient for the phylogenetic study. Any further inferences derived from poorly assembled or potentially contaminated species' genomes were made with great caution.

Identifying the species provenance of each dataset was carried out with BLAST rDNA gene alignment and, in some cases, Kraken database mapping. BLAST alignment of fundamental conserved genes is the most common approach to species identification when a reference genome is not available. The decision was made to use 26S (28S) and 18S regions of the rDNA as query sequences to match against target genome assemblies. The additional ITS region would have added further to confirming or rejecting species identity in this project and is recommended. However while there is a large number of ITS gene sequences available on GenBank, some of the older ones must be used with caution as misclassification of species and strains via phenotypic identification is a possibility.

One of the issues with using such a BLAST approach includes selecting an appropriate sequence identity cut-off for a match. Ideally species identity would be indicated by a 99-100% identity score between query and target sequences, although in some cases this could be too high or low. A 2016 large scale barcode sequencing project found 9.5% of species' LSU regions were indistinguishable and so a second region, the ITS, was added which brought down the proportion of unidentified strains to 3% (Vu et al. (2016)). In the study presented here, the 18S region was used instead of the ITS region as it is a longer gene and although it is more highly conserved than the ITS region it was hoped this would add to the ability to distinguish between species with very similar rDNA genes. A lower limit of 99% sequence identity, however, does not allow for much variation in these regions. These are highly conserved genes but there can still be higher than expected variation between strains within a species. Lowering the species threshold below 99% may help identify a species but it must be done with caution. In this study a threshold of 95% was selected, with the vast majority matching 99% or 100% to the expected species. One species matched 91% to the 26S region which is believed to be too low a threshold but was kept in the dataset and noted as likely misclassified. At present the source of the missclassification is unknown. If time had permitted this strain or another, which was re-sequenced unsuccessfully, would have been ordered from another collection and sequenced.

The BLAST approach requires prior knowledge of the identity of a given species. If a high quality match to expected species' genes is not achieved then one must take another approach to identify the species. Mapping one's genome to a large genome database is becoming a popular approach, as exemplified by software such as Kraken. At the time of Kraken quality control on an initial species set a fungal database was unavailable, hence the generation of one here. As of 2020, a fungal database was recently made available with all RefSeq fungal genomes. This could be further added to with GenBank assembled genomes if more yeast species were needed. Doing so would require some form of species identity check (such as the BLAST process used here) as a number of these genomes are misclassified as was found personally. With a high quality database one can at the very least classify by genus and possibly even identify contaminated samples. Kraken also has the potential to be used for identifying hybrid species and to give their proportional make up. Preliminary findings testing hybrid genomes *Saccharomyces pastorianus* and *Saccharomyces bayanus* highlighted this potential use (data not shown). There is confidence in the identity of all but one of the species in this dataset through the use of the BLAST and/or Kraken approach. With regard to contamination, it is noted that minor contamination of some assemblies is still possible, as is also possibly the case with many assemblies in the public domain. Going forward, taking a metagenomics approach to genome assembly may be an interesting way to proceed.

The lower confidence in the species identity and possible contamination of a subset of strains led to the gathering of information on k-mer distributions, genome size, gene counts, GC content and BUSCO gene counts. It would be expected that a mixed species sequence dataset would, depending on the proportion of each species' DNA present, have a larger than expected genome size and more genetic words or k-mers. The genome sizes of all 76 datasets were all recorded along with k-mer distributions with the intention of identifying any contaminated samples. Genome size alone is not sufficient to do so as yeast are known to undergo hybridisation and whole genome duplication events. Further-

more, strain representatives of some species had not been sequenced before, so reliable size estimates were not available. The k-mer counting tool Jellyfish was used to assess k-mer distribution, including unique, distinct, total k-mers as well as the max count of a k-mer. No correlation was found between unique k-mer or max count of a k-mer and increasing genome size (data not shown). A correlation between distinct k-mers and genome size and total k-mers and genome size was, however found ($r^2 = 0.6248$ and $r^2 = 0.8553$ respectively) (See Figures A.1 and A.2 of the Appendix). The plot of total k-mers and genome size showed one clear outlier, NCYC3345 (*Saccharomycodes ludwigii*, clade 14) with more than double the number of total k-mers compared to genome size. One hypothesis for this finding is contamination, which would require further investigation.

Although Jellyfish has the potential to detect contamination, when the expected genome size and breadth of k-mer/genetic diversity is unknown for a species, it can be hard to definitively prove contamination. The effects of poor assemblies on k-mer measures were also assessed by comparing N50 scores to the various k-mer measures, with no correlations found (data not shown here). Ideally, a study with known mixed sample sequences would be undertaken to more clearly show the usefulness of this approach in detecting contaminated sequence.

The BUSCO analysis showed that 64 out of 76 species had more than 90% of complete BUSCO genes. The assemblies with particularly low complete BUSCO gene counts may be so for a number of different reasons. One factor is a poor genome assembly, which can result in fragmented and even missing BUSCO genes, particularly for low coverage datasets where contigs have not spanned, or spanned fully, all gene sequences in a genome. The proportion of fragmented genes or missing genes and number of contigs or N50 was assessed here again but showed no correlation in this dataset. Another possibility is that the BUSCO geneset does not reflect a common core of all species of the group Saccharomycetales. For some low BUSCO gene count strains, including publicly available ones (*Hanseniaspora uvarum* and *Hanseniaspora guilliermondii*), many genes were missing although the genome assemblies appeared of a good quality. This leads one to question

what findings are a data issue and what are real biological signals. The BUSCO approach did add confidence to datasets with lower scores for other QC statistics but also raised questions regarding the genomes of these low gene count species. The BUSCO data was explored further in the Comparative Genomics chapter (Chapter 4).

2.5 Conclusions

Quality control of sequencing data is extremely important in order to make accurate biological inferences. There are many places along the DNA sequencing pipeline where error can occur. Taking a multi-pronged stringent approach to quality control, as shown here, can greatly reduce the chance of the incorporation of a contaminated or incorrect strain. The steps taken here identified a number of misclassified, contaminated and low quality sequencing datasets from in-house sequenced and publicly available datasets, reflecting the need for this crucial step. Ultimately, this approach identified 75 of the 76 genomes as the expected species, with uncertainty surrounding the final species dataset. In the next chapter, entitled Core Genome of a Saccharomyces complex dataset, we look in greater detail at the genes which have been conserved across such a diverse dataset.

Chapter 3

Core Genome of a Saccharomyces complex dataset

3.1 Summary

- The core genome of forty NCYC Saccharomyces complex species is investigated.
- Different sequence read mapping and assembly approaches are assessed.
- Core proteins at varying thresholds of similarity are identified.
- Phylogenetic signal between different core protein sets are assessed.

3.2 Introduction

Since the proliferation of genome sequencing in the biological sciences, the concepts of core genome, pan genome and accessory genome have become frequently discussed and analysed. What is the difference between them? The core genome represents the genes present in all strains of a species or taxonomic grouping (e.g. clade) (See Figure 3.1). It typically includes housekeeping genes for cell envelope or regulatory functions. The pangenome is the entire gene set of all strains of a

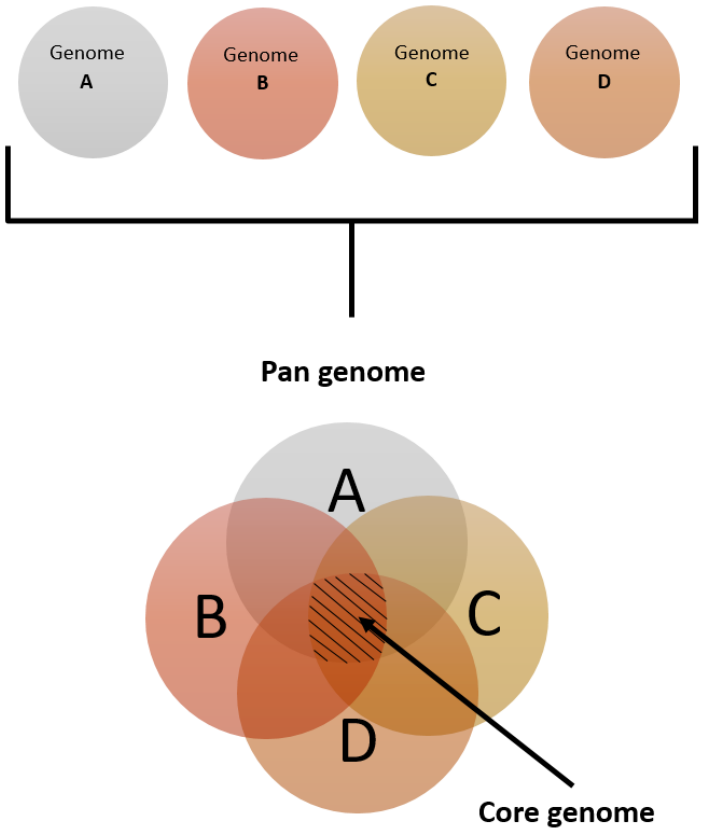


Figure 3.1: Pan, core and accessory genome. The pan genome is all genes present in all strains (Union of all genomes). The core genome is the genes shared in all strains (Intersection of all genomes). The accessory genome is the genes not present in all strains.

species. Where genes are present only in some strains of a species this is considered the variable or accessory genome. These genes include those present in two or more strains or genes unique to a single strain only, for example, genes for strain specific adaptation such as antibiotic resistance in bacteria.

What is the advantage of identifying the core, pangenome or accessory genome of a group of strains? This can aid in characterizing strains or species by their individual gene set (e.g. detecting virulence factors or commercially advantageous traits only present in one particular strain of a species) which may help with developing vaccines or anti-fungal drugs for that specific strain or species. The accessory genome in particular was found to be an important contributor to pro-

tein antigens in bacteria (Mora et al. (2006)). This implies that for many bacterial species, a protein-based universal vaccine would only be possible by including a combination of antigens from the core and the accessory genomes. It can also be useful for detection, identification and tracking of new strains in metagenomic samples based on their individual gene subset of the species pangenome.

Looking at the core genome may also shed light on questions such as what are the types of core genes shared by eukaryotic species? One may be interested in the size of a species pangenome, of a clade core genome, or of the eukaryotic pangenome. There are complications when estimating these gene sets for eukaryotes including the abundance of transposable elements, hybridisation, whole genome duplication, the presence of large gene families and the relative incompleteness of genomic sequences, particularly of those containing numerous repeats.

A few studies on fungal core and pangenomes have been undertaken to date. Hsiang and Baille in 2005 compared *Saccharomyces cerevisiae* to distant fungal species (and non-fungal species) and found that 17 of the core genes appeared to be unique among the fungi studied (Hsiang and Baillie (2005)). Of these 17 genes, two were found to be involved in protein biosynthesis, two in transport, and one in sporulation. A more recent study looked at the pangenome of within and across fungal species including *S.cerevisiae*, *Candida albicans*, *Cryptococcus neoformans var. grubii* and *Aspergillus fumigatus* (McCarthy and Fitzpatrick (2019)). Using Gene Ontology enrichment to look at the fungal core genomes of this set of species, the authors showed that many housekeeping biological processes, such as translation, nucleic acid metabolism and oligopeptide metabolism, were significantly over-represented in each species ($P<0.05$). Furthermore, molecular function terms for enzymatic and nucleic acid binding activity were also significantly over-represented.

Among all eukaryotes presently sequenced, ascomycetous yeasts are arguably one of the most well-described phyla with the pangenomes of *Saccharomyces cere-*

visiae, *Candida glabrata*, *Candida albicans* as well as *Schizosaccharomyces* species having been studied (Peter et al. (2018), Carreté et al. (2018), Rhind et al. (2011), Gabaldón and Fairhead (2019)). In 2018, Peter *et al.* identified 4,940 core open reading frames (ORFs) across 1,011 *S. cerevisae* genomes. These isolates were sourced from around the world and were from domesticated, wild, or human origin (mainly clinical). This sequencing effort enabled the determination that Chinese and Taiwanese strains were closer to *Saccharomyces paradoxus* and to the root of the *Saccharomyces* genus than strains from any other origin, strongly supporting a single out-of-China origin for *S. cerevisiae*, that subsequently spread all over the planet (See Figure 3.2). As mentioned previously, eukaryotic genomes may harbour introgressed genomic segments and undergo horizontal gene transfer (HGT), both of which were indeed seen in Peter et al. (2018) with 913 and 183 cases found of these two phenomena respectively. Half of the HGT ORFs could be traced to other Saccharomyces complex species belonging to the Torulaspora or Zygosaccharomyces clades. These yeasts are known to share similar environmental fermentative niches, which likely favored frequent transfer of genetic material between species.

Defining the core, pan and accessory genomes of a group of organisms relies on establishing the orthologs amongst them. Orthologous genes diverge from a most recent common ancestor (MRCA) due to speciation. Parologous genes diverge from MRCA due to duplication. There are three types of methods for identifying orthologs: tree-based, graph-based and hybrid methods. Tree-based methods infer orthologs and paralogs by comparing phylogenetic trees estimated from homologous (common ancestry) gene sequences to species trees. Graph-based methods use pairwise alignments of genes to determine similarity between proteins. Hybrid methods use a combination of tree- and graph-based methods and are the most frequently used. In core or pan-genome analysis, the sequence unit for the modeling could be for example genes, clusters of orthologous groups (COGs) (Tatusov et al. (1997)), coding sequences (CDS), proteins, concatenated genes etc. There are a number of software approaches available to identify the core genome depending on the desired input or output and model parameters.

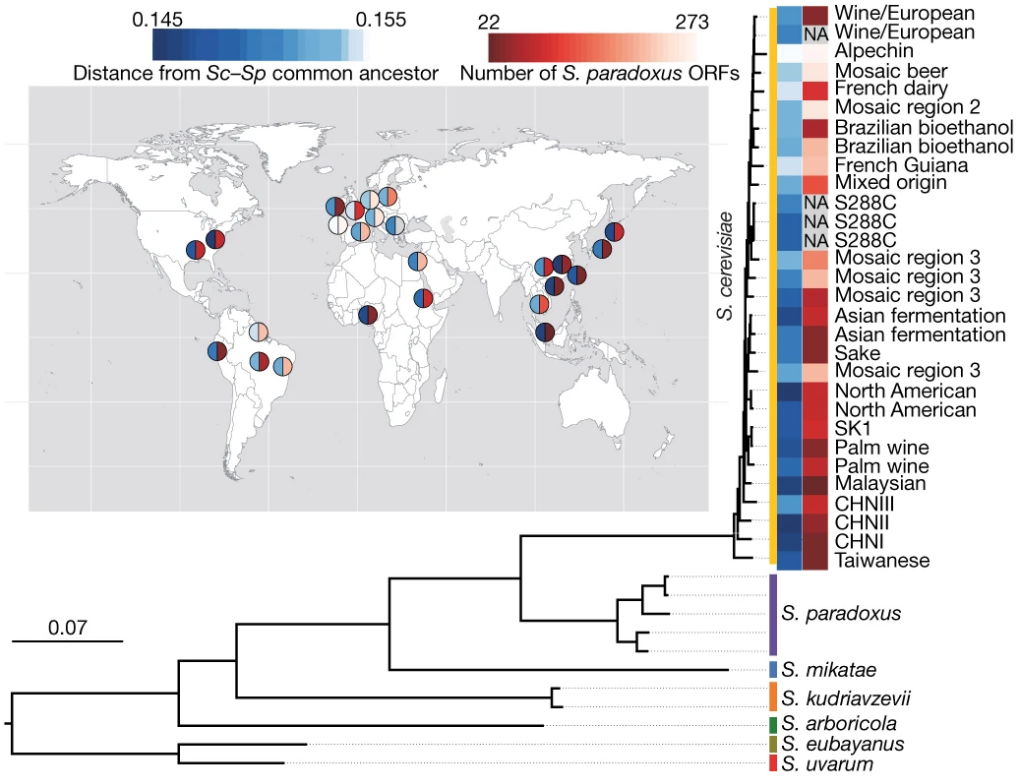


Figure 3.2: Figure taken from Peter et al. (2018) Figure 2. Maximum-likelihood rooted tree of the *Saccharomyces senso scripto* species, based on the alignment of 2,018 concatenated conserved genes. Heat maps display the distance from the last common ancestor of *Saccharomyces cerevisiae* (Sc)–*Saccharomyces paradoxus* (Sp) (white–blue), and the number of introgressed *S. paradoxus* ORFs (white–red). The map shows the geographical origins of the strains.

BPGA (Chaudhari (2016)), Roary (Page et al. (2015)), PanOCT (Fouts et al. (2012)) and PGAP (Zhao et al. (2012)) are just some commonly used software for identifying orthologous gene clusters in microbial genomes.

In the past, bacterial comparative genomic analyses started by exploiting ∼0.07% of a genome (16S rRNA) (Woese (1987)), later on using up to ∼0.2% of the genomic information (Multi Locus Sequence Typing) (Maiden et al. (1998)), and more recently up to 100% of the information exploiting the pangenome (Tettelin et al. (2005)). The largest amount of genomic information (whole genome) is potentially ideal for estimating an accurate phylogenetic tree, but could a core genome be sufficient when the computational resources are not available? If so, what threshold of ortholog identity should be chosen? These questions are investigated in this chapter.

These prior studies collectively show that identifying the core genome of a large and diverse group of species has many uses. It can aid in understanding the phylogenetic history of a group of intra- or inter-species strains. In yeasts it also has the potential for identifying essential genes which have uses within industry or within medicine as anti-fungal targets. There are a number of approaches to identifying core genes, depending on the input data, but a surprisingly small number of studies have been carried out on a diverse yeast dataset to date. The aim of this chapter was to shed light not just on the diversity of Saccharomyces complex species but also the genomic similarities across the species. Forty species across 11 clades and an outgroup were used for this study. Different sequence read mapping and assembly approaches were assessed to help identify the number and identity of the core genes. A clustering algorithm, BPGA was also used to identify core genes at varying thresholds of similarity. Finally, these genesets were used to build phylogenetic trees and were compared to a whole proteome tree to assess differences in phylogenetic signal.

3.3 Methods

3.3.1 Dataset

Whole genome sequences from forty Saccharomyces complex yeast species across 11 clades and one outgroup species (NCYC18: *Wickerhamomyces anomalus*) from within the National Collection of Yeast Cultures (NCYC) were selected for the analysis (See Figure B.1 of Appendix for species list). All of these genomes are included in the larger 76 species dataset described in the previous chapter, with the exception of NCYC2739 (*Hanseniaspora uvarum*). This genome was later removed from the full dataset as it had an unacceptably low 26S identity match of 87% to the reference genes from this species and was replaced by a publicly available version of the species (AWRI3580; GCA001747055.1).

3.3.2 Mapping-based core genome prediction

Trimmed sequence reads (see previous chapter for details) from each of the forty-one strains were mapped to two different reference genomes, *Saccharomyces cerevisiae* (S288c; Accession: GCA000146045.2) and *Candida glabrata* (CBS138; Accession: GCF000002545.3) with Stampy (Lunter and Goodson (2011)) v1.0.31 (default settings) and NextGenMap (Sedlazeck et al. (2013)) v0.5 (default settings). Different divergence parameters were tested with Stampy (default = 0.001), ranging from 0.001 to 0.1, to see whether this affected the proportion of reads mapping to the reference genome. However, little difference in the numbers of mapped reads was observed between parameter settings (data not shown), so a final value of 0.001 was used. Samtools (Li et al. (2009)) v1.9 was used to extract the mapped reads for counting and further processing.

The genome assembler software ABySS (Simpson et al. (2009)) v1.9.0, with the -k = 64 option was used to assemble the trimmed and sorted paired-end reads which had mapped to the *Candida glabrata* and *Saccharomyces cerevisiae* reference genomes into gene-based contigs. The Trinity RNA-Seq assembler (Grabherr et al. (2011)) v2.6.5 was also used to assemble gene-based contigs from the same paired-end reads due to its focus on shorter genomic segments. The

align and estimate abundance Perl script was used following Trinity RNA-Seq assembly, with *kallisto* selected as the estimation method. The AUGUSTUS gene prediction software was used to extract and translate coding genes from these gene-based contigs. The identities and locations of annotated proteins within the reference genomes were obtained from the Saccharomyces Genome Database (https://www.yeastgenome.org/) and the Candida Genome Database (http://www.candidagenome.org/) respectively. The predicted protein sequences from the gene-based contigs were then queried against these reference protein sequence sets using *blastp* and the number of unique protein hits counted.

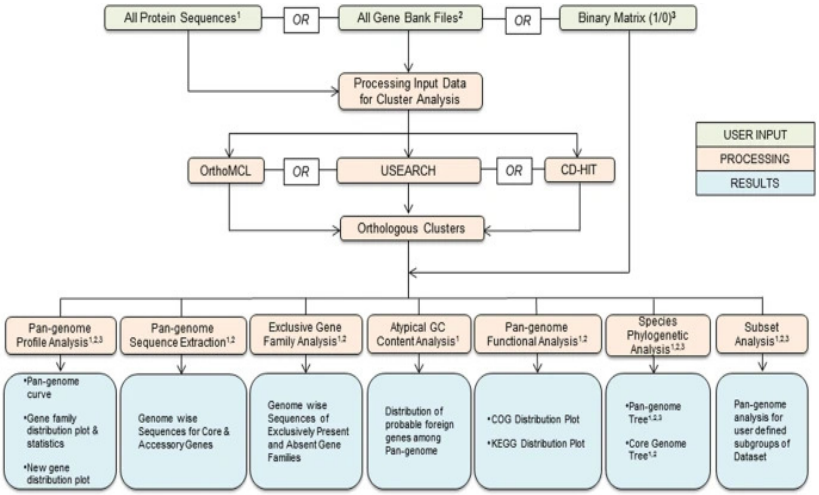


Figure 3.3: A depiction of the BGPA workflow, showing the various input format, clustering algorithm and output choices. Taken from Chaudhari (2016) Figure 1.

3.3.3 BPGA-based core genome prediction

The AUGUSTUS gene prediction software was used to estimate and translate coding sequences from the 40 Saccharomyces complex species. This dataset was subsequently used as input to the Bacterial Pan Genome Analysis (BPGA) toolkit (Chaudhari (2016)) v1.3 for core genome estimation. BPGA is a Perl software tool that performs clustering between genes or proteins in the individual input strains or species using third-party software. Clustering can be performed by one of three dependant tools (OrthoMCL, USEARCH or CD-HIT) and at a chosen level of sequence identity (default value 50%). The output of the clustering

step is presented as a tab delimited gene or protein presence/absence matrix (the pan-matrix) which can then be used for a range of downstream analyses, including pan genome phylogeny estimation and COG/KEGG gene assignment. Figure 3.3 illustrates the BGPA analytical workflow. Within this study, the USEARCH clustering tool was selected for inter-protein sequence comparison and five different sequence identity cut-offs - 50%, 75%, 80%, 85% and 90% - were assessed.

3.3.4 BLAST annotation of core proteins

Both the assembled mapped reads and BPGA 'core' proteins from the two core genome prediction pipelines were queried against the relevant coding annotation database using blastp (Altschul et al. (1990)) v2.2.30 in order, where possible, to determine their identity and putative function. All accepted matches had e-values ≤ 0.

3.3.5 Core protein tree-building

Five phylogenetic trees were built with FFP (Sims et al. (2009a)) v3.19 (k = 14, amino acid alphabet) of the 40 species dataset and outgroup with varying numbers of 'core proteins'. The choice of k-mer length is explained in detail in the Phylogenetic comparison chapter (Chapter 5, section 5.4.3). Each dataset comprised one of the BPGA 'core' protein sets, corresponding to the chosen sequence identity cut-off, with the exception of the outgroup which contained the full proteome (n = 6,043). Trees were viewed with Figtree (Rambaut and Drummond (2009)) v1.4.2 and compared to an FFP amino acid alphabet tree which used the full proteome (average n = 5,438).

3.3.6 Tree comparison metrics

The Robinson-Foulds distance metric (see Introduction chapter), both weighted and unweighted, were obtained for each pairwise comparison of the FFP proteome and core protein set Newick trees using the dendropy library (Sukumaran and Holder (2010)) in Python (v.2.7.12). The Kendall–Colijn metric,

both weighted and unweighted, was obtained with the treescape package (Jombart et al. (2015),v.1.10.18) in R (R Development Core Team (2008),v.3.3.2).

3.4 Results

3.4.1 Read mapping

As part of the phylogenetic method comparison discussed in Chapter 5, a SNP tree approach was undertaken, where SNPs were identified by mapping sequence reads to a common reference genome, with the resulting variable sites used to estimate a phylogeny. The results indicated that a large number of sequence reads were unable to map to the S288c reference genome. Because of that observation, the reads in this analysis were also mapped to a second reference genome, *Candida glabrata* CBS138. Initially, reads were mapped with Stampy using the default settings. Five genomes from across the Saccharomyces complex as well as the outgroup species were assessed. As shown in Table 3.1 a significant drop in the proportions of sequence reads mapping to the reference genome was observed, for the most part decreasing with evolutionary divergence from *Saccharomyces cerevisiae*. To make sure the read mapping software used was not partly or mainly responsible for the poor mapping, the diversity setting of the mapper (Stampy) was tested, with very little difference found (data not shown), even when accounting for a greater potential evolutionary distance from the reference species. A second mapper was also assessed, NextGenMap, but yielded similar results (data not shown). A closer examination of the reads of one strain, NCYC388 (*Candida glabrata*), that mapped to the S288c genome showed that the largest proportions of reads mapped to the mitochondrial chromosome (40%) and to chromosome 12 (13.6%), indicating a variable conservation of the genome among chromosomes between the two species, and potentially across the species set.

3.4.2 Assembler comparison

Next, the identities of the reference proteins to which the various sequence read datasets mapped were assessed, plus how the choice of genome assembler affected

Species	ID	Clade	% Reads mapped to S288c genome
Saccharomyces cerevisiae	NCYC78	1	98.3
Kazachstania lodderae	NCYC1417	2	25.5
Naumovozyma castellii	NCYC2898	3	21.4
Candida glabrata	NCYC388	4	15.7
Zygosaccharomyces mrakii	NCYC2489	8	12.3
Hanseniaspora osmophila	NCYC31	13	17.3
Wickerhamomyces anomalus	NCYC18	Outgroup	10.5

Table 3.1: Percentage of sequence reads of six Saccharomyces complex strains and outgroup (NCYC18) mapped to the *Saccharomyces cerevisiae* S288c reference genome with Stampy (Lunter and Goodson (2011)) v1.0.31.

the results. Once again using the clade 4 species *Candida glabrata* (NCYC388) as an example, the mapped reads were assembled into contigs using the ABySS and Trinity RNA-Seq assemblers. Both assemblers were tested to assess if one would assemble a greater number of genes than the other. Putative identities of the proteins within the assembled contigs were found using AUGUSTUS and *blastp* comparison to annotated protein databases. Similar, but not identical, numbers of nuclear-encoded proteins resulted from both approaches. Trinity RNA-Seq resulted in 533 proteins, of which 509 were nuclear-encoded proteins and ABySS resulted in 460 proteins, of which 443 were nuclear-encoded proteins. Four hundred and eleven nuclear-encoded proteins (81% of Trinity RNA-Seq- and 93% of ABySS-predicted nuclear-encoded proteins respectively) were found to share a protein name between both sets (See Figure 3.4).

3.4.3 BPGA core proteins

A different approach to looking at the number and identity of core proteins in the 40 Saccharomyces complex species dataset was also undertaken. BPGA, a pan-genome pipeline, was used to find proteins present in all species at a specified sequence identity level. The program took the AUGUSTUS-predicted amino acid

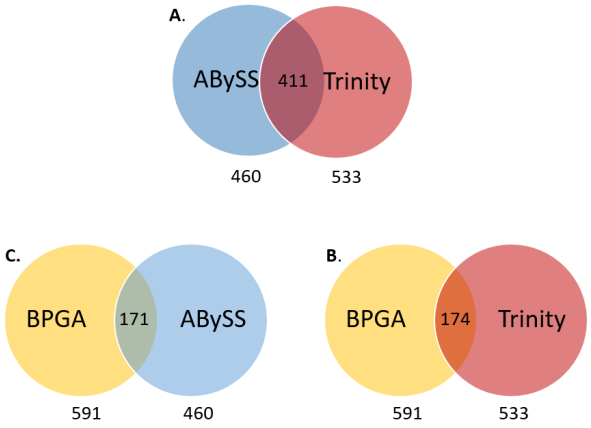


Figure 3.4: Total protein counts and shared protein names of three 'core' protein sets. A. ABySS- and Trinity RNA-Seq-assembled proteins shared 411 protein names. B. BPGA and ABySS-assembled proteins shared 171 protein names. C. BPGA and Trinity RNA-Seq-assembled proteins shared 174 protein names.

sequences of each species' whole genome as input and clustered proteins between datasets at 50%, 75%, 80%, 85% and 90% sequence identities, resulting in 591, 82, 38, 19 and 5 core proteins respectively. The 591 proteins found at the 50% sequence identity level within the NCYC78 genome (*S. cerevisiae*) were then annotated with *blastp*, as described in the Methods section. This number does not account for core proteins encoded by the mitochondrial genome as AUGUSTUS predicts only nuclear genes. Theses 591 proteins were then compared to the 443 ABySS-assembled and 509 Trinity RNA-Seq-assembled nuclear-encoded proteins which had been identified from the S288c mapped reads. The BPGA proteins had 171 proteins in common with ABySS-assembled reads and 174 with Trinity RNA-Seq-assembled reads (See Figure 3.4).

The BLAST-predicted identities of the 591 proteins found by BPGA can be found, by chromosome, annotated according to the *S. cerevisiae* S288c reference genome, within Table 3.2. A closer look at the 19 proteins found at the 85% sequence identity level showed that these proteins are predicted to be involved in fundamental cellular processes including protein synthesis, cell structure and metabolism (See Table 3.3).

Chr. size	230218	813184	316620	1531933	576874	270161	1090940	562643	439888	745751	666816	1078177	924431	784333	1091291	948066
Chr. number	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**	**XIV**	**XV**	**XVI**
protein	EFB1	UTP20	ADP1	PST2	ARG5	RET2	GSC2	PRP8	SPO22	URA2	TOR2	SSQ1	PFK2	TOP2	RPA190	YPL109C
	DRS2	LYS2	PWP2	DBP10	SPF1	GCN20	SLH1	MSR1	NEO1	HAM1	FAS1	YSH1	PLB2	RIA1	MYO2	MOT1
	PMT2	PYC2	SPB1	SUB2	VAC8	GSY1	TAF1	GUT1	KGD1	CPA2	MAK11	MEF1	TSL1	KRE33	RPO31	TRM44
	FUN12	MIS1	HIS4	RPO21	RNR1	FRS2	SEC15	GAR1	SSL2	MTR4	MYO3	VIP1	GCV2	YNL247W	ELG1	NEW1
	ACS1	VMA2	ADY2	CDC53	RPL12A	ACT1	RTG2	SMF2	YIL067C	MRS3	YKT6	DNM1	ERB1	IMP4	RPB2	RPA135
	CYS3	GRS1	RSA4	DNF2	RSP5	DUG1	CHC1	ERG7	THS1	NPA3	MCD4	EMG1	TUB1	SSB2	RFC4	CDC60
	RBG1	ILS1	FEN2	ARO1	RAD3	LPD1	ADE6	QNS1	CCT2	ACO2	YPK1	SEC13	AMD1	CIT1	ALG6	VTC3
		MCM2	NFS1	YCF1	MET6	TUB2	UTP22	YHR020W	DPH1	KAR2	UBA1	YEF3	ERG5	DBP2	RET1	PMA2
		RPL4A	THR4	SLY1	GLC3	RPN11	PUP2	PUT2	LYS1	CCT5	DHR2	CBF5	HSC82	LYS9	DIS3	PRP46
		SEC18	PGK1	SNF1	TRP2	SEC53	GND2	DED81	RPL16A	SSC1	YKL091C	HSP104	ADH3	MLS1	RTC5	MRD1
		RAD16	KRR1	NOP1	HOM3		VAS1	SLT2	HIS5	ILV3	OAC1	MCM5	ADE17	NOP2	ALA1	YME1
		TKL2	LEU2	BPL1	ARB1		POX1	NMD3	LYS12	CCT7	RPT1	RIX7	CPR3	LEU4	TUF1	SEC23
		IFA38	ELO2	COP1	UTP7		DBF2	ERG11	CFD1	CCT3	TPK3	DRS1	PRC1	NOG2	DBP5	MNN9
		HIS7		CYM1	ILV1		HNM1	UBA4	RPL34B	ATP2	TRM2	GPN3	GAS1	LAT1	TOP1	NOP4
		URA7		POL3	GDI1		PFK1	GPA1	RPS24B	MET3	GFA1	JLP1	YPT7	ZWF1	EFT1	DBP1
		PGI1		SAC6	NUG1		RAD54	APE4	MMF1	ARP3	VPS1	PDC5	NDI1	AAH1	CYT1	NOG1
		ATP1		CDC48	SAH1		PMR1	RRP3		BNA3	FBA1	DPS1	IMD4	RIO2	PRO2	RPN7
		PDX3		HSP78	PDA1		SKN1	ARG4		ADO1	KAE1	FRS1	RRB1	ADE12	FAA1	ELP3
		ACH1		TSR1	PMI40		SEC27	ENO2		RNR2	MAE1	ALT1	ALO1	RPD3	IDH2	RPS9A
		CDS1		RPT3	RAD51		STT3	DBP8		ERG20	SDH1	HSP60	GUA1	MVD1	ITR2	HTS1
		SHM1		BCS1	DMC1		PRP43	CDC12		ASF1	PGM1	MAP1	HAS1	SIS1	CIR2	FUM1
		TPS1		TCP1	PUP3		LEU1	BAT1		RPS4A	DCW1	UBI4	TOM40	RFC3	HTZ1	GLR1
		UGA2		SUP35	URA3		MES1	DYS1		ARG3	CCP1	NOP56	ERG13	RPC34	ADE2	HRR25
		ALG7		LYS4	GPP2		TRP5	LEU5		LIA1	UGP1	SOF1	RPS16A	HHT2	MBF1	RVB2
		SUP45		HOM2	HIS1		GUS1	THR1		SUI2	TCD2	SEC61	ARG7	LST8	DGK1	VPS4
		OLA1		NHX1	RPS26B		ZUO1	TRR2		MIR1	OSH6	SHM2	PRE8	OCA1	MET7	CAR1
		ARO4		KRS1	NSA2		LSG1	RPF1		RPE1	TMA19	ADE13	YML6	RPS3	NOP58	TAZ1
		PDB1		RLI1	SPT15		ASN2	VMA16		RPS5	MDH1	HOG1	ERG6	RPS7B	PYK2	RPS6A
		RIM2		NHP2	RPS8B		VHT1	IMP3		TIM17	AUR1	NMA1	ERG2	RHO2	YTM1	SUA7
		FUS3		MNN10	NOP16		ERG4	RPL27A		RPS14B	PRS1	CAR2	GTR1	SUI1	RPT4	RRD2
		PRE7		LCB2	YPT31		GCN5	TIM10		RPS22A	RPF2	MET17	TEM1	HHF2	NPT1	GLN1
		RIB1		HEM1	RIP1		NOP7			RPL17B	APN1	AAT2	COQ5		WRS1	CBP3
		PRS4		GGC1	VMA3		CDC55			SOD1	GPM1	ILV5	YHM2		LIP5	LSP1
		ATP3		DLD2	SNU13		ARO8				MTD1	CPR6	YMR099C		ARG1	RPL5
		PET9		URH1	HYP2		UGA1				YKL069W	VMA6	YMR226C		HEM15	RPC40
		TRM7		SES1	CYC7		CYS4				GPX1	RPP0	RPL13B		CPA1	PRE2
		IPP1		SIT4			OLE1				MET14	TAL1	RPS1B		SER1	SUI3
		RPL19A		RMD1			ERG1				ARC19	PNP1	URA10		KTR1	COX11
		POL30		RVB1			EMP24				AIM29	NNT1	PRE5		ORT1	ISU1
		RPL32		ADA2			HXK2					DPH5	TSA1		GSP2	RPL7B
		RPL21A		RPT2			TFG2					HCR1	APT1		RPL3	TIF6
		RPB5		PPH22			YGR210C					SDH2	RPL15B		CDC21	RPL1A

		CMD1	IDP1			RPT6					RPL10	RPS17A		GPN2	NIP7
		UBC4	SAM2			PHB1					RLP24			LSC1	VMA11
		NHP6B	HEM12			TYS1					RPL6B			CKA2	YAH1
		HTB2	PSA1			HGH1					RPL37A			FSF1	RPO26
			HEM3			RPS0A								RPL18A	FCY1
			GET3			ARO2								BRX1	RPS23B
			HEM13			ERG26								PRE10	RHO1
			TRM8			RBG2								RPL25	
			PHO13			NBP35								CAT5	
			ARF1			HEM2								PUP1	
			BMH2			ERG25								RPS15	
			RPS13			PHB2								RIB4	
			GUK1			RPL11B								PFY1	
			TPI1			BUD32								RPS12	
			YRB1			PRE9								RPS10A	
			ADK1			RPS2								RPB11	
			HNT1			RPL26B									
			RIB3			RPL24B									
			RPB7			RPL28									
			FCF1			RPL30									
			SDH7			RPS25A									
			UBC13			VMA7									
			UBC9												
			RPS11A												
			RPL31A												
			QCR7												
			RPP1A												
No. proteins:	7	46	13	69	36	10	64	31	16	33	39	46	43	31	58	49
Gene density:	$3.0 \times e^{-5}$	$5.7 \times e^{-5}$	$4.1 \times e^{-5}$	$4.5 \times e^{-5}$	$6.2 \times e^{-5}$	$3.7 \times e^{-5}$	$5.9 \times e^{-5}$	$5.5 \times e^{-5}$	$3.6 \times e^{-5}$	$4.4 \times e^{-5}$	$5.8 \times e^{-5}$	$4.3 \times e^{-5}$	$4.7 \times e^{-5}$	$4.0 \times e^{-5}$	$5.3 \times e^{-5}$	$5.2 \times e^{-5}$

Table 3.2: Five-hundred and ninety-one conserved proteins found at the 50% sequence identity level with BPGA (Chaudhari et al. (2016)) across 40 Saccharomyces complex species, by chromosome, with protein names given according to *blastp*-predicted orthology to proteins within the *S.cerevisiae* S288c reference genome.

protein name	protein function
EFT2/Elongation Factor 2	Catalyses ribosomal translocation during protein synthesis
GSP2/GTP binding protein	Nuclear organisation, RNA processing and transport
HHF2/HistoneH4	Chromatin structure
TCP1	Subunit of T complex, protein folding and actin cytoskeleton maintenance
RBL15B	Ribosomal 60S subunit protein L15B: protein production
RBL12A	Ribosomal 60S subunit protein L12A: protein production
RBL11B	Ribosomal 60S subunit protein L11B: protein production
RBL10	Ribosomal 60S subunit protein L10: protein production
RBL3	Ribosomal 60S subunit protein L3: protein production
RBL1A	Ribosomal 60S subunit protein L1A: protein production
TEF1/Transcription Factor 2B	Catalyses ribosomal translocation during protein synthesis
ELP3	Subunit of Elongator complex
Actin	Structural protein involved in cell polarization, endocytosis, and other cytoskeletal functions
HTA2/Histone2A	Chromatin structure
RPD3	Histone deacetylase, chromatin structure
TIF6/Translation Initiation Factor 6	Constituent of 66S pre-ribosomal particles
SAH1/S-Adenosyl-l-Homocysteine hydrolase	Lipid and methylation metabolism
GLC7	Type 1 Serine/Threonine protein phosphorylation catalytic subunit: glycogen metabolism sporulation, mitotic progression
VMA3/Vascular Membrane Atpase	Vascuolar acidification, metal ion homeostatis

Table 3.3: Names and putative functions of nineteen proteins found conserved at 90% (bold) and 85% sequence identity with BPGA (Chaudhari et al. (2016)) across 40 Saccharomyces complex species.

3.4.4 Core proteins and phylogenetic tree building

The effects of varying numbers of core proteins on phylogenetic tree accuracy was then investigated. This was done by building FFP amino acid trees of the 40 species with the proteins found at the five levels of sequence identity (50%: 591, 75%: 82, 80%: 38, 85%: 19 and 90%: 5 proteins) and comparing them to the tree estimated using the full protein set (average n = 5,438). The full proteome (6,043 proteins) of NCYC18 was used as an outgroup for all trees. As shown in Figure 3.5 and Table 3.4, a tree built with all proteins (amino acid sequences excluding mitochondria) results in the same topology as a tree built with 591 or 82 'core' proteins, indicating that the phylogenetic signal remains down to inclusion of only ∼1.5% of the number of reference proteins. However, when also taking branch length into account, the distance from the full protein set tree grows as the number of proteins used is reduced. As the number of proteins is reduced below 82, both the topology-only and tree-based metrics increase, though the relationship is not monotonic (See Table 3.4).

Tree comparison metric	Topology only measure		Topology and Branch length measure	
Trees	RF-Unweighted	KC-Unweighted	RF-Weighted	KC-Weighted
Core 591 vs proteome	0	0	4.862	0.9816
Core 82 vs a.a	0	0	8.782	1.664
Core 38 vs a.a	10	17.292	9.838	1.736
Core 19 vs a.a	8	2.646	12.236	1.921
Core 5 vs a.a	30	68.007	11.432	1.869

Table 3.4: Tree comparison metrics; Robinson-Foulds (RF) Unweighted and Weighted, Kendall-Colijn metric (KC) Unweighted and Weighted. FFP 20-letter amino acid alphabet trees of 40 Saccharomyces complex strains and outgroup with varying protein content. Proteome tree (average n = 5,438), BPGA identified core protein set trees; 591 (50% sequence identity), 82 (75% sequence identity), 38 (80% sequence identity), 19 (85% sequence identity), 5 (90% sequence identity) proteins.

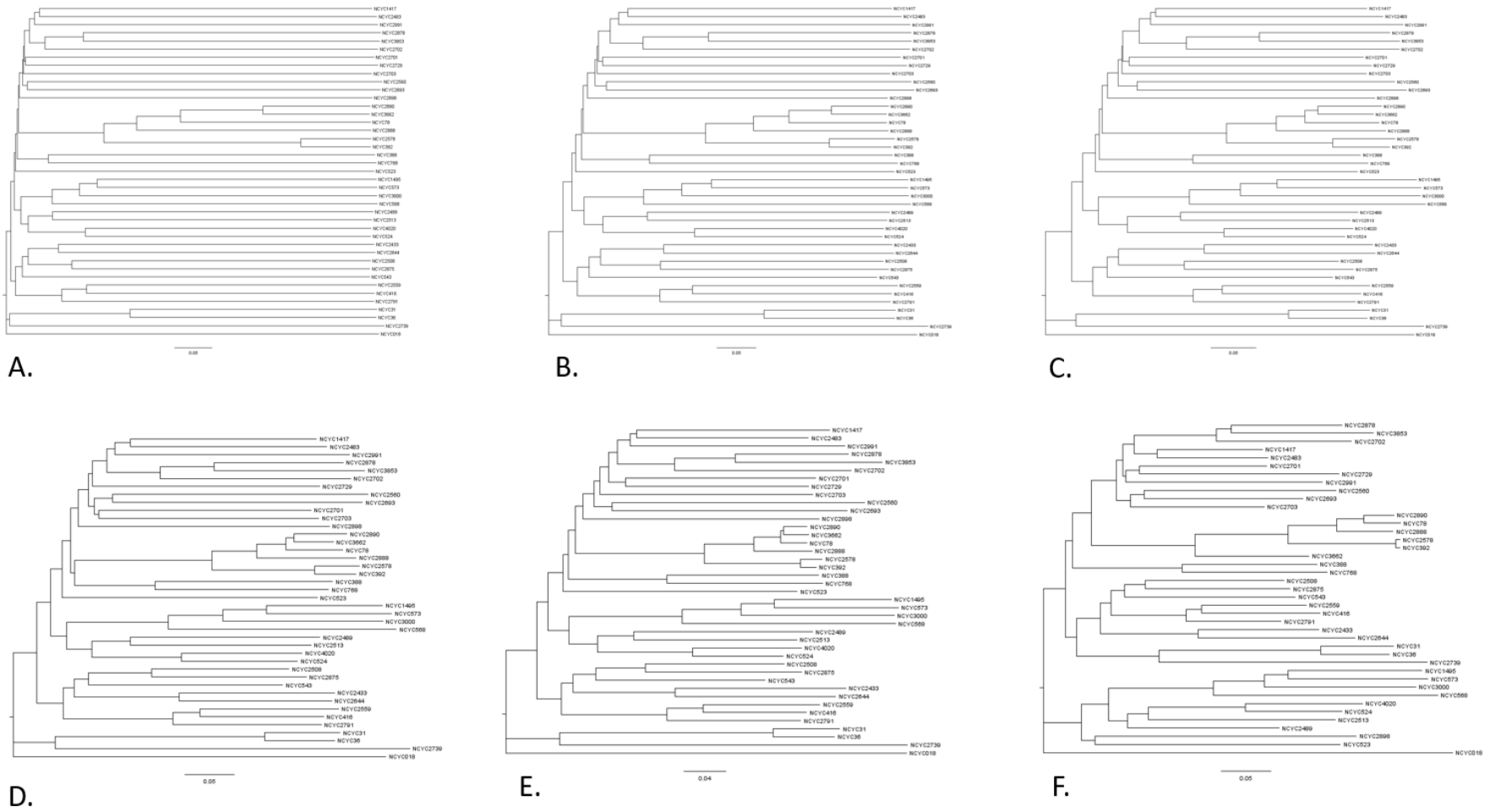


Figure 3.5: FFP 20-letter amino acid alphabet trees of 40 Saccharomyces complex species and outgroup (NCYC18) with varying protein content protein sets. A. Whole proteome tree (average n = 5,438), B. 591 proteins at 50% sequence identity, C. 82 proteins at 75% sequence identity, D. 38 proteins at 80% sequence identity, B. 19 proteins at 85% sequence identity, B. 5 proteins at 90% sequence identity.

3.5 Discussion

Whilst the comparative genomics chapter (Chapter 4) will highlight the sometimes vast genomic differences between species of the Saccharomyces complex, this chapter identifies the similarities. Core protein sets were identified both by mapping strain-specific reads to two reference genomes and by using a third-party software pipeline.

In the first analysis, Stampy was the initial software of choice for mapping sequence reads to a reference genome due to its prior successful use in projects within the NCYC. However, a SNP-based tree produced for the Phylogenetic comparison study contained a subset of taxa with very long branches, requiring further investigation here. The proportion of S288c mapped reads from a *Candida glabrata* (NCYC388, clade 4 species) genome was assessed at different divergence settings with a very low number being mapped. A second mapper was also tested, NextGenMap, which resulted in a similar number of mapped reads. Even when repeating the analysis with a different reference genome, the number of NCYC78 (*S. cerevisiae*) reads which mapped to the *C. glabrata* reference genome (CBS138), a highly similar proportion of mapped reads was observed. These results indicated strongly that the observed mapping bias was simply a reflection of the evolutionary diversity of the Saccharomyces complex rather than a software issue.

The next question was if this low mapping was repeated across the dataset. The proportion of mapped reads from five genomes from across the complex, as well as the outgroup species, were assessed. A maximum of 25.5% of sequence reads from clades other than clade one were found to map to the S288c reference genome. Whilst this is a small set of species which does not cover the majority of the Saccharomyces complex species it seemed reasonable to hypothesise that these results would be reflective of those in the full dataset, particularly in light of the phylogenetic trees built. The observed gradual decrease in mapping as the evolutionary distance from the reference genome increased also makes biological sense. A closer look at the locations of the mapped reads on the S288c genome was made next for the *C. glabrata* (NCYC388) representative. This examination found that the largest proportions of reads mapped to the mitochondrial chromosome (40%) and chromosome 12 (13.6%). Conserved genes are known to be found on both chromosomes. The genes present on the mi-

tochondrial genome of *S.cerevisiae* include those related to energy production such as ATP synthase subunits and cytochrome c oxidase subunits. There are also two rRNAs (21S and 15S) and 24 tRNAs (Foury et al. (1998)) present within this genome. Chromosome 12 of *S.cerevisiae* is also known to harbour essential genes including 22 nuclear tRNA.(Johnston et al. (1997)).

The next step was to find the identities of those proteins to which these reads were mapping. This was done by assembling the mapped reads into contigs, predicting the proteins with Augustus and using *blastp* to count the number of proteins. Two different assembly tools were tested, Trinity RNA-Seq and ABySS, to see if there would be a difference in the resulting number of proteins identified by the two pieces of software. ABySS-assembled contigs resulted in fewer unique proteins being identified by *blastp* than Trinity RNA-Seq (460 and 533 respectively). While the two approaches were found to share 411 common proteins, using Trinity RNA-Seq appeared to offer superior results, likely due to the gene-based nature of the datasets.

Next, these core protein sets were compared to the 591 proteins identified by BPGA resulting in a much lower than expected number of common proteins (ABySS-assembled: 171 proteins, Trinity RNA-Seq-assembled: 174). One would expect a much higher match between the protein sets but one explanation for the low matching may be the BLAST annotation step. The approach chose the first protein match to confer a protein name which is not always the correct match. Variation in protein names could also account for further differences. A future comparison should involve a more sophisticated approach to account for this issue.

The BPGA program identified proteins present in all species at a specified sequence identity (50%: 591, 75%: 82, 80%: 38, 85%: 19 and 90%: 5 proteins). A closer look at the 19 proteins found at the 85% sequence identity level showed that these proteins are involved in fundamental cellular processes including protein synthesis, cell structure and metabolism. The five essential proteins conserved at 90% sequence identity were EFT1, RPL15B, GSP2, HHF2 and TCP1. EFT1 is an elongation factor which catalyses ribosomal translocation during translation. RPL15B is the the large 60S ribosomal protein which, along with the 40S subunit, makes up a ribosome and performs protein synthesis in yeast. GSP2 is a GTP binding protein which is involved

in the maintenance of nuclear organization, RNA processing and transport. HHF2, or Histone H4, is a core histone protein required for chromatin assembly and chromosome function. Finally, TCP1 is a subunit of chaperonin-containing T-complex. Chaperonin is a heat-shock protein and the T-complex mediates protein folding and is also involved in actin cytoskeleton maintenance.

For a phylogenetic tree topology to reflect true evolutionary relationships as accurately as possible, the dataset from which it is inferred requires the correct balance of conservation and divergence. A core of conserved proteins ensures the backbone of the tree will likely reflect the true organismal relationship. More variable proteins ensure that different taxa can be properly distinguished from one another. The affects of varying numbers of core proteins on phylogenetic tree accuracy was assessed here by building phylogenetic trees of the 40 species with the BPGA core protein sets found at the five levels of sequence identity (50%: 591, 75%: 82, 80%: 38, 85%: 19 and 90%: 5 proteins). The resulting FFP phylogenetic trees were then compared to the full protein set tree (average n = 5,438). As the number of proteins was increased and decreased the phylogenetic trees changed. The lowest number of proteins, 5 proteins at 90% sequence identity, gave the least topologically similar tree to the full proteome tree. As those five proteins are highly conserved, the sequences are likely not different enough to provide enough data and therefore phylogenetic signal for the FFP algorithm to provide accurate results. However, including just 82 proteins (75% sequence identity) led to a tree topology identical to that achieved from the full proteome.

3.6 Conclusions

Studies of the core genome(s) of yeast are lacking and have the potential to highlight academically interesting and industrially- and clinically-relevant proteins. There are many challenges in correctly identifying the core genome of a species or a clade of species. When looking at a clade, how representative a strain is of that species and are there sufficient species or diversity of species to make the correct inferences are all questions that require further investigation. Choosing the most appropriate and accurate software or pipeline for the study as well as being able to take account of the

complications inherent in eukaryotic genomes such as transposons and hybridisations for example is also required.

Here, core protein sets from 41 Saccharomyces complex yeast datasets were identified both by mapping strain-specific reads to two reference genomes and by using a third-party software pipeline. These different approaches, as shown in this study, can result in differing core protein sets which highlights the need for testing a number of approaches. Given the current uncertainty in the overlap of results obtained by the BPGA and ABySS/Trinity approaches, a BPGA analysis would be recommended. A 75% sequence identity threshold would likely achieve a dataset that was both representative of the evolutionary signal within the genome(s) while also optimising speed of analysis. As expected, the proteins identified here were found to undertake functions which are essential to the survival of the species. Also, the affect that different numbers of these core proteins had on phylogenetic tree topology was clearly shown in this study and shows the need for deep consideration of which datasets to use for phylogenetic studies. The next chapter delves further into the genomic diversity contained within the Saccharomyces complex dataset.

Chapter 4

Comparative genomics of a Saccharomyces complex dataset

4.1 Summary

- The seventy-five Saccharomyces complex species and outgroup are quality filtered for this comparative genomics study.
- The genome statistics, including genome size, gene counts, BUSCO gene counts, proportion coding genome and GC content of the filtered dataset are compared.

4.2 Introduction

Comparative genomics can provide a highly detailed view of how organisms are related to one other at the genetic level and more generally how they are similar or differ in terms of their genomic organisation, composition and features. The evolution of genome content, traits and phylogenetic patterns can be uncovered in such studies. One may compare genome size, number of chromosomes or genomic ploidy, the number of genes and the GC content, for example, of a number of strains or species. This approach has been used to compare and contrast species from across the tree of life from vertebrates to yeast and has highlighted conserved elements present in all (Siepel et al. (2005)). Comparison of the fruit fly genome with the human genome revealed that about sixty percent of genes are conserved, including two-thirds of human cancer-related genes (Myers et al. (2000)). As such, the fruit fly has been used as a model organism for cancer research showing the usefulness of comparative genomics

studies (Mirzoyan et al. (2019)).

As sequencing technology is growing easier and less expensive, it is being used for a whole range of applications in agriculture, biotechnology, and zoology as a tool to tease apart the often subtle differences among animal and plant species. The first eukaryote whole genome to be sequenced was *Saccharomyces cerevisiae* in 1996 (Goffeau et al. (1996)), followed by a number of academically and industrially important species (Dujon and Louis (2017)). In the Saccharomycotina clade, a number of the Saccharomyces species including ones used in the food and alcohol industry were sequenced next. This helped distinguish new clades and species as well as the large scale of hybridisation in these species (Kellis et al. (2004), Gabaldón (2020)). Since then the number of yeast genomes being sequenced has continued to increase. One of the first population genomics studies in yeast (55 *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* strains) was published in 2009 (Liti et al. (2009)) followed, in 2018, by a study with 1,011 *Saccharomyces cerevisiae* strains (Peter et al. (2018)). These genomic analyses revealed a history of yeast domestication and the mechanisms that have contributed to these species' adaptation to anthropogenic environments.

Today, more than 1,500 yeast species have been described within two phyla (the Ascomycetes and Basidiomycetes) and classied in a variety of lineages (Kurtzman et al. (2011)), of which only a subset has been studied at the genomic level so far (Spencer and Spencer (2013)). Whole genome sequences of distantly related yeast species from across the Saccharomycotina have become available and much has been learned by comparing these species. The depth of genomic diversity across the Saccharomycotina is much deeper than was initially expected and, as shown in Figure 4.1, is on par with levels seen in animals and plants (Shen et al. (2018)). An early yeast comparative genomics studies was performed in 2000 with 13 different species which had been sequenced at low coverage, as permitted by the technology of the time, and compared to *S.cerevisiae* (Souciet et al. (2009)). This work gave the first quantitative estimates of the evolutionary spectrum covered by the Saccharomycotina based on sequence divergence between orthologous genes (Malpertuy et al. (2000)) and loss of microsynteny (Llorente et al. (2000)). In 2009, a comparison of the complete genomes of five Saccharomyces complex species (*Lachancea thermotolerans*, *Lachancea kluyveri*, *Zygosaccharomyces rouxii*, *Kluyveromyces lactis* and *Eremothe-*

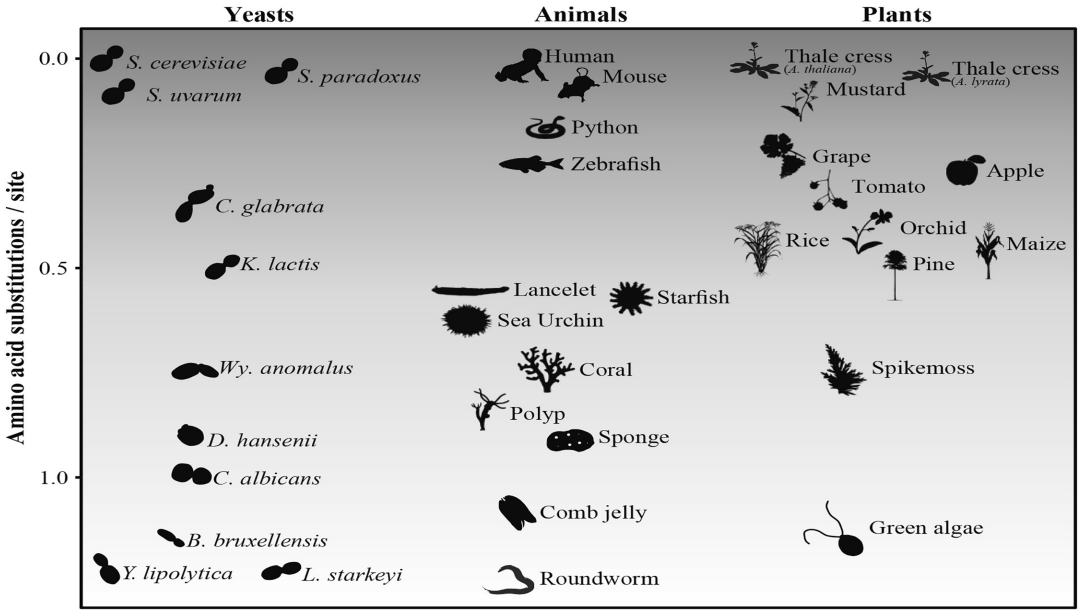


Figure 4.1: Figure taken from Shen et al. (2018) shows levels of evolutionary sequence divergence within the budding yeast subphylum are on par with levels observed in animals and plants. The phylogenetic distance (in terms of amino acid substitutions/site) between iconic species in budding yeasts (*Saccharomyces cerevisiae*), animals (*Homo sapiens*), and plants (*Arabidopsis thaliana*) and other representative species in each lineage is shown. For each lineage, the phylogenetic distance was estimated from a concatenated Maximum Likelihood tree inferred from analysis of 295 single-copy BUSCO genes.

cium gossypii) was undertaken (Souciet et al. (2009)). This study found 3,300 protein families were shared between the species and there was also a high degree of conserved synteny. By 2010, more than 20 Saccharomycotina species had been fully sequenced (Dujon (2010)). This number has now risen to more than 100 (ignoring the many hybrids) completely assembled sequences and permanent drafts with sufficiently limited numbers of scaffolds.

Yeast have a very broad range of ecologies with diverse metabolisms that have yet to be exploited by the biotechnology industry or explored by science. In recent years even more non-model yeast species have begun to be sequenced and comparative genomics studies have been undertaken (Hittinger et al. (2015), Wolfe et al. (2015), Riley et al. (2016), Dujon and Louis (2017)). These studies of diverse species have shown the similarities at the genetic level and the wide-scale prevalence of gene sharing through hybridisation, introgression and horizontal gene transfer, especially in domesticated yeast.

The aim of this study was to get an overall genomic picture of 75 Saccharomyces complex strains (and an outgroup), some of which have no publicly available genome to date. Genomic statistics assessed include: genome size, gene counts, the proportion of coding genome, whole genome GC content and BUSCO gene counts. The study highlights a number of 'genomic outliers' which may have particularly interesting evolutionary histories.

4.3 Methods

4.3.1 Dataset selection

The genomes of a set of 75 Saccharomyces complex yeast strains plus an outgroup strain were sequenced as described in Chapter 2. This 76 species genome dataset was then filtered to avoid potentially spurious conclusions being formed about a species dataset as a result of its poor genome assembly. Two datasets were created, a) one where the proportion of fragmented BUSCO genes was less than 4% and b) one where the proportion was less than 2%. CBS8763 was also removed from both datasets as the species identity was at the time unknown. This process resulted in a dataset of size 67 genomes and another of size 58 genomes. The strains removed from the 67

species set (bar CBS8763) also had N50 scores less than 20,000bp (See Figure C.1 in the Appendix). In the more conservative 58 species set, the ten removed strains included those with relatively average N50 scores amongst the full dataset.

4.3.2 Genomic profile

The genomic profile of each genome was assessed in different ways including, with the use of a custom Python script, assessing the size of the genome, number of genes, proportion of coding genome and GC content. Different k-mer count measures were assessed with Jellyfish (Marçais and Kingsford (2011)) v.2.0. BUSCO (Simão et al. (2015)) v.3 gene counts were also compared for each genome.

4.3.3 Tree annotation

The 18 taxa not included in the final 58 species dataset were deleted from the 76 species BUSCO gene tree generated for the phylogenetic comparison study in Chapter 5 using iTOL (Letunic and Bork (2006)) v5.7. The removed taxa and nodes are represented as dots on the resulting tree. The tree was then annotated according to various genomic features using iTOL.

4.4 Results

4.4.1 The dataset

The number of fragmented BUSCO genes was taken as a measure of assembly quality. Eight of the nine grey-shaded strains seen in Table 4.1 were removed from the dataset for this comparative genomics study as they have more than four percent fragmented BUSCO genes (chosen quality cut-off threshold) (See Figure C.1 in the Appendix). The ninth removed strain was CBS8763, which was believed to be misclassified. The nine yellow-shaded strains are those included in the sixty-seven species dataset which have between two and four percent fragmented BUSCO genes. The unshaded strains are the fifty-eight species which all have less than 2% fragmented BUSCO genes.

Next, the average genome statistics of both datasets were compared (See Table 4.2). The average genome size, number of genes and GC content was found to be highly similar between the two datasets whilst the coding genome measure was iden-

tical (68%). The BUSCO gene information varied to a degree which likely reflects the removal of the nine datasets with higher fragmented gene counts.

Clade	Strain ID	Species name	Fragmented BUSCO genes	N50
1	CR85	*Saccharomyces kudriavzevii*	0.7%	862,320
1	NCYC2578	*Saccharomyces bayanus*	3.2%	48,168
1	NCYC2888	*Saccharomyces mikatae*	0.8%	157,774
1	NCYC2890**	*Saccharomyces cariocanus*	0.7%	141,541
1	NCYC3662*	*Saccharomyces paradoxus*	0.8%	66,736
1	NCYC392	*Saccharomyces pastorianus*	0.9%	36,566
1	NCYC78	*Saccharomyces cerevisiae*	0.7%	280,243
2	DBVPG7206	*Kazachstania turicensis*	7.7%	3,678
2	NCYC1417**	*Kazachstania lodderae*	0.8%	31,719
2	NCYC2449*	*Kazachstania telluris*	1%	9,946
2	NCYC2450	*Candida humilis*	10.5%	12,684
2	NCYC2483**	*Kazachstania piceae*	1.2%	43,101
2	NCYC2560**	*Kazachstania sinensis*	1.2%	47,906
2	NCYC2693	*Kazachstania servazzii*	1.8%	38,946
2	NCYC2701	*Kazachstania viticola*	0.6%	46,943
2	NCYC2702	*Kazachstania kunashirensis*	1.2%	81,552
2	NCYC2703	*Kazachstania martiniae*	1%	90,007
2	NCYC2729	*Kazachstania africana*	0.4%	61,914
2	NCYC2827	*Kazachstania rosinii*	1.2%	37,466
2	NCYC2878**	*Kazachstania barnettii*	1.4%	80,952
2	NCYC2991	*Kazachstania spencerorum*	1.3%	89,048
2	NCYC3853	*Kazachstania bulderi*	7.9%	5,335

2	NCYC814**	*Kazachstania exigua*	1.1%	10,883
2	NRRLY1556	*Kazachstania unispora*	1.3%	159,570
2	NRRLY17245	*Kazachstania transvaalensis*	1.8%	133,191
3	CBS421	*Naumovozyma dairenensis*	1%	109,578
3	NCYC2898	*Naumovozyma castellii*	0.6%	103,769
4	CBS4332	*Candida castellii*	3%	13,774
4	CBS7729	*Nakaseomyces bacillisporus*	1.3%	301,794
4	NCYC388	*Candida glabrata*	0.8%	357,606
4	NCYC768	*Nakaseomyes delphensis*	1.8%	211,289
5	CBS4417	*Tetrapisispora phaffii*	0.5%	48,426
5	CBS6284	*Tetrapisispora blattae*	1.2%	286,437
5	CBS8762**	*Tetrapisispora arboricola*	0.7%	141,586
5	CBS8763	*Tetrapisispora nanseiensis*	1.1%	117,592
5	NRRLY27309	*Tetrapisispora iriomotensis*	0.8%	317,713
6	NCYC2754**	*Vanderwaltozyma yarrowii*	1.1%	7,727
6	NCYC523	*Vanderwaltozyma polyspora*	0.7%	186,992
7	NCYC1495	*Zygosaccharomyces bisporus*	0.9%	18,213
7	NCYC2403*	*Zygosaccharomyces mellis*	0.7%	73,822
7	NCYC2789**	*Zygosaccharomyces lentus*	1.1%	97,590
7	NCYC3000	*Zygosaccharomyces kombuchaensis*	1%	61,715
7	NCYC568	*Zygosaccharomyces rouxii*	0.4%	1,059,696
7	NCYC573	*Zygosaccharomyces bailii*	1.1%	44,422

8	NCYC2489	*Zygotorulaspora mrakii*	0.8%	363,292
8	NCYC2513	*Zygotorulaspora florentinus*	0.8%	196,060
9	NCYC4020	*Torulaspora delbrueckii*	0.5%	230,122
9	NCYC524	*Torulaspora pretoriensis*	0.5%	207,604
9	NCYC820*	*Torulaspora globosa*	0.6%	528,175
9	NRRLY1549	*Torulaspora microellipsoides*	0.5%	300,364
9	NRRLY17532	*Torulaspora franciscae*	0.3%	434,022
10	CBS6340	*Lachancea thermotolerans*	0.9%	1,513,537
10	NCYC2508	*Lachancea fermentati*	0.6%	136,750
10	NCYC2644	*Lachancea waltii*	1.1%	113,908
10	NCYC2875	*Lachancea cidri*	0.9%	23,555
10	NCYC543	*Lachancea kluyveri*	0.5%	28,713
11	CBS4438	*Kluyveromyces aestuarii*	1.3%	310,612
11	CBS8778*	*Kluyveromyces nonfermentans*	1.8%	193,381
11	NCYC2559	*Kluyveromyces dobzhanskii*	1%	29,646
11	NCYC2791	*Kluyveromyces marxianus*	4%	9,917
11	NCYC416	*Kluyveromyces lactis*	0.6%	918,166
11	UCD54210	*Kluyveromyces wickerhamii*	3%	36,691
12	ATCC58844	*Eremothecium sinecaudum*	1.3%	1,398,029
12	CBS106.43	*Eremothecium ashbyi*	22.7%	2,328
12	CBS109.51	*Eremothecium gossypii*	1.2%	303,511
12	DBVPG7215	*Eremothecium cymbalariae*	0.8%	1,193,613

12	NCYC1563	*Eremothecium coryli*	1.9%	17,267
13	AWRI3580	*Hanseniaspora uvarum*	2.6%	1,289,090
13	CBS2592	*Hanseniaspora occidentalis*	2.9%	33,465
13	CBS285	*Hanseniaspora lindneri*	4.7%	18,587
13	NCYC31	*Hanseniaspora osmophila*	3.4%	239,914
13	NCYC36	*Hanseniaspora vineae*	3.1%	199,177
13	NCYC4006	*Hanseniaspora valbyensis*	8.3%	2,667
13	UTAD222	*Hanseniaspora guilliermondii*	3.9%	91,417
14	NCYC3345	*Saccharomycodes ludwigii*	7.1%	3,746
Outgroup	NCYC18	*Wickerhamomyces anomalus*	2.7%	28,699
Averages:			**2.1%**	**217,552**

Table 4.1: N50 and percentage of fragmented BUSCO genes for 75 Saccharomyces complex species plus outgroup. Nine grey-shaded strains were excluded from all analyses as they had greater than 4% fragmented BUSCO genes or were potentially misclassified. 10 yellow-shaded strains had more than 2% fragmented BUSCO genes and were excluded from the final 58-species dataset. Asterisks denote strains or species for which no genome assembly is currently publicly available: *= Newly sequenced strain, **= Newly sequenced species.

Finally, the k-mer statistics (unique and distinct) found with the Jellyfish software also differed between the two datasets. These initial findings indicated that although gross genomic statistics varied between the two datasets, more fine-grained counts, such as DNA k-mers were more strongly influenced by genome choice. As a consequence, the more conservative 58 species dataset was chosen for a closer comparative genomics study.

4.4.2 BUSCO statistics

The first annotated phylogenetic tree of the final 58 species dataset shows the proportion of complete (single-copy and duplicated), fragmented and missing BUSCO genes for each individual strain (Figure 4.2). Whilst the number of fragmented gene counts were used to aid in the removal of potentially poorly assembled genomes for this dataset, the number of complete and missing genes is also often assessed to the same end. These counts can also highlight variations from the expected number of orthologous genes in a genome. Due to dataset filtering, all genomes have less than 2% fragmented genes but a small number of the genomes are missing more than 3% of the expected 1,711 BUSCO genes (Average = 1.7%). These four strains are NRRLY17245 (*Kazachstania transvaalensis*, clade 2), ATCC58844 (*Eremothecium sinecaudum*, clade 12), CBS7729 (*Nakaseomyces bacillisporus*, clade 4) and NCYC1563 (*Eremothecium coryli*, clade 12). See Table C.2 of the Appendix for all BUSCO counts.

Next, the complete BUSCO genes were investigated in more detail looking into those present in single-copy and duplicated states (See Figure 4.3). The average number of duplicated BUSCO genes for this set was 5.8% (median = 0.6) (See Table C.2 of Appendix). Seven strains in this set had more than 11% of BUSCO genes in duplicate with the remaining 51 below 3.9%. The highest numbers of duplicated genes in this dataset was 77% from CBS4417 (*Tetrapisispora phaffii*, clade 5) followed closely by NCYC814 (*Kazachstania exigua*, clade 2) at 72.6%.

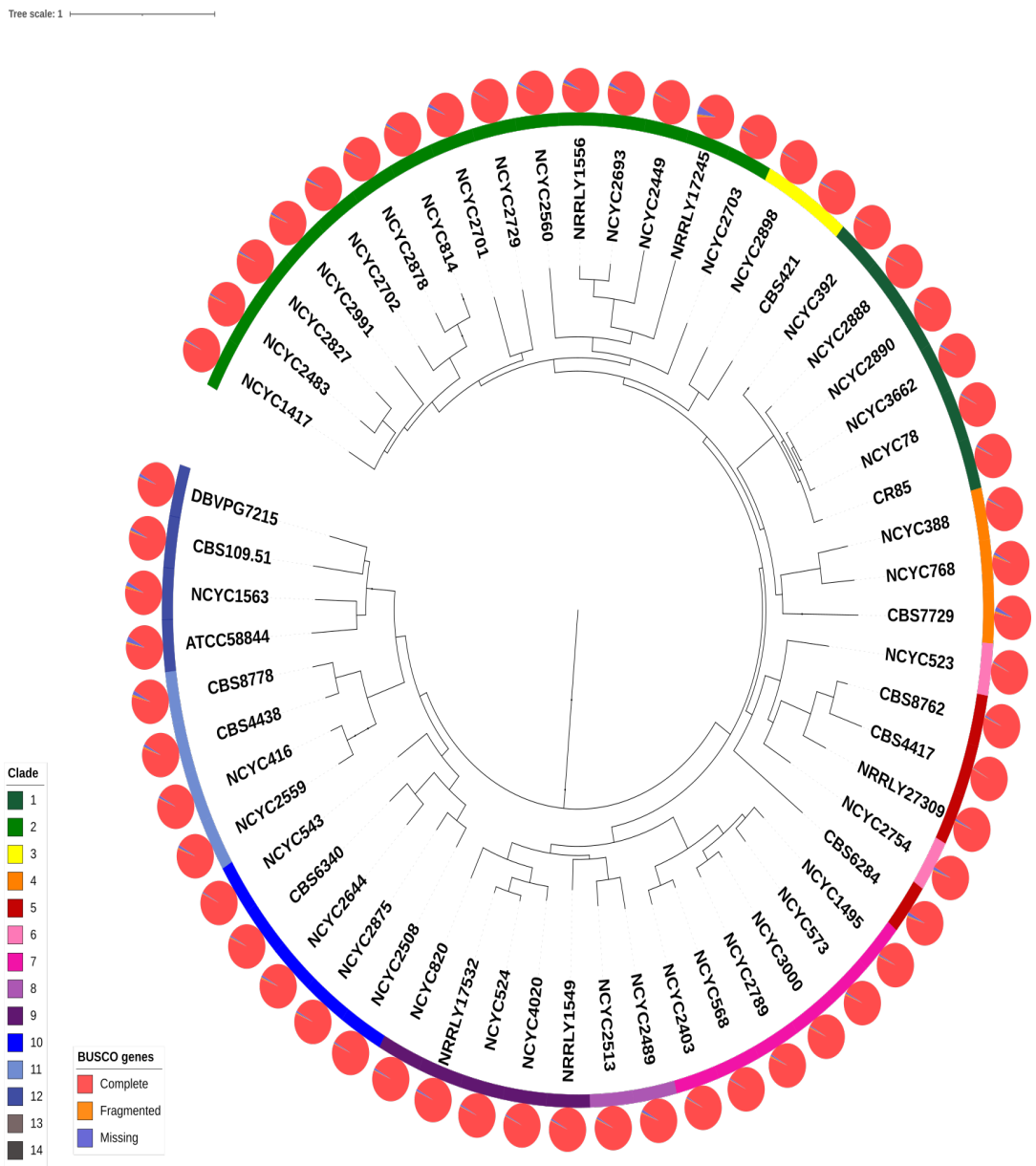


Figure 4.2: Maximum likelihood phylogenetic tree of 58 Saccharomyces complex species annotated with percentage of 1,711 genes present as complete single-copy (red), fragmented (green) and missing (blue) BUSCO genes. Clade annotation is given compared to the clade ordering seen in Kurtzman and Robnett (2003).

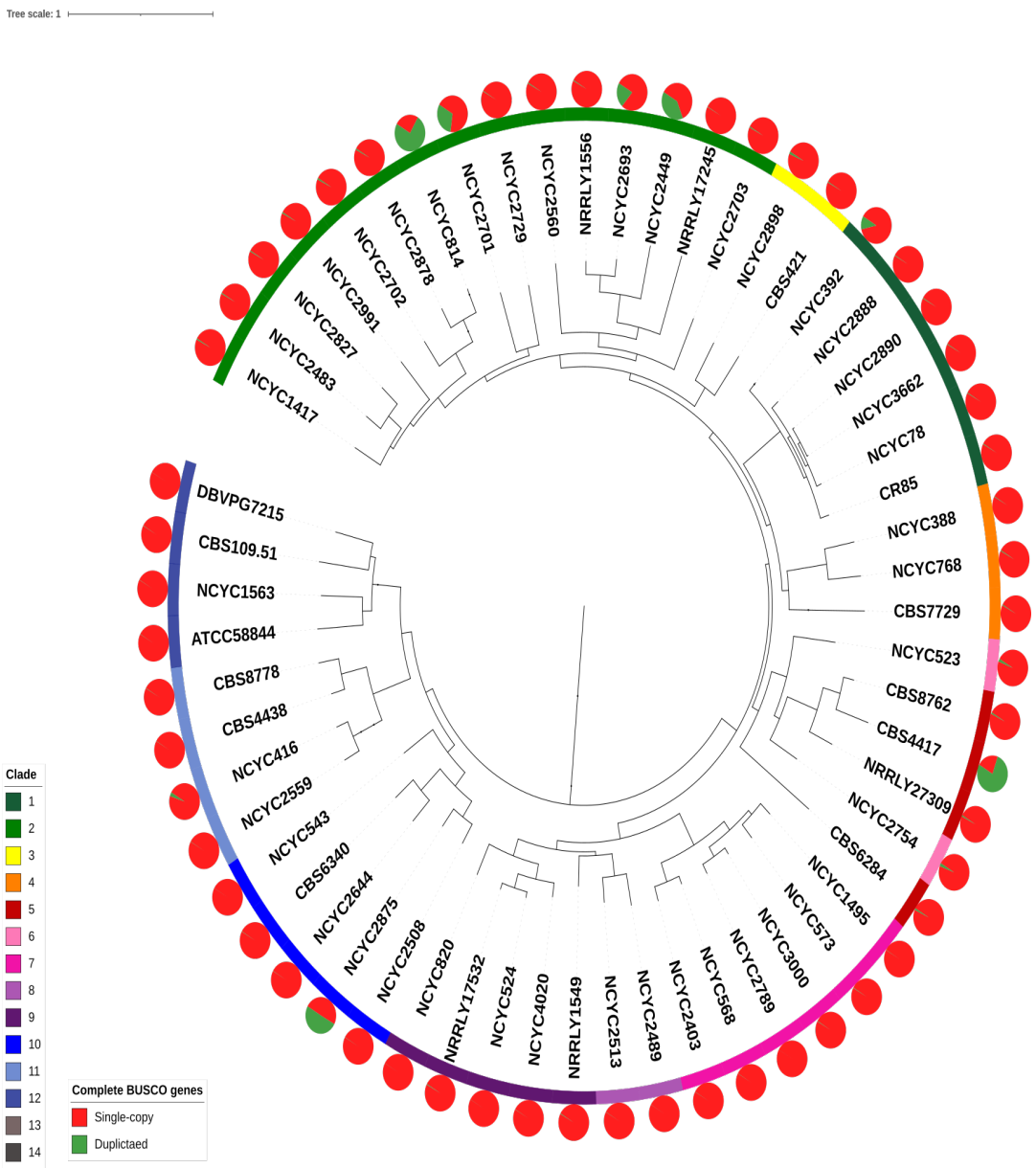


Figure 4.3: Maximum likelihood phylogenetic tree of 58 Saccharomyces complex species annotated with percentage of 1,711 genes present as complete single copy (red) and complete duplicated (green) BUSCO genes. Clade annotation is given compared to the clade ordering seen in Kurtzman and Robnett (2003).

Genomic statistics

Dataset	Genome size	No. of genes	Coding (%)	GC (%)
67 species	12941378	5875	68	39.10
58 species	12882114	5785	68	39.32

BUSCO statistics

Dataset	Complete	Single	Duplicated	Fragmented	Missing
67 species	95.10%	89.1%	6.00%	3.70%	1.24%
58 species	97.34%	91.53%	5.80%	1.71%	0.96%

Jellyfish k-mer statistics

Dataset	Unique	Distinct	Max count
67 species	8439817	10037733	3476
58 species	8602693	10159861	2663

Table 4.2: Average genome statistics of two Saccharomyces complex species datasets, of sizes 67 and 58 strains respectively.

4.4.3 Key genome statistics

The average genome size was found to be 12,848,934bps and ranged from 8,783,618bps to 29,850,499bps (See blue bars in Figure 4.4). Four strains were found to have genome sizes above 20Mbps including NCYC2449 (*Kazachstania telluris*, clade 2) at the top of this list with 29,850,499bps, NCYC814 (*Kazachstania exigua*, clade 2) with 26,437,421bps, CBS4417 (*Tetrapisispora phaffii*, clade 5) with 23,960,103bps and NCYC2875 (*Lachancea cidri*, clade 10) with 21,335,293bps. At the other end of the spectrum lies the smallest genomes of CBS109.51 (*Eremothecium gossypii*, clade 12) with 8,783,618bps and ATCC58844 (*Eremothecium sinecaudum*, clade 12) with 8,922,988bps.

The average number of genes was found to be 5,785 with numbers ranging from as high as 11,243 down to 4,131 (See green bars in Figure 4.4). Four genomes were found to be well above the average with the highest (11,243 genes) being NCYC2449 (*Kazachstania telluris*, clade 2), followed by NCYC2875 (*Lachancea cidri*, clade 10) which had 11,001, CBS4417 (*Tetrapisispora phaffii*, clade 5) with 10,479 and NCYC814 (*Kazachstania exigua*, clade 2) with 10,380 - the four longest genomes listed above. The three genomes with the lowest number of genes were clade 12

species including CBS109.51 (*Eremothecium gossypii*) with 4,131 genes, ATCC58844 (*Eremothecium sinecaudum*) with 4,299 genes and DBVPG7215 (*Eremothecium cymbalariae*) with 4,531 genes. These three low gene-count genomes included the two shortest genomes noted above.

Unsurprisingly, a strong positive correlation between genome size and the number of genes was found ($r^2 = 0.918$, see Figure C.2 in the Appendix) and is reflected in the resulting genome and gene statistics. A sizeable but weaker positive correlation was also found between genome size and the number of distinct k-mers ($r^2 = 0.7206$, see Figure C.3 in the Appendix). For both plots, the variance in the gene/ k-mer count statistic increased as genome size increased.

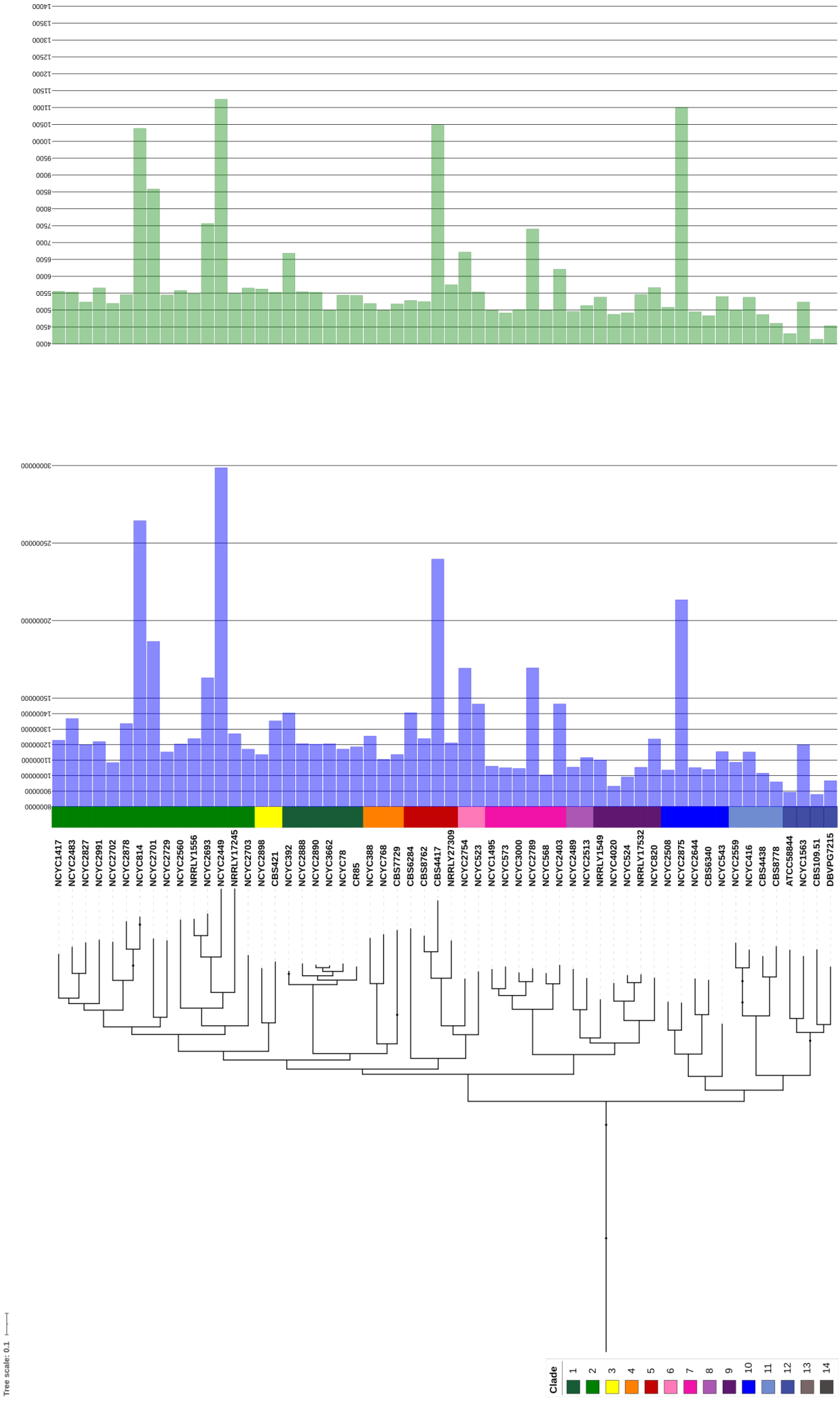


Figure 4.4: Maximum likelihood tree of 58 Saccharomyces complex species estimated from 1,711 BUSCO genes and annotated with genome size (blue bars) and number of AUGUSTUS-predicted genes (green bars). Clade annotation is given compared to the clade ordering seen in Kurtzman and Robnett (2003).

The percentage of each genome which is coding was assessed next (See Figure 4.5). The average was 68% and ranged from 77.33% down to 54.5%. NCYC4020 (*Torulaspora delbrueckii*, clade 9) had the highest proportion (77.33%) followed closely by NRRLY17532 (*Torulaspora franciscae*, clade 9) at 76.86%. NCYC2754 (*Vanderwaltozyma yarrowii*, clade 6) had the lowest proportion of coding genome (54.5%) followed by NCYC2789 (*Zygosaccharomyces lentus*, clade 7) with 55%. Nonetheless, these proportions are all high for eukaryotic genomes, highlighting the general streamlining of yeast genomes.

The average GC content in the dataset was 39.3% and ranged from 31.73% to 51.96% (See Figure 5.11). NCYC820 (*Torulaspora globosa*, clade 9) had the highest GC content with CBS109.51 (*Eremothecium gossypi*, clade 12) and NCYC2789 (*Zygosaccharomyces lentus*, clade 7) close behind at 51.81% and 51.73% respectively. The lowest value of 31.73% was seen in CBS6284 (*Tetrapisispora blattae*, clade 5). In general, GC content can be seen to vary quite significantly across the Saccharomyces complex tree and even within individual clades.

4.5 Discussion

This study has helped uncover new information about the genomes of strains and/or species for which no genomic details have yet been made publicly available. This information may be of interest from an evolutionary biology point of view as well as potentially highlighting new species for industrial application. Some of the species with particularly interesting genomic features include *Zygosaccharomyces lentus*, *Eremothecium gossypii* and *Eremothecium sinecaudum* as well as species of the Hanseniaspora clade which will be discussed in further detail below.

Choosing the strains and genome assemblies to include in this analysis proved challenging. The number of fragmented BUSCO genes was chosen as a proxy measure of assembly quality. First, a dataset excluding strains with more than four percent of the BUSCO genes fragmented was assessed. Then an even more conservative dataset which excluded strains with more than two percent was assessed. The average genome statistics of both datasets were compared and the 58 species set was selected for further analysis. This conservative approach was necessary while dealing

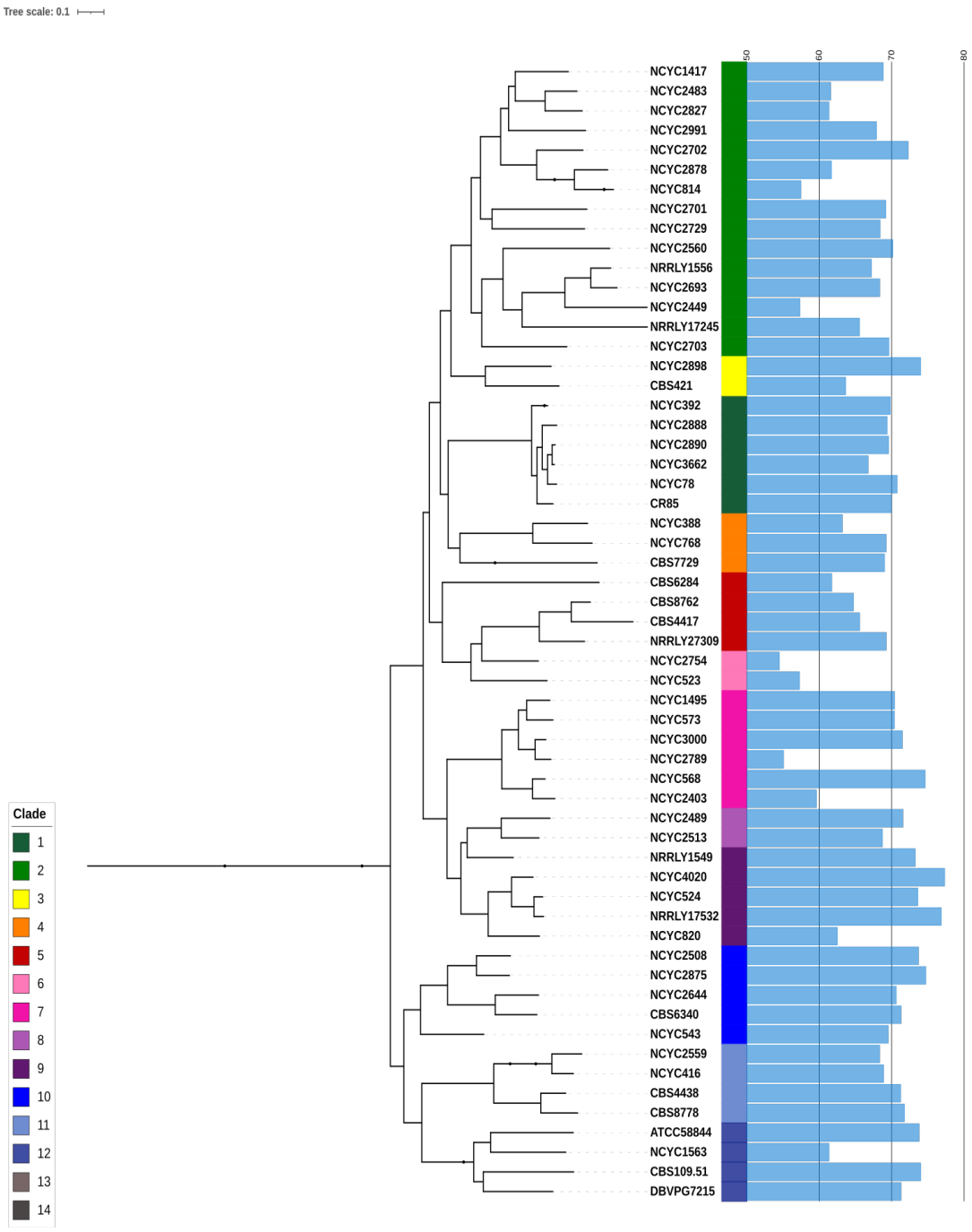


Figure 4.5: Maximum likelihood phylogenetic tree of 58 Saccharomyces complex species estimated from 1,711 BUSCO genes and annotated with percentage coding genome (blue bars). Clade annotation is given compared to the clade ordering seen in Kurtzman and Robnett (2003).

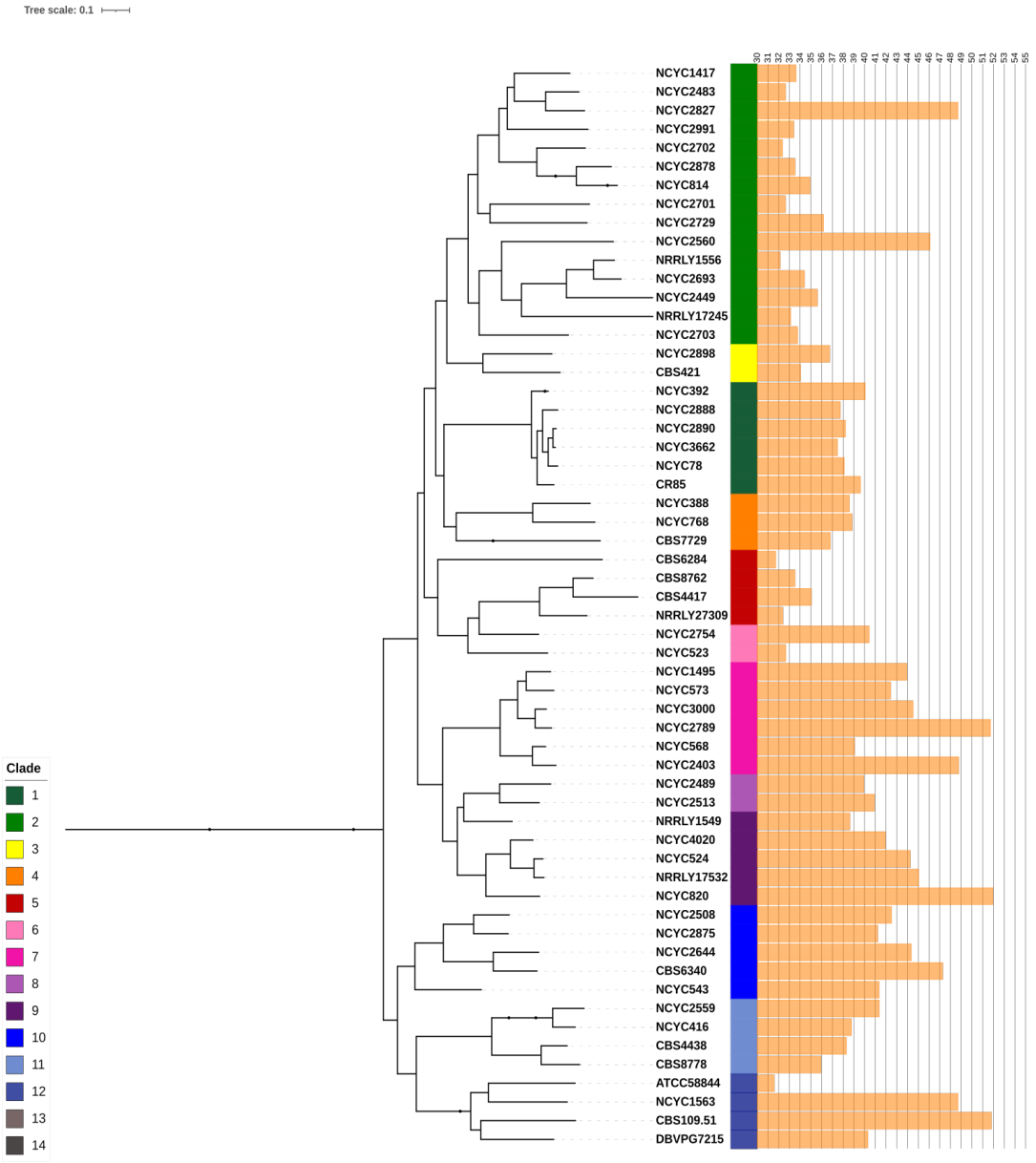


Figure 4.6: Maximum likelihood phylogenetic tree of 58 Saccharomyces complex species and outgroup *Wickerhamomyces anomalus*, estimated from 1,711 BUSCO genes and annotated with whole genome Guanine-Cytosine content. Clade annotation is given compared to the clade ordering seen in Kurtzman and Robnett (2003).

with low-cost high-throughput assemblies. When long-read sequencing can be done more cheaply and with high accuracy in all genomic regions the true genomic characteristics of all 76 species could be unearthed. Until this time, this low-cost short read approach was undertaken with estimates of key genomic parameters made with the resulting data.

Within the conservative 58 species dataset there is still reason to believe that a number of the genomes may indeed be contaminated, particularly when looking at gene duplication. Seven genomes showed a particularly large number of BUSCO gene duplications (>11%). Whilst genome duplication or hybridisation could explain these findings in theory, such events are rare and are usually followed by quite rapid gene loss (Gao and Innan (2004), Naseeb et al. (2017)). Another genome statistic which requires further inquiry is the proportion of coding genome in each dataset. This ranged from 77.33% down to 54.5% in this study, with the average being 68%. Yeast genomes are compact and this study reflects that fact with the exception of a few. It is biologically unlikely that these few yeast have such a large amount of non-coding DNA but the genomes would need to be re-sequenced to confirm this hypothesis.

Zygosaccharomyces lentus (NCYC2789), a clade 7 species, is one of the higher quality genome assemblies (N50 = 97,590) in this dataset and also has one of the largest genome lengths at 16.94Mbps. The strain also has some other interesting genome statistics including a large number of genes (7,399), a smaller than average coding genome of 55% and a high GC content of 51.7% (third highest in the set). This strain also has the third highest GC content at the third codon position in this 58 species set at 38%. This genome assembly is for a species which has no known publicly available genome to date. The strain was isolated from spoiled orange juice and deposited in the NCYC collection in 1996. This is a species which is important to the food industry because of its resistance to commonly used food preservatives and its ability to grow well at 4°C (Steels et al. (1999)). In future, it will be interesting to investigate genes potentially underlying these useful traits. Many industrial traits are thought to be driven by gene copy number expansion, particularly in sub-telomeric regions of the genome. The high observed gene count may provide useful candidates for such an analysis.

At the other end of the genome size spectrum lies a clade 11 species, *Eremothecium gossypii* (CBS109.51). This strain was found to have the smallest genome at 8.8Mbps, 74% coding and the least number of genes at 4,131. This compact genome is also reflected in the k-mer statistics with it having the second lowest number of Distinct k-mers and fourth lowest maximum count for a k-mer. This strain also has the second highest percentage of GC content at 51.8% and highest percentage of which is at the third codon position (39%). A representative of this species was also sequenced in a recent study Shen et al. (2018), in that case the strain ATCC10855. The genome there was found to be 9Mbps in length, with 4,327 genes and 51.79% GC content, statistics highly similar to those found in this study. Using a custom Kraken database, CBS109.51 was found to be 97.14% identical to ATCC10855 at the read level. *Eremothecium gossypii*, also known as *Ashbya gossypii*, is a filamentous fungus which was first described in 1929 as a cotton pathogen transmitted by sucking insects. In addition to cotton, it infects other agricultural crops such as citrus fruits. The species is also used as a model to study filamentous growth due to its small haploid genome and is used in industry for the production of riboflavin (Dietrich et al. (2004)).

A group of species which collectively appears to have lost a large proportion of its genes lie within the Hanseniaspora clade (seven species in this study). Whilst these strains were not included in the 58 species dataset, five out of seven were included in the 68 species dataset and are nonetheless worthy of note. The proportions of missing BUSCO genes within these five Hanseniaspora strains were found to range from 11.9% to 46.2%. The proportions of fragmented genes ranged from 2.6% to 3.9% and duplicated genes ranged from 0.1% to 1.7%. The proportions of BUSCO genes along with the values of the other genome statistics were similar to those found for the same species sequenced in a recent study (Shen et al. (2018)).

Recently, multiple genome-scale phylogenies of species in the budding yeast subphylum Saccharomycotina showed that certain species in the yeast genus Hanseniaspora are characterized by very long branches [Shen et al. (2016), Riley et al. (2016), Shen et al. (2018)], which are reminiscent of the very long branches of fungal hypermutator strains (Rhodes et al. (2017)). If indeed these strains are high mutators then perhaps this could explain the large loss of genes. Hanseniaspora species are

found in high abundance on mature fruits and in fermented beverages (Albertin et al. (2016)), especially on grapes and in wine must (Jordão et al. (2015), Montero et al. (2004)). It has been found that even with the use of *S. cerevisiae* starter cultures in wine production, Hanseniaspora species, particularly *Hanseniaspora uvarum*, can achieve very high cell densities, in certain cases comprising greater than 80% of the total yeast population, during early stages of fermentation (Hendler et al. (2017)), suggesting exceptional growth capabilities in this environment.

A genomic study of Hanseniaspora genus species by Steenwyk *et al.*, (Steenwyk et al. (2019)) found that compared to *S. cerevisiae*, 748 genes were lost from two-thirds of Hanseniaspora genomes examined, with a lineage dubbed the fast evolving lineage (FEL) of yeasts (including *H. uvarum*) having lost an additional 661 genes. In contrast, a slow evolving lineage (SEL) were found to have lost only an additional 23 genes. Both lineages were found to have lost major cell cycle regulators and DNA damage checkpoint genes. Also, the average GC contents for FEL yeasts (33.10%), SEL yeasts (37.28%), and all other Saccharomycotina yeasts (40.77 $\pm$ 5.58%) were found to be significantly different. Given the findings in these prior studies, the exceptional numbers of missing BUSCO genes within the strains sequenced in this study should be investigated further.

This study has presented ten genomes in the conservative 58 species dataset that are of relatively good assembly quality (N50$>$ 30,000) and for which no genome sequence is known to be publicly available to date (See strain list in Table C.3 of the Appendix). Furthermore, six of these assemblies are the first sequenced representative of a species. The assemblies of one new species and two new strains are of a particularly high quality (N50$>$ 100,000) - the species *Saccharomyces cariocanus* (NCYC2890) and new strains of *Torulaspora globosa* (CBS820) and *Kluyveromyces nonfermentans* (CBS8778). The new species genome, *Saccharomyces cariocanus*, is a *Saccharomyces sensu stricto* (clade 1) species with similar genome statistics to other species in the complex. The genome is of length 12Mbps, the coding genome proportion is 69.6%, the estimated GC content is 38.22% and 5,526 genes were predicted for this dataset. It also has 98% of the expected BUSCO genes present, of which 97.5% are in single copy. NCYC2890 is the type strain for the species and was first described in 2000. It was isolated originally from a fruit-fly in Brazil and is very closely related to North

American *Saccharomyces paradoxus* species (Naumov et al. (2000), Liti et al. (2006)).

4.6 Conclusions

Much information learned from comparative studies of yeast is applicable across eukaryotic organisms. More analyses such as these are increasing with the burgeoning of genome sequencing and will likely answer even more questions about evolution as well as have implications for industry. This analysis has given some first insights into the wide genomic diversity within this set of species as well as highlighted a number of particularly unusual yeast genomes. Also, new information about the genomes of 10 strains and/or species, for which no genomic details have yet been made publicly available, has been uncovered here. This study has also clearly shown the need for high quality sequencing data for these types of studies in order to tell apart real biological phenomena, such as whole genome duplication events, and contaminated sequences. Many questions remain about the genomes of the Saccharomyces complex species and this will require high quality sequencing data and large-scale comparisons, which is very evident from this study. In the next section, the full 76 Saccharomyces complex species dataset is used to compare different phylogenetic tree building approaches.

Chapter 5

Comparison of phylogenetic methods for a Saccharomyces complex dataset

5.1 Summary

- Five phylogenetic trees are built from a dataset of seventy-five Saccharomyces complex species and an outgroup using different approaches.
- The resulting trees are compared to each other and the tree in Figure 5.3 using two computational measures (the Robinson-Foulds and Kendall-Colijn metrics).
- The FFP method is investigated further for potential biases in the approach.

5.2 Introduction

The reconstruction of an evolutionary tree helps us to think more clearly about the differences between species and allows us to analyse them in a statistical sense. Phylogenetic trees can be built from different types of homologous biological data such as conserved genes (e.g. ribosomal DNA), Single Nucleotide Polymorphisms (SNPs) and more recently whole genomes. As discussed in the Introduction chapter, there are two approaches for phylogenetic inference, character-based approaches and distance-based approaches. Character-based approaches typically involve aligning the sequences of one or more homologous genes and use methods such as Maximum Parsimony (Sober

(1983)), Maximum Likelihood (Felsenstein (1981)) or Bayesian inference (Huelsenbeck et al. (2001)) for tree building. Distance-based approaches convert a character matrix into a distance matrix that represents the evolutionary distances between all pairs of species. The phylogenetic tree is then inferred from this distance matrix using an algorithm such as Neighbor-Joining (Saitou and Nei (1987)).

Whilst character-based approaches require a sequence-alignment step, distance-based methods could use gene, SNP or k-mer frequencies for example. A comparison of these different types of phylogenetic inference methods was carried out in 2018 by Lees *et al.*, (Lees et al. (2018)). Four approaches, alignment-based, partial-alignment, distance-based with alignment and distance-based without alignment, were compared. Phylogenetic trees were built from simulated bacterial genome datasets and, as shown in Figure 5.1, found RAxML (Stamatakis (2014)) and IQ-Tree (Nguyen et al. (2015)) along with a close reference alignment to be the most accurate and efficient approaches.

5.3 Alignment-free phylogenetic approaches

Next Generation Sequencing (NGS) has resulted in very large datasets (e.g. whole genome sequences rather than sequences of just one or a few genes) for thousands of species and strains across the tree of life. This data quantity can make the sequence alignment step required by character-based phylogenetic approaches extremely difficult, both in computational and human time. As a result, new alignment-free approaches have started to be developed which compare word frequencies in sequences. These methods are much more efficient than alignment-based methods (as indicated in Figure 5.1) as aligning two sequences takes time proportional to their total sequence length whilst word frequencies, which most alignment-free methods use, can be calculated in linear time (Vinga and Almeida (2003)). Many different alignment free sequence comparison tools have been developed for phylogenetics as well as for mapping, assembly and metagenomics (Zielezinski et al. (2017), Zielezinski et al. (2019)).

A recent comparison of alignment-free (AF) sequence comparison tools was carried

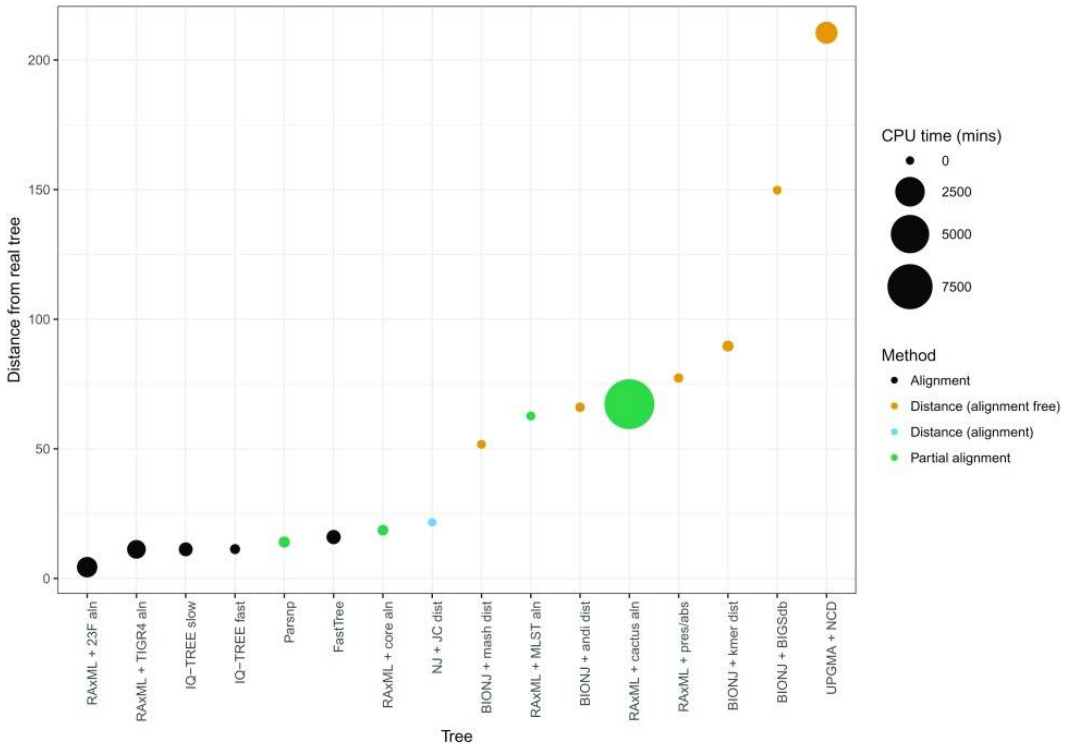


Figure 5.1: Figure from Lees et al. (2018) (Figure 2). Plot shows ordered accuracies, in terms of distance from the true tree, and the CPU time required for tree estimation for each of 16 methodological approaches.

out in 2019, showing the large number of approaches being developed as an alternative to traditional methods for handling large datasets (Zielezinski et al. (2019)). Some variation in accuracy of the different software was found. To test the genome-based phylogenetic analysis approaches, the authors used both assembled and unassembled sequence (read) datasets as well as datasets of varying size and species. Twenty two AF tools (70 tool variants in total) had their performance tested in phylogenetic inference using complete mitochondrial genomes from 25 fish species, with the greatest accuracy being achieved by nine AF tools which generated almost identical tree topologies (normalised Robinson-Foulds (nRF) = 0.05) and 39 other tool variants held joint second position with nRF = 0.09. These results indicate that most AF methods tested infer trees in general agreement with the reference tree of mitochondrial genomes.

The authors also tested the performance of 16 tools (61 tool variants) on the larger bacterial genomes of *Escherichia coli/Shigella* and nuclear genomes of plant species (Zielezinski et al. (2019)). This resulted in a lower performance than was seen for the mitochondrial genome trees and differences in top performing tools for the bacterial versus plant datasets. Some of the software tested were developed for closely related organisms and so were found to be the best performing tools for the bacterial dataset, yet performed poorly for the plant dataset. One such example is the Phylonium (Fabian and Bernard (2019)) software which achieved an nRF value of 0.04 for the bacterial dataset but an nRF value of only 0.64 for the plant genomes. The best-performing tools for the plant data were found to be co-phylog (Yi and Jin (2013)), mash (Ondov et al. (2016)) and Multi-SpaM (Dencker et al. (2018)), all of which had an nRF value of 0.09. Overall, across the two datasets, the best performing tools were co-phylog, mash, Skmer (Sarmashghi et al. (2019)), FSWM (Leimeister et al. (2017)) and FFP (Sims et al. (2009a)). In addition, previous testing of co-phylog on NCYC sequence datasets (Dr Jo Dicks, pers. comm.) found that while the software ran well on a small strain dataset, there was an exponential increase in computation time, with a 94-strain dataset taking ∼6 weeks to run.

5.3.1 Feature Frequency Profiles

One highly popular method of alignment-free phylogenetic analysis is Feature Frequency Profiles (FFP). The FFP software is a collection of utilities for implementing the FFP methods of phylogenetic comparison, it is suitable for viral to mammalian-scale genomes and has been used successfully with viral, bacterial, fungal and mammalian sequences in the past (Wu et al. (2009), Sims and Kim (2011), Choi and Kim (2017),Choi and Kim (2020)). FFP calculates the frequencies of features (e.g. DNA 'words' such as 'AATT') of a suitable length in one genome and compares them to the analogous frequencies in other genomes. The distance between word frequency distributions can be calculated using the Jensen-Shannon distance (Lin (1991)) and this can then be used to build a phylogenetic tree by third-party software such as PHYLIP's neighbor program (Felsenstein (1989)) or BIONJ (Gascuel (1997)). Feature filtering can be done if one wanted to remove features of low or high frequency from the comparison.

The first step in building a phylogenetic tree with FFP is creating a feature frequency profile of each genome. The ideal k-mer/feature length will differ from species to species and can be discovered by running the FFP program with a range of k-mer lengths, plotting the Robinson-Foulds distances of k-mer vs k-mer +1 trees and then picking one of the k-mers within a range of topological convergence. Once a k-mer length has been decided upon, a feature frequency profile can be created (Figure 5.2, Step 1). This can be done for a given DNA sequence using an RY alphabet where R is Purines (A or G) and Y is Pyrimidines (C ot T) or using the four letter ACGT alphabet. A 20 letter alphabet can also be used if working with amino acid sequences.

To count the frequencies of each feature in the genome, a sliding window of length l is run through the sequence from position 1 to n - l + 1, with a step of size 1. Large genomes, which consist of multiple chromosomes, are represented by a collection of assembled chromosomes and others are just a collection of unassembled contigs. When counting, l-mers continue over the whole genome, but the sliding window is not allowed to span over sequencing gaps. For nucleotide sequences, the counts are tabulated in the vector C_l for all possible features of length l,

$$C_l = <C_{l,1} \ldots C_{l,K}>$$

where K, the number of all possible features, is 4^l and 4 is the alphabet size. A similar process is adopted for amino acid sequences, the only difference being the alphabet size (20) and the potential number of features (20^k). The next step of the algorithm removes the feature labels and aligns features by column (Figure 5.2, Step 2). Then the relative frequency of each feature is calculated (Figure 5.2, Step 3). The raw frequency counts are normalized to form a probability distribution vector or FFP,

$$F_l = C_l / \sum_i C_{l,i}$$

giving the relative abundance of each l-mer. This normalization procedure removes small genome length differences as a factor in the comparison. Next, a distance matrix can be created with Jensen–Shannon Divergence (Figure 5.2, Step 4) which can then be used to create a phylogenetic tree with PHYLIP's neighbor program or BIONJ (Figure 5.2, Step 5). The Jensen-Shannon Divergence (JSD) is a method of comparing two or more probability distributions, producing finite and symmetric values. For a series of probability distributions $P_1, P_2, ..., P_n$, JSD is defined as:

$$\mathrm{JSD}_{\pi_1,\ldots,\pi_n}(P_1, P_2, \ldots, P_n) = H\left(\sum_{i=1}^{n} \pi_i P_i\right) - \sum_{i=1}^{n} \pi_i H(P_i)$$

where $\pi_1, \ldots, \pi_n$ are weights selected for the probability distributions $P_1, P_2, \ldots, P_n$ and $H(P)$ is the Shannon entropy for distribution P. For the two-distribution case used here:

$$P_1 = P, P_2 = Q, \pi_1 = \pi_2 = \frac{1}{2}$$

The Shannon entropy is defined as:

$$H(X) = \sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$

The efficiency of the software varies depending on which alphabet (2, 4 or 20-letter) is being used. A tree built from the two letter alphabet is the most efficient in computational terms but has less precision and is missing information compared to a four letter or twenty letter alphabet. The four letter DNA alphabet uses the most information from within the input sequences but early analyses indicate has less precision than use of the 20 letter amino acid alphabet which may make the amino acid alphabet preferable where such datasets are available. Also, previous analyses of NCYC genomes have indicated a key potential problem with the FFP 4-letter method: a phenomenon termed "GC-attraction" (similar to the well-known

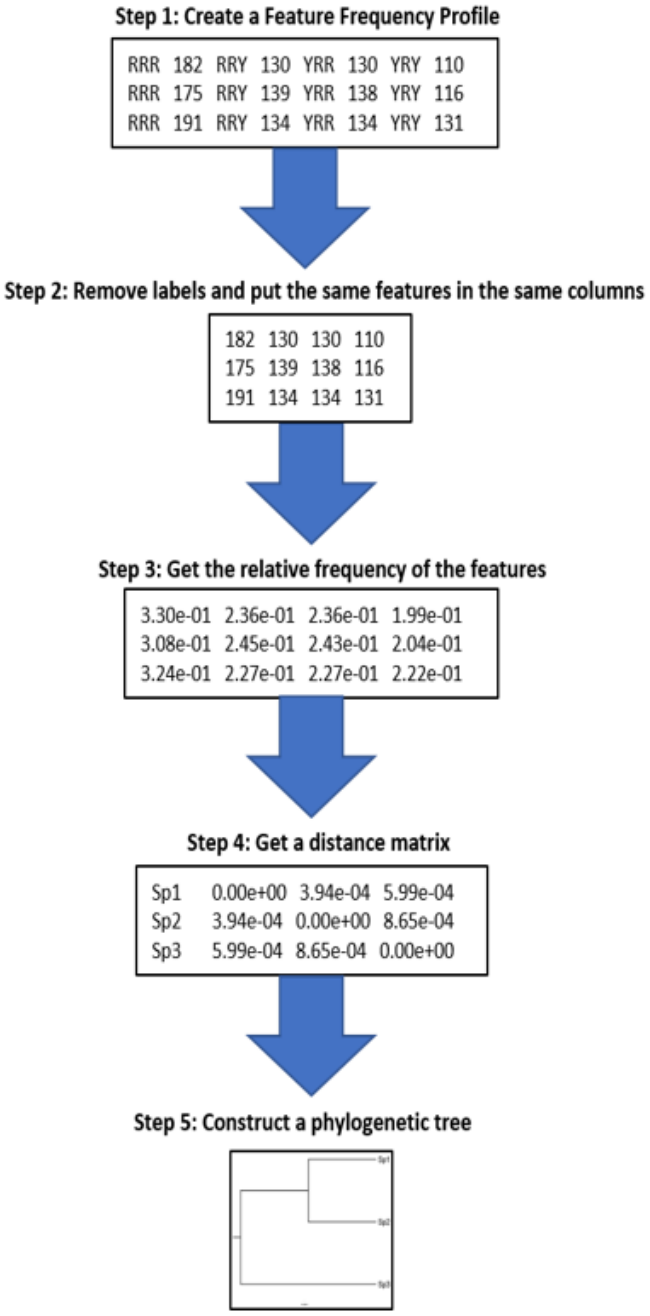


Figure 5.2: The five steps of Feature Frequency Profile (FFP) phylogenetic tree estimation: Step 1) Count the occurrences of distinct 'words' in the input sequences; Step 2) Remove word labels; Step 3) Normalise the word frequencies; Step 4) Calculate a pairwise distance matrix from the word frequency distributions (e.g. using the Jensen-Shannon Divergence); Step 5) Estimate a phylogenetic tree from the pairwise distance matrix.

long-branch attraction (Carmean and Crespi (1995), Bergsten (2005))) in situations where GC content differs markedly between taxa. The present FFP algorithms are essentially written for any type of string data, whether words in a book or sequences in a genome, as long as the alphabets are of a certain type. Taking biology more explicitly into account in these algorithms may result in a more accurate tree.

As mentioned previously, the FFP approach has been used for comparing mammalian, prokaryote and fungal genomes in the past (Sims et al. (2009b), Jun et al. (2010), Sims and Kim (2011), Wang and Ash (2015), Choi and Kim (2017)). The resulting topologies were found to be highly similar to those of established trees. The mammalian study used the RY-alphabet approach and also included a comparison of the evolutionary signal from genic and non-genic sequence (Sims et al. (2009b)). The authors found that the non-genic portion of the mammalian genomes contained evolutionary information that is similar to their genic counterparts. A phylogenetic study of 884 prokaryotes which used the 20-letter amino acid alphabet was undertaken in 2010 (Jun et al. (2010)). In 2011, the RY-alphabet approach was used again to build a tree from 38 *Escherichia coli/Shigella* group species. More recently, in 2017 a fungal tree of life analysis was published which used the FFP 20-letter amino acid alphabet approach to build a tree which consisted of the whole proteomes of 244 unique fungal species and 71 protozoan species (Choi and Kim (2017)). The study compared findings from a gene-tree approach to that of the FFP approach. Similar, but not identical, trees resulted. One of the most recent studies using the FFP 20-letter amino-acid alphabet approach was in 2020 which used the whole-proteomes of 4,023 taxa to build a tree of life (Choi and Kim (2020)). The findings of this study, which challenged the previous consensus view of the tree of life, was followed by a phylogenetic tree comparison study by Li *et al.*, (Li et al. (2020)). The FFP approach was compared to concatenation and coalescence approaches using a real dataset and a simulated dataset. The simulated dataset consisted of one-hundred genes which were simulated under a 50-taxon balanced tree. The comparison found that the FFP approach did not perform as well as the standard approaches. Consequently, there is no current consensus view of the efficacy of the FFP approach.

As there is a wide variety of approaches to phylogenetic tree building, choosing the most appropriate method for one's question and data can be challenging. As

discussed in the Introduction chapter, there are also different tree comparison metrics which look at similarities in topology alone or both topology and branch length. The Mantel test (Mantel (1967)) and the Euclidean metric take both topology and branch lengths into account whilst the Robinson-Foulds (Robinson and Foulds (1981)) and Kendall-Colijn (Jombart et al. (2015)) metrics have both a weighted (including branch lengths) and an unweighted (topology only) option. A 2015 study which evaluated tree comparison metrics concluded that the best metric depends on whether branch-length information is of interest (Kuhner and Yamato (2015)).

In conclusion, phylogenetic tree building is key to uncovering the evolutionary history of a taxonomic group of organisms. There are a number of different approaches and software available to date. Choosing the most appropriate approach will depend on the data available as well as computational resources. In the age of big-data, large numbers of whole genome sequences are becoming widely available. With this, building trees by traditional approaches that involve an alignment step becomes challenging if not infeasible. This problem has given rise to a number of alignment-free approaches. The FFP method, an efficient k-mer based approach, is both computationally straightforward and widely-used. While FFP has been used successfully in many real data analyses, there have been suggestions of poor performance or systematic biases for some datasets.

The aim of this chapter is to compare different phylogenetic approaches for a key yeast dataset. FFP, SNP and BUSCO phylogenetic trees were constructed from a 75 Saccharomyces complex genome (plus outgroup) dataset. The results of these analyses were compared using the Robinson-Foulds and Kendall-Colijn metrics. The FFP approach was investigated further to see whether a hypothesised GC bias for the 4-letter alphabet was present in this dataset and whether genome length affected the accuracy of results.

5.4 Methods

5.4.1 Dataset

Three datasets for each strain of 75 Saccharomyces complex species plus outgroup strains were used in this analysis: a) sequence reads, b) genome assemblies, c) AUGUSTUS-predicted coding gene sets. All methods for dataset generation were described in previous chapters with the exception of the pseudo-read generation of seven datasets which is described in Section 5.4.4.

5.4.2 Kurtzman and Robnett tree estimation

Sequence files from a classic study of Saccharomyces complex yeast species (Kurtzman and Robnett (2003)) were input to PAUP (Swofford (2001)) v4.0168 to generate a maximum parsimony tree with 100 bootstrap replicates, a process mirroring that conducted by the authors. The tree was generated from all 76 species (including outgroup) which consisted of 4,962 characters, of which 929 were parsimony informative.

5.4.3 FFP tree estimation

First, the optimal feature (i.e. word or k-mer) length of a pilot dataset of 11 yeast strain genome assemblies was established using the *ffpvprof* and *ffpreprof* options in the FFP v3.19 program (not available in 2v3.0). The *ffpvprof* option counted the usage of words of length 3 to 30 that occurred a minimum of three times. This was used to determine the lower limit for word length. The *ffpreprof* option calculated the relative entropy between observed and expected frequencies of words for a range 3 to 30 using an l-2 Markov Model. This was used to determine the upper limit for word lengths. The optimal feature length range was found to be between 11 and 26. This was followed by the generation of six two-letter DNA (RY) FFP trees ranging from a k-mer length of 10 to 15. The trees were visually assessed to see when the tree topologies converged. An optimal k-mer length of 14 was decided for further analysis (See Appendix Figure D.1).

FFP trees of the full 76 yeast species dataset were then constructed using the Jensen-Shannon Divergence matrix between feature frequency profiles of word length 14 for all 3 alphabets with FFP (Sims et al. (2009a)) 2v3.0. The FFP amino acid tree

was produced with the 20-letter alphabet (no classing selected) from AUGUSTUS-predicted and translated coding gene datasets. The FFP four- and two-letter DNA alphabet trees were constructed with default settings from genome assembly datasets. The output matrices were converted to phylogenetic trees with the neighbor program of PHYLIP (Felsenstein (1989)) v3.695 (default settings). All trees were viewed and annotated with iTOL (Letunic and Bork (2006)) v5.7. A bootstrap option was available in FFP v3.19 but failed to run on a yeast dataset of this size.

5.4.4 MLST SNP tree estimation

Single Nucleotide Polymorphism (SNP) trees were generated for the same 76 species dataset. Of the 14 publicly available genomes used in this dataset seven had sequence read datasets in addition to genome assemblies. For the remaining seven datasets, for whom only assemblies were available, pseudo read datasets were generated with pIRS (Hu et al. (2012)) v2 with parameters -x 50 (sequence coverage), –no-subst-errors (no substitution errors) and –no-indels (no indels). Forward and reverse paired trimmed reads for each strain were mapped to the *Saccharomyces cerevisiae* S288c reference genome (Accession number: GCF000146045.2) with Stampy (Lunter and Goodson (2011)) v1.0.31. Samtools (Li et al. (2009)) v1.9 was used to order and index the Stampy produced SAM files. Picard (Broad-Institute (2018)) v2.9.4 with parameter AddOrReplaceReadGroups was used to assign all the reads in a file to a single new read-group. This was followed by running FreeBayes (Garrison and Marth (2012)) v1.2.0-4 for variant prediction against the *S. cerevisiae* reference, output in vcf format.

SNP calling was completed using a custom R (R Development Core Team (2008)) script. High quality binary SNPs were identified across the whole genome first, followed by the filtering of SNPs to just those present in Multi-Locus Sequence Typing (MLST) genomic regions, genes and other genomic elements used to unambiguously determine the taxonomic identity of an organism. In yeast, the MLST regions include rDNA genes (18S, 26S, Internal Transcribed Spacer, 5.8S), a protein-encoding gene (EF-1α) and Mitochondrially-encoded genes (15S rDNA, COX2). A bootstrapped SNP tree was generated with the PHYLIP (v3.695) SeqBoot program (1,000 bootstrap replicates), Gendist program (all alleles option), Neighbor program (neighbour joining option, multiple dataset option) and consense (default setting). The tree was

viewed and annotated with iTOL v5.7.

5.4.5 BUSCO tree estimation

A phylogenetic tree of 1,711 BUSCO genes across the 76 species dataset was constructed as described in the previous chapter using the Saccharomycetales lineage option. This approach largely followed that used in Shen et al. (2018). Initially all potential paralogs or additional spurious genes which were the result of a bug in the BUSCO software (also acknowledged by another user) were removed using a custom Python script. Each gene was then aligned for all species using the MAFFT multiple sequence alignment software (Katoh et al. (2002)) v7.529 (E-INS option, which allows for large unalignable regions), followed by trimming of alignments with the trimAl (Capella-Gutiérrez et al. (2009)) v3 software (gappy out option). Next, all genes of length less than 167 amino acid sites and genes which were present in less than 50% of strains were removed, again with a custom Python script. Species' genes were also removed if the length of the gene was less than half the size of the average for that gene across the dataset. The resulting dataset consisted of 1,541 genes. IQ-Tree (Nguyen et al. (2015)) v1.7 was used to construct a maximum-likelihood phylogenetic tree under a single-partition LG+G4 model (options used: subtree prune regraft 4, mlacc 2slownni) which was found optimal for a similar dataset in a recent study (Shen et al. (2018)). Next the tree was bootstraped 1,000 times to assess the statistical support for the tree and concatenated with IQ-Tree (-nt AUTO, -m LG+G4). iTOL v5.7 was again used for tree annotation.

5.4.6 Tree comparison metrics

The Robinson-Foulds distance metric, both weighted and unweighted, were obtained for each comparison to the Kurtzman and Robnett tree for all FFP generated and BUSCO Newick trees using the dendropy library (Sukumaran and Holder (2010)) in Python v.2.7.12. The Kendall–Colijn metric, both weighted and unweighted, was obtained with the treescape package (v.1.10.18, Jombart et al. (2015)) in R (v.3.3.2, R Development Core Team (2008)). Topology measures alone of both metrics were obtained for the MLST SNP tree as the consensus tree did not have branch lengths but rather consensus/bootstrap values in the Newick tree.

5.4.7 Data simulation

A short simulation project was undertaken to assess any correlation between FFP tree accuracy and sequence length. The *Saccharomyces cerevisiae* S288c reference genome (RefSeq Accession number: GCF000146045.2) was split into 17 individual chromosomes. Each chromosome was used as the ancestral sequence to simulate 41 sequences over a given phylogenetic tree with the Seq-Gen software (Rambaut and Grass (1997)) v.1.3.4. The General-Time-Reversible model of evolution (Waddell and Steel (1997)) and a random seed number of 13 were selected for all runs. A previously constructed MLST SNP tree consisting of 40 Saccharomyces complex strains and an outgroup was used as the true tree over which to generate the sequences. Seventeen FFP trees were built using each alphabet approach with the 17 chromosome simulated datasets. The FFP amino acid tree approach required an AUGUSTUS prediction and translation step prior to tree building. The Robinson-Foulds distance between each chromosome tree and the true tree was assessed. A k-mer length of 14 was used for all FFP trees.

5.5 Results

Phylogenetic trees estimated from 75 Saccharomyces complex species (plus outgroup) using five different computational approaches were compared to the 2003 Kurtzman and Robnett tree topology. Figure 5.3 shows the original tree from the 2003 paper. While the original Newick tree was unavailable for this study, the PAUP files from which the tree was originally generated were provided by a colleague (Dr. K. T. Huber, pers. comm.). The tree resulting from the process described in the Methods section can be seen in Figure 5.4. The newly estimated tree consisted of the same number of characters and parsimony informative characters as the published tree (4,962 and 929 respectively). While the clading and clade ordering are highly similar to those of the published tree, the topology, tree length and other measures varied slightly from the original. In the original, the tree length was 5135, the consistency index (CI) was 0.329, the retention index (RI) was 0.63 and rescaled consistency index (RC) was 0.208. For the newly generated tree, tree length was 5245, CI was 0.322, RI was 0.62 and RC was 0.2. One explanation for these differences could be the use of a different version of the PAUP software to generate the tree. The original tree used version 4.063 whilst version 4.0168 was used for this study.

Metric			RF-Unweighted	RF-Weighted	KC-Unweighted	KC-Weighted
FFP-2	vs	KR	125	12957	541	10095.5
FFP-4	vs	KR	109	12931	344	10094.9
MLST	vs	KR	77	-	258	-
FFP-20	vs	KR	63	**12928**	93	10094.6
BUSCO	vs	KR	**49**	12939	**78**	**10084**

Table 5.1: Comparison between five phylogenetic trees and the newly estimated Kutzman and Robnett tree for 75 Saccharomyces complex species plus outgroup. Trees: Kurtzman and Robnett tree (KR); FFP 2-letter RY alphabet (FFP-2), FFP four-letter DNA alphabet (FFP-4), FFP 20-letter amino acid alphabet (FFP-20), BUSCO core gene (BUSCO) and MLST SNP (MLST). Tree comparison metrics: Robinson-Foulds (RF - Unweighted and Weighted) and Kendall-Colijn (KC - Unweighted and Weighted).

The trees resulting from FFP (2-, 4- and 20-letter alphabets), BUSCO alignment and MLST SNP tree approaches were compared to the Kurtzman and Robnett tree in Figure 5.3. The trees can be seen in Figures 5.5 to 5.9 and the measures of similarity to the KR topology are shown in Table 5.1. The least similar to the original tree on all measures was the tree estimated by the FFP RY-alphabet approach (RF-unweighted = 125, KC-unweighted = 541) whilst the BUSCO tree was most similar in three out of the four measures (RF-unweighted = 49, KC-unweighted = 78, KC-weighted = 10,084). There were 6,300 SNPs in the MLST dataset. The MLST SNP tree was the third most similar tree topologically to the original tree (RF-unweighted = 77, KC-unweighted = 258), behind the FFP amino acid alphabet tree. The FFP amino acid tree was found to be the most similar FFP alphabet tree with an RF unweighted distance of 63 and KC unweighted distance of 93.

While similarities to the Kurtzman and Robnett tree for the same taxa can be seen in varying degrees across all three FFP alphabet trees (See Figures 5.5, 5.6 and 5.7), the different alphabet choices have resulted in distinct tree topologies. The 20-letter amino acid alphabet tree (See Figure 5.7) was found to be the most similar to

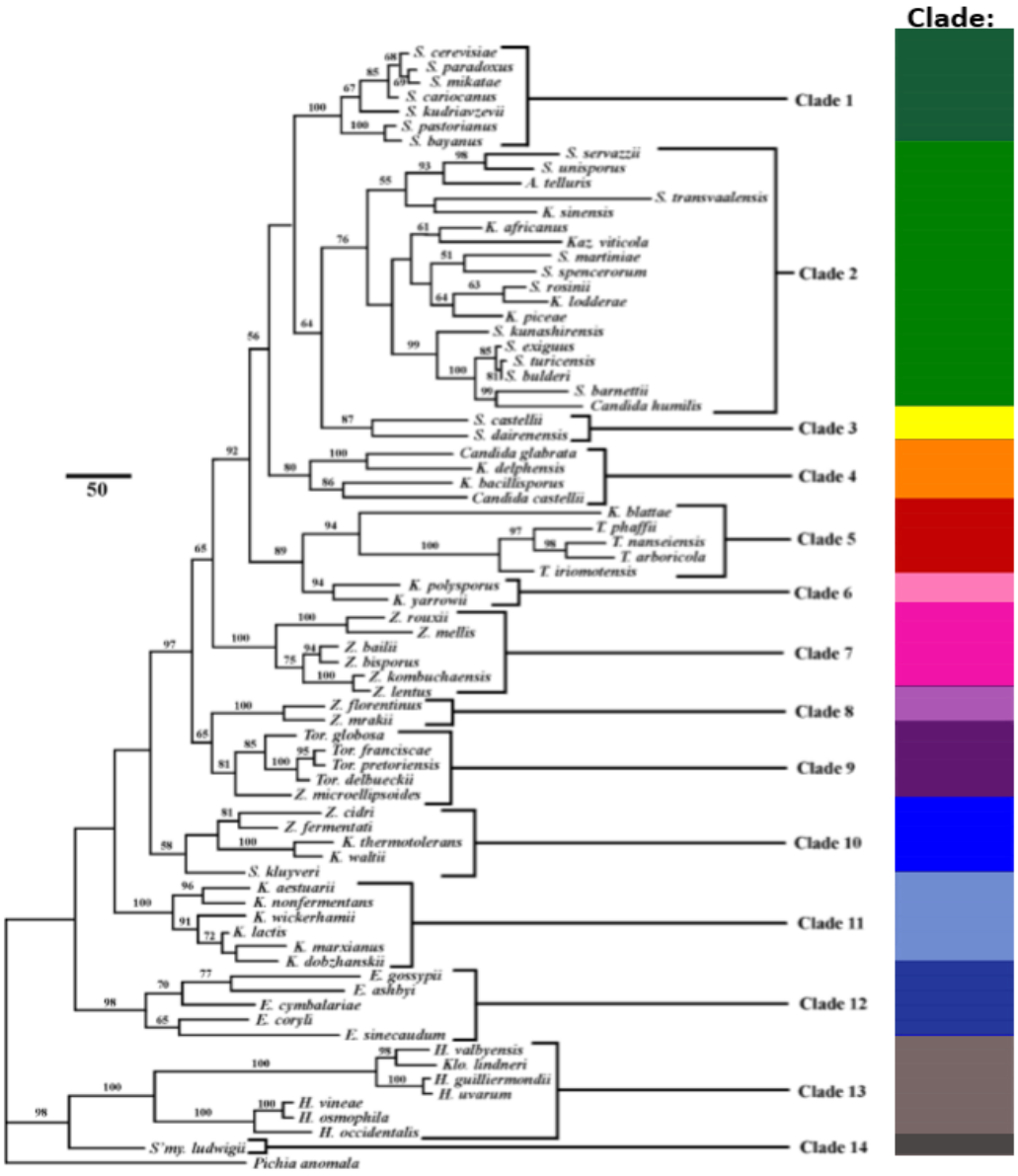


Figure 5.3: Originally published Kurtzman and Robnett tree of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (formally *Pichia anomala*) (Kurtzman and Robnett (2003)).

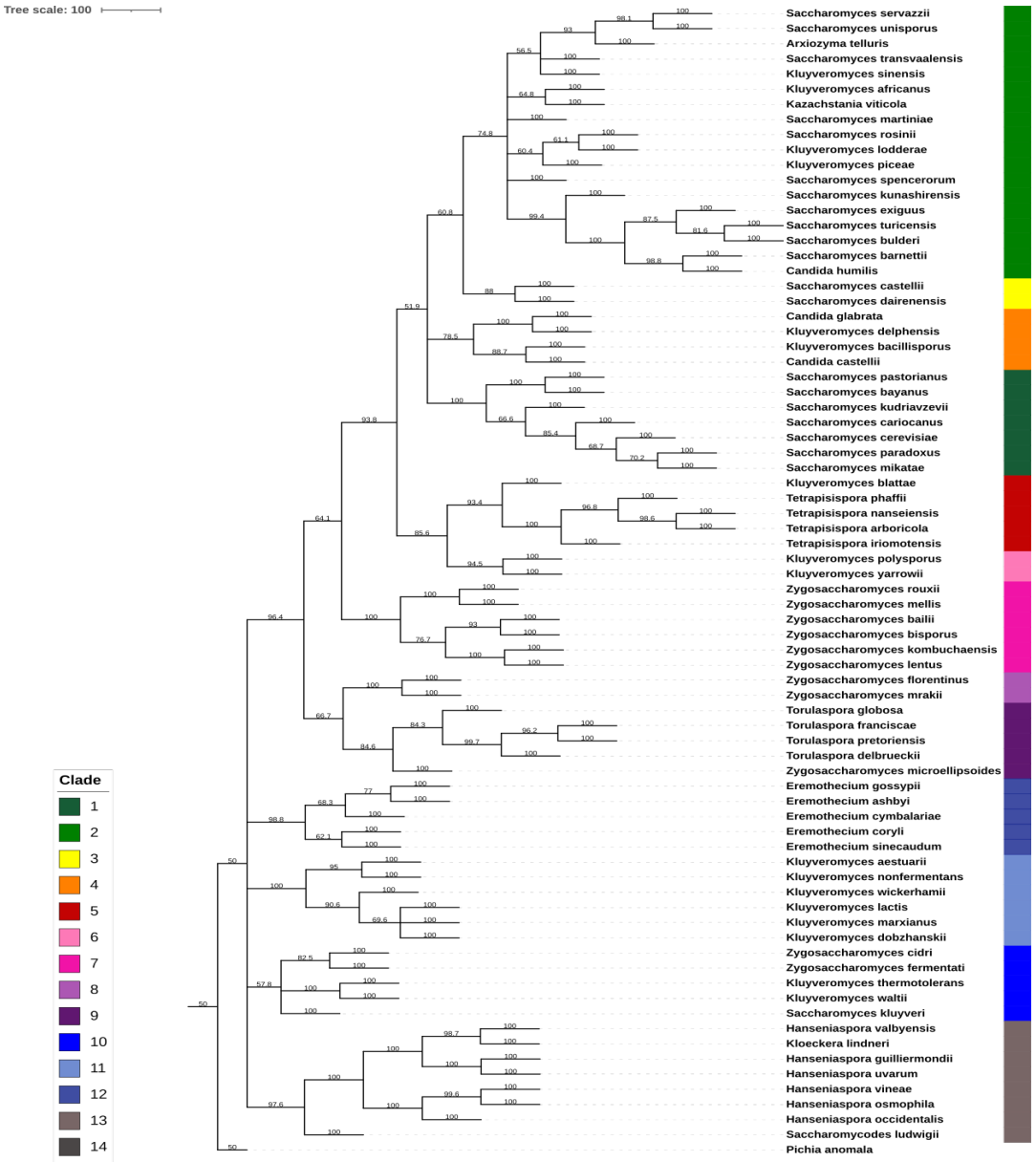


Figure 5.4: Newly estimated Kurtzman and Robnett tree of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (formally *Pichia anomala*) (Kurtzman and Robnett (2003)). Clade annotation is given compared to Figure 5.3.

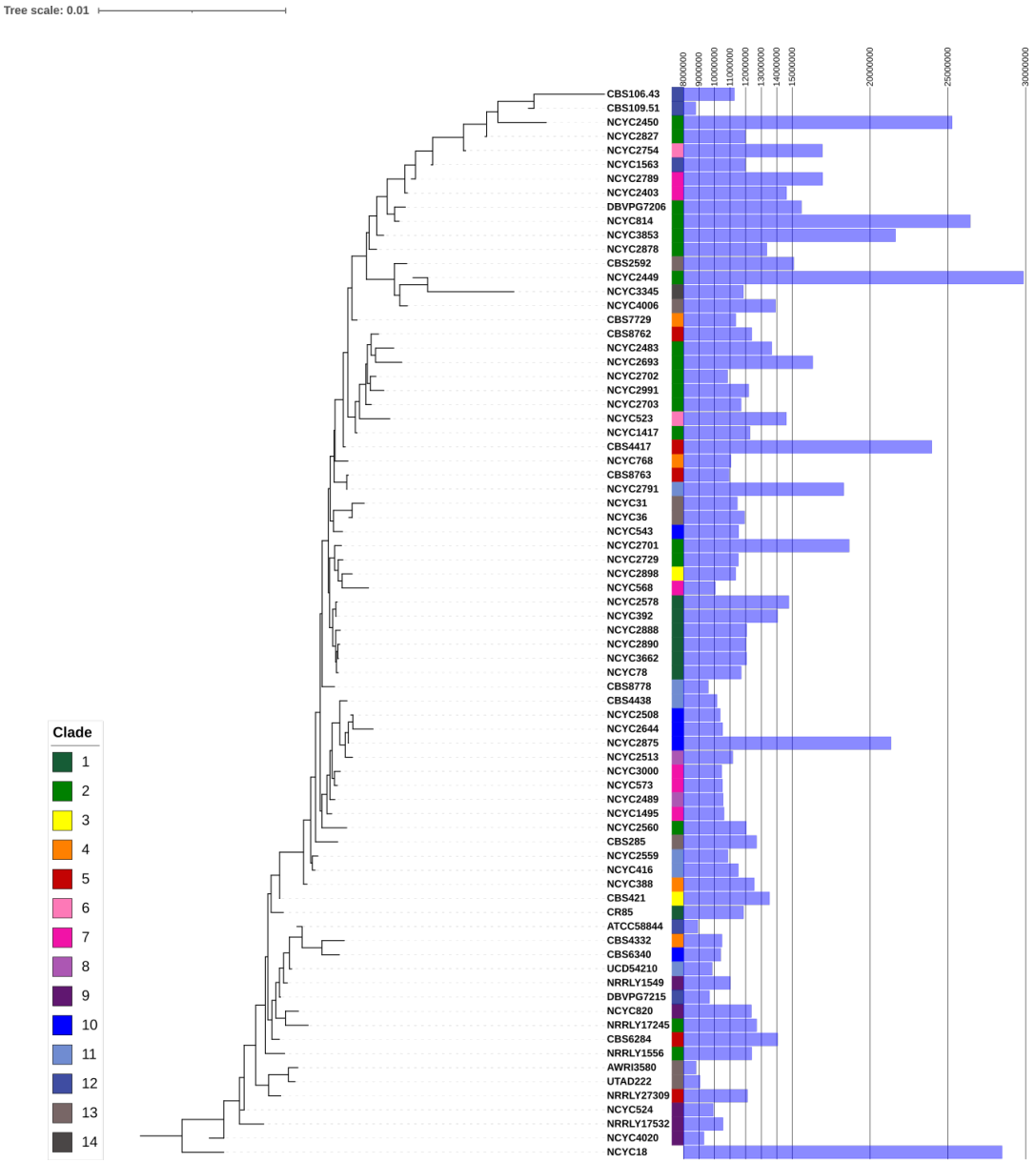


Figure 5.5: FFP 2-letter RY alphabet tree (Purines and Pyrimidines) of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (NCYC18). Clade annotation is given compared to Figure 5.3. Whole genome sizes are shown alongside each species as blue bars.

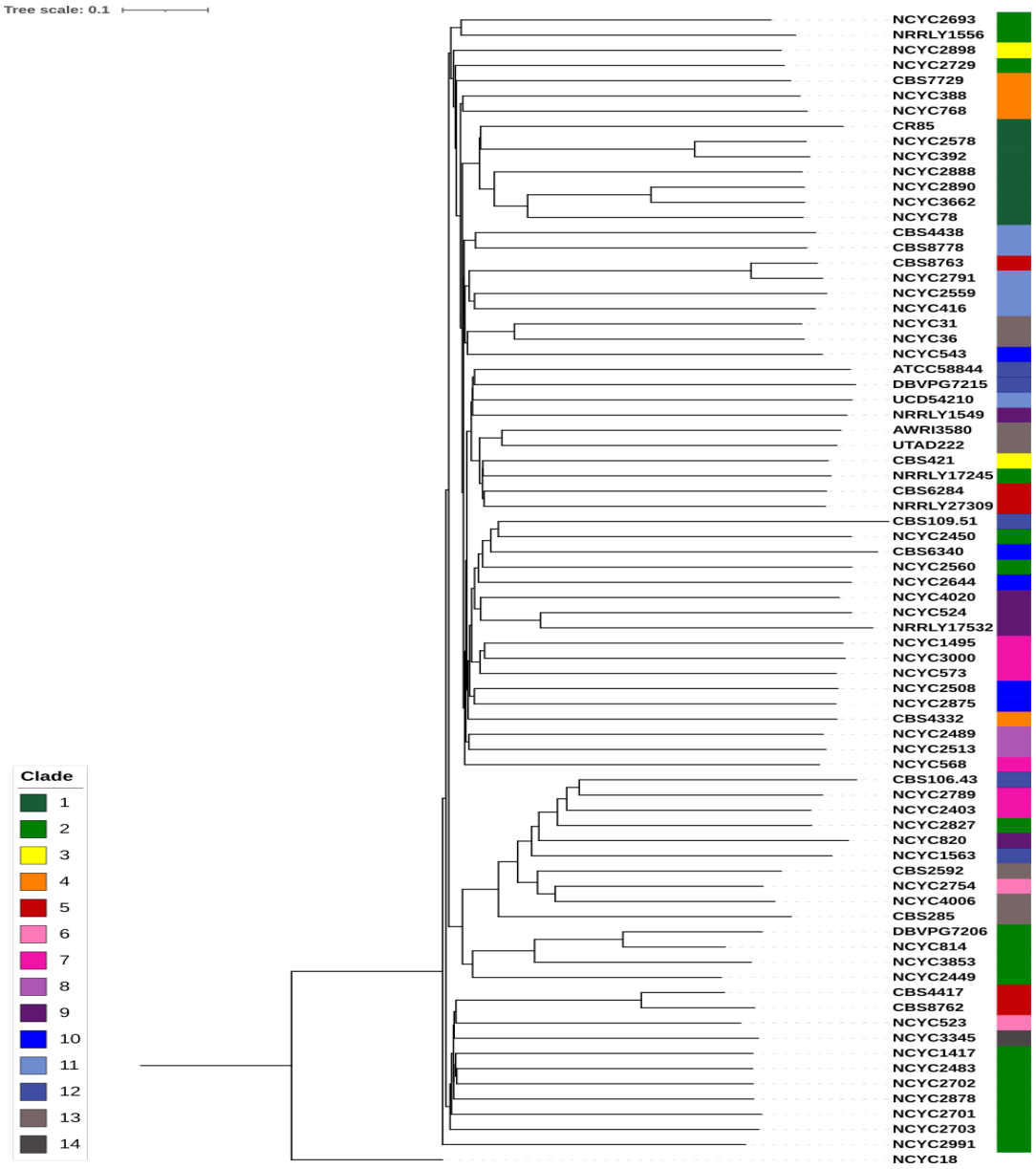


Figure 5.6: FFP 4-letter ACGT alphabet tree of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (NCYC18). Clade annotation is given compared to Figure 5.3.

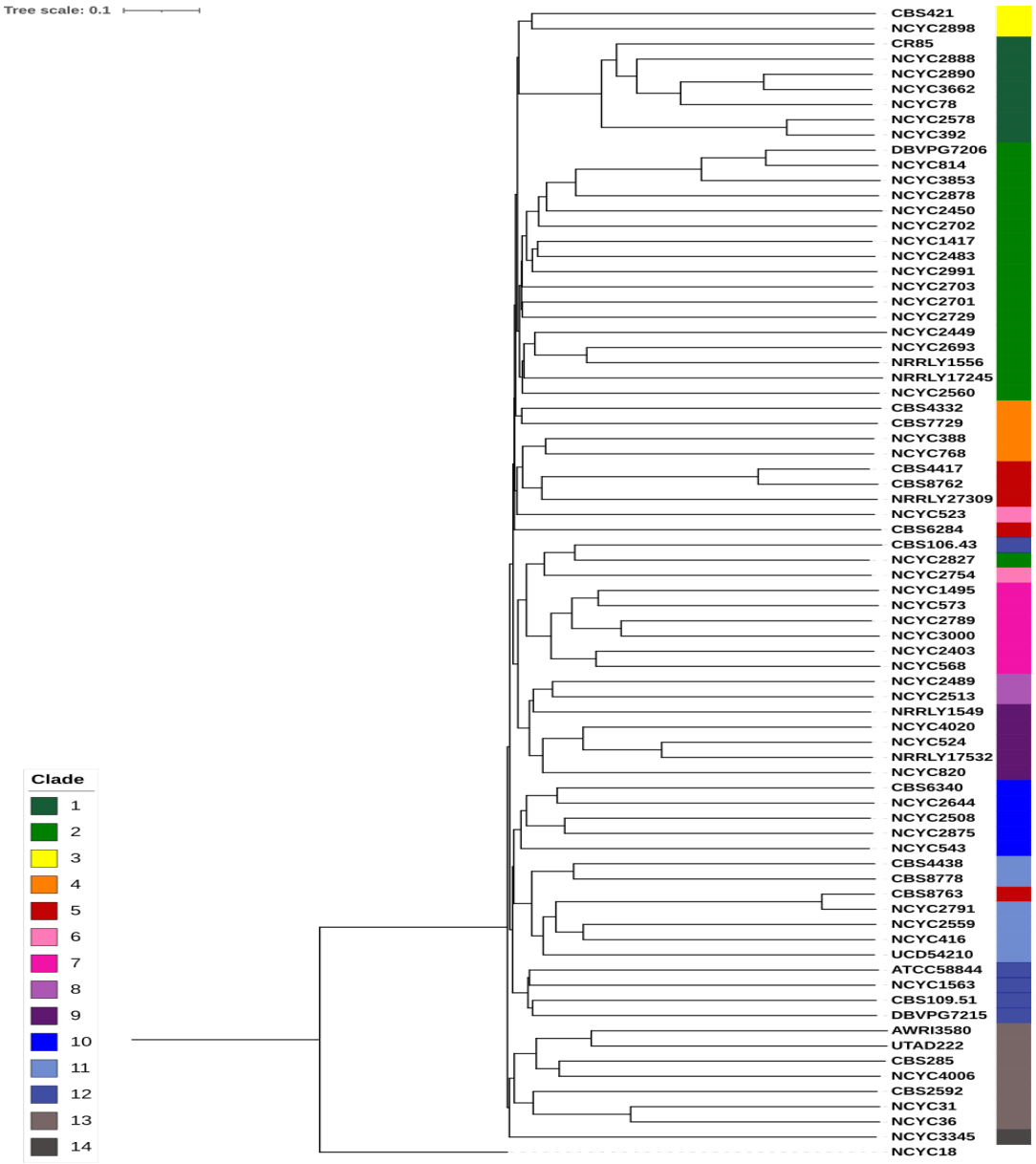


Figure 5.7: FFP 20-letter amino acid alphabet tree of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (NCYC18). Clade annotation is given compared to Figure 5.3.

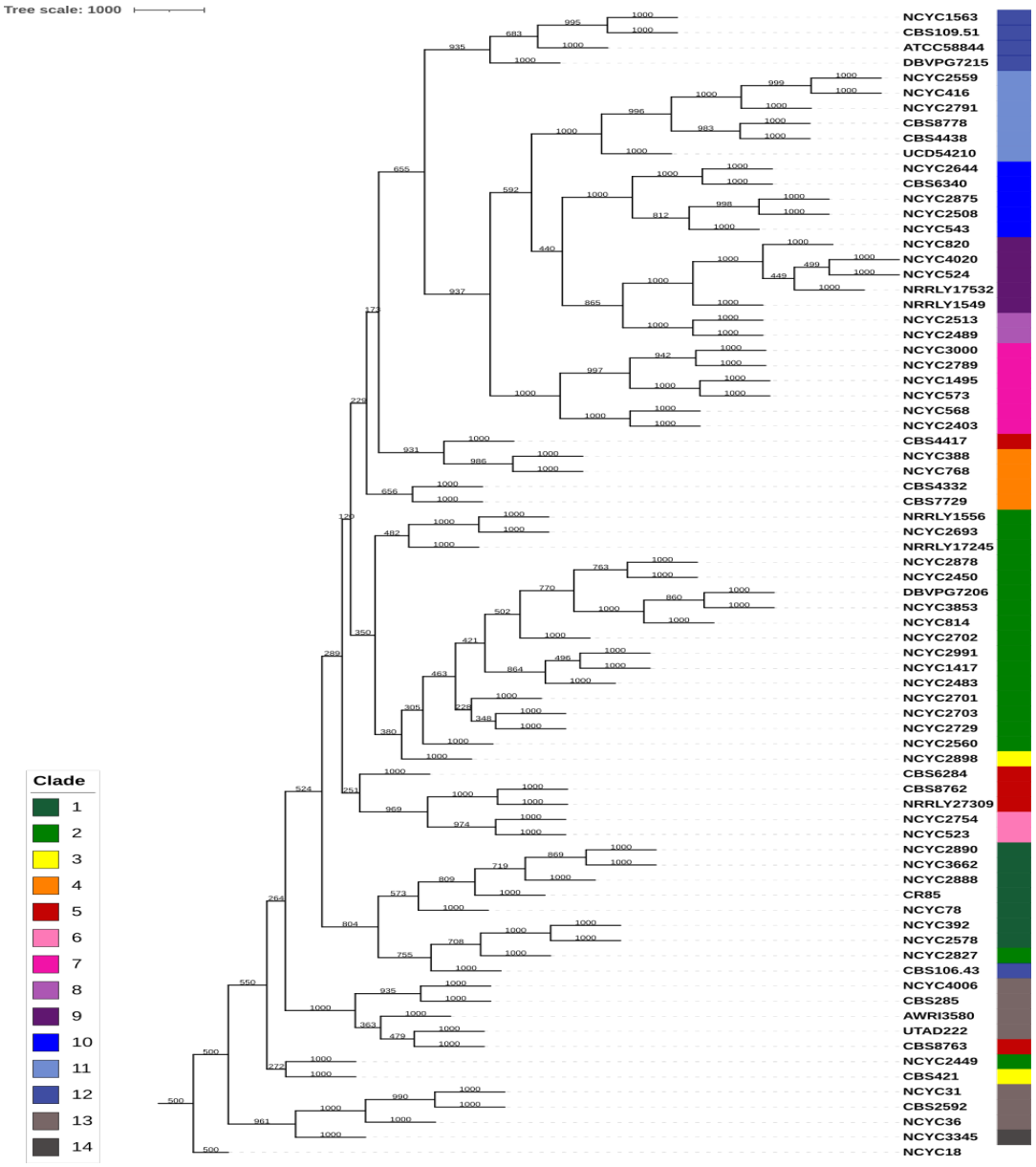


Figure 5.8: MLST SNP tree of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (NCYC18). Clade annotation is given compared to Figure 5.3.

Figure 5.9: BUSCO gene tree of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (NCYC18). Clade annotation is given compared to Figure 5.3. Log-likelihood of consensus tree is -38964070.90.

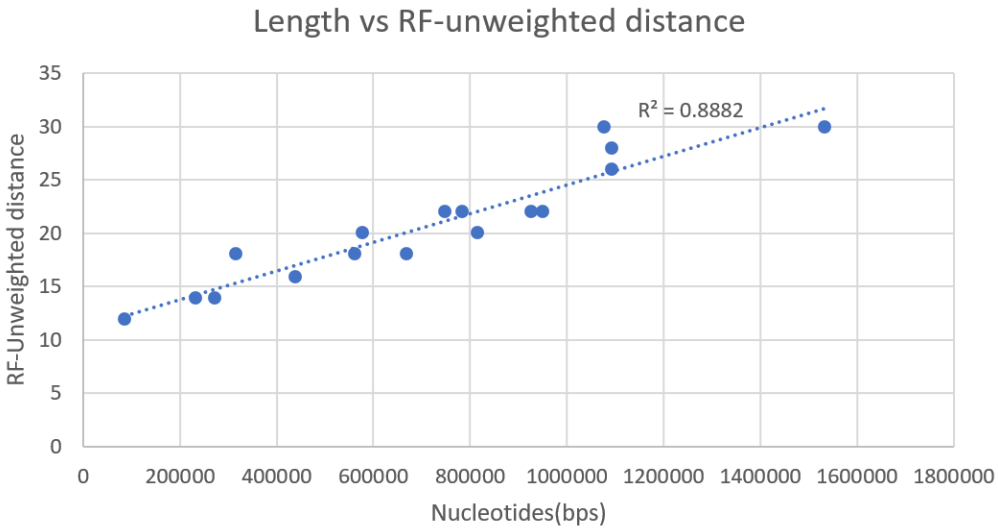


Figure 5.10: A plot of sequence length vs Robinson-Foulds unweighted distance for seventeen FFP two-letter alphabet (RY) trees shows a clear positive correlation between these two factors.

the Kurtzman and Robnett clading and clade ordering, whilst the RY-alphabet tree (See Figure 5.5) has the least congruence. Also, long branches were highly frequent within the FFP 20-letter amino acid and 4-letter DNA alphabet trees.

The variation in FFP tree topology was investigated further. An initial assessment of the affect of sequence length on tree accuracy was undertaken through a simulation study. Seventeen FFP trees were estimated using each of the three alphabet approaches, one for each of seventeen simulated chromosomal datasets evolved over a known phylogenetic tree, and were then compared to the true tree with the Robinson-Foulds distance. The resulting 4-letter DNA and amino acid alphabet trees showed no correlation between sequence length and tree accuracy (See Figures D.2 and D.3 of the Appendix) but a clear positive correlation was found for the RY-alphabet tree ($R^2 = 0.8882$) (See Figure 5.10). The genome size annotation of the 76 species FFP RY-alphabet tree in Figure 5.5 also appears to show moderate signs of clustering between genome size and tree topology.

Next, the association between genomic GC content and FFP tree topology were assessed with the full 76 species set as shown in Figure 5.11 (BUSCO and GC), Figure

5.12 (FFP AA and GC) and Figure 5.13 (FFP ACGT and GC). Both the BUSCO and FFP AA trees indicate that the distribution of GC content among strain genomes is essentially random (i.e. is not correlated with the tree topology). In contrast, the topology of the FFP ACGT trees appears highly influenced by GC content, with clusters of species with similar GC values.

5.6 Discussion

The results of all five tree building approaches investigated here showed some similarity to the expected Kurtzman and Robnett topology but there was significant variation between them. The complex BUSCO approach, which takes a longer time to perform compared to the other approaches tested here, estimated the tree found to be the most similar to the expected topology. The second most similar tree topology was that seen within the FFP amino acid alphabet tree. The two approaches are highly different (k-mer based vs concatenated genes) as well as the data used. The BUSCO tree was built from 1,541 aligned orthologous genes whilst the amino acid tree used word frequencies extracted from the full unaligned proteomes. As discussed in the Core genome chapter, deciding on what number of genes is best for building the most accurate tree can be crucial. The BUSCO approach examined orthologous genes filtered by steps to remove spurious sequences. As noted previously, the assembly quality of a number of genomes was low and this in turn can affect the genes predicted by AUGUSTUS. With the full proteome, spurious sequences may have played a part in the tree mislocation of at least one case, that of CBS106.43. In the BUSCO tree the signal was likely sufficient to place the strain in the expected clade but in the amino acid tree, which relied upon a larger and potentially more divergent dataset than the BUSCO tree, the noise of spurious genes likely affected its correct position in the tree. This is another benefit to the BUSCO concatenated gene approach along with the highest similarity scores to the expected topology.

Interestingly, the observed poor assembly of CBS106.43, which is indeed likely *Eremothecium ashbyi*, was found beside or close to NCYC2827 *(Kazachstania rosinii* of clade 2) in all trees other than the BUSCO tree. One hypothesis for grouping with NCYC2827 could be contamination of NCYC2827 with CBS106.43, as they were se-

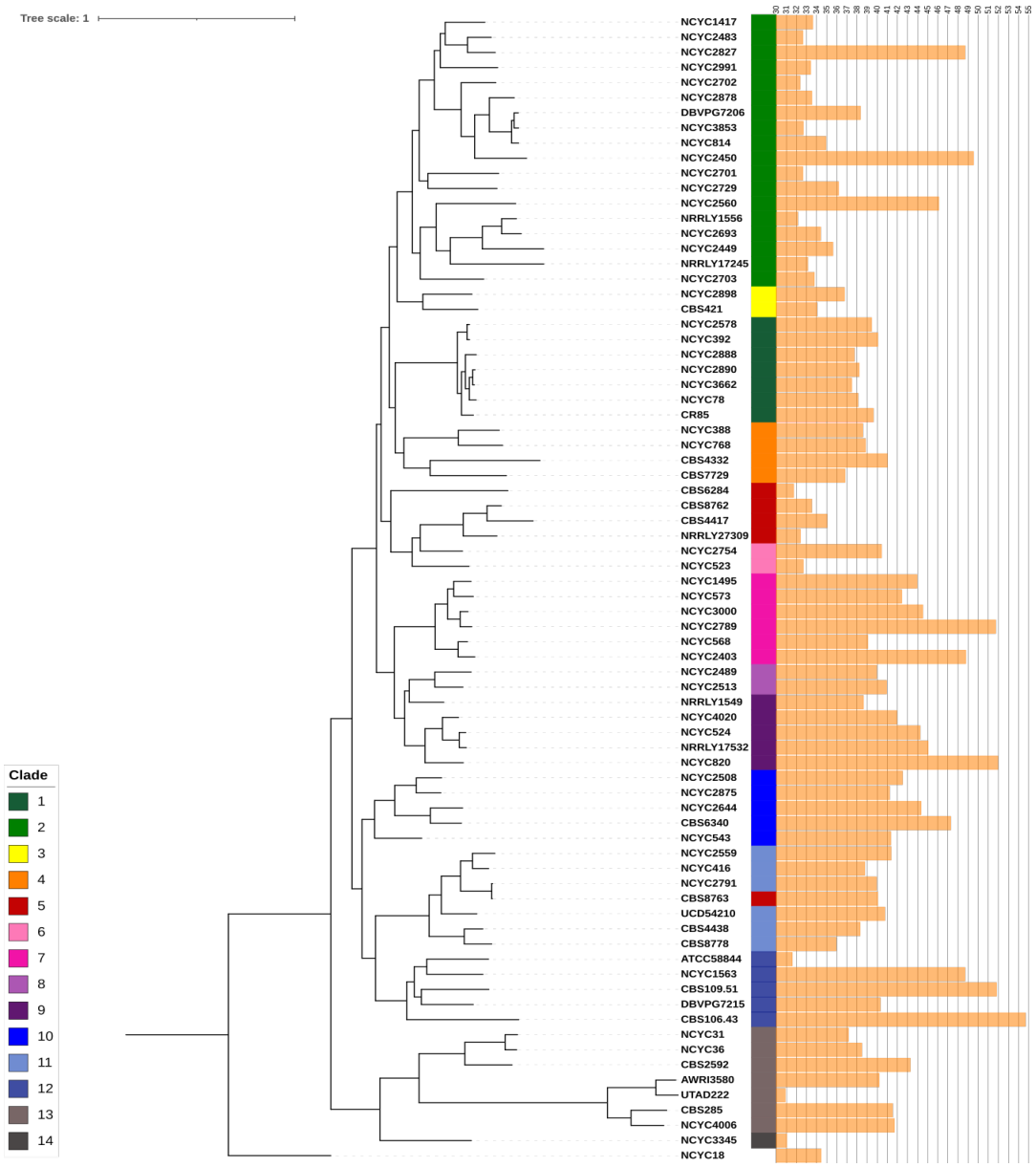


Figure 5.11: BUSCO gene tree (n = 1,711) of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (NCYC18). Whole genome GC contents are shown alongside each species as orange bars.

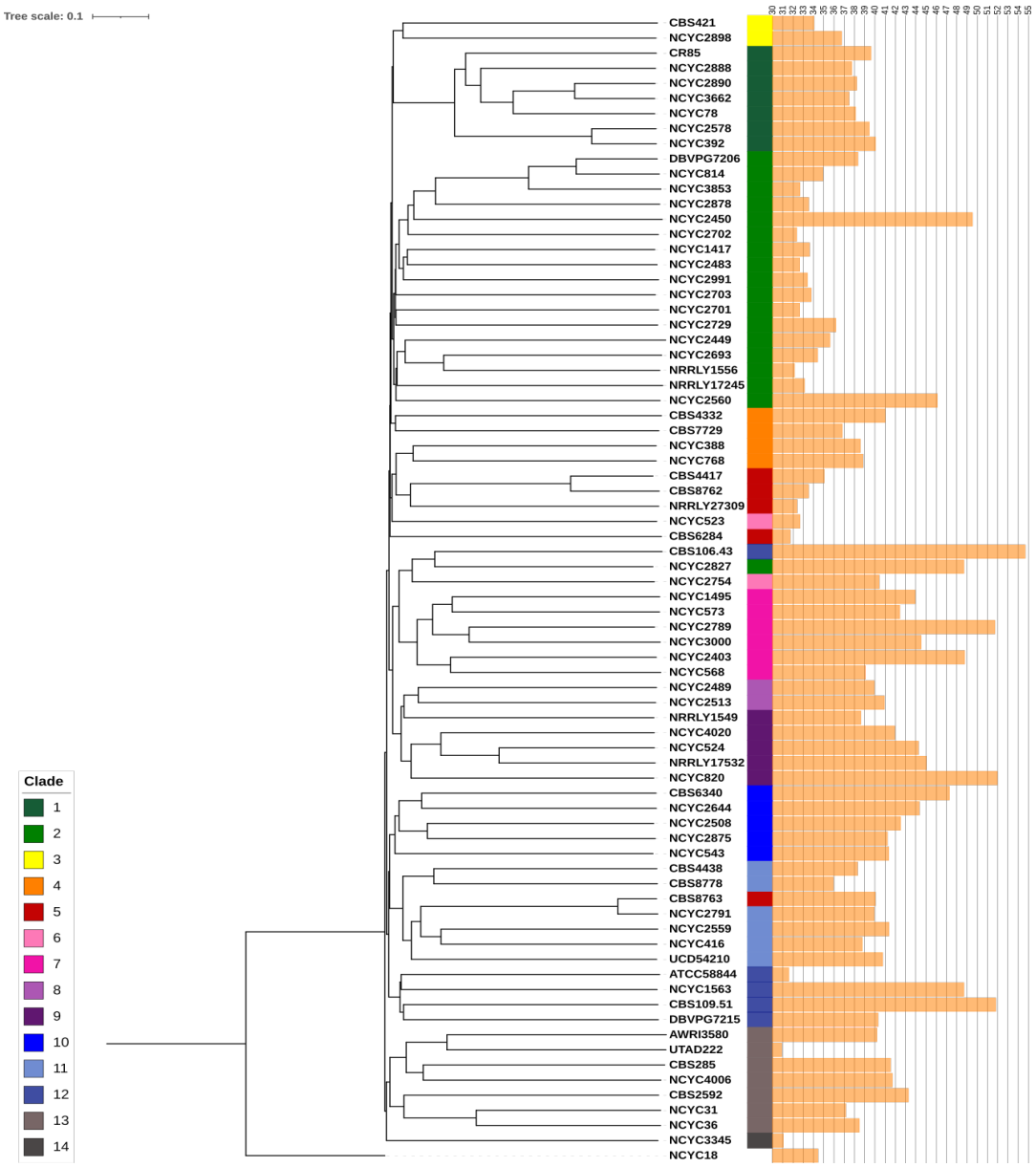


Figure 5.12: FFP tree using the twenty-letter amino acid alphabet of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (NCYC18). Whole genome GC contents are shown alongside each species as orange bars.

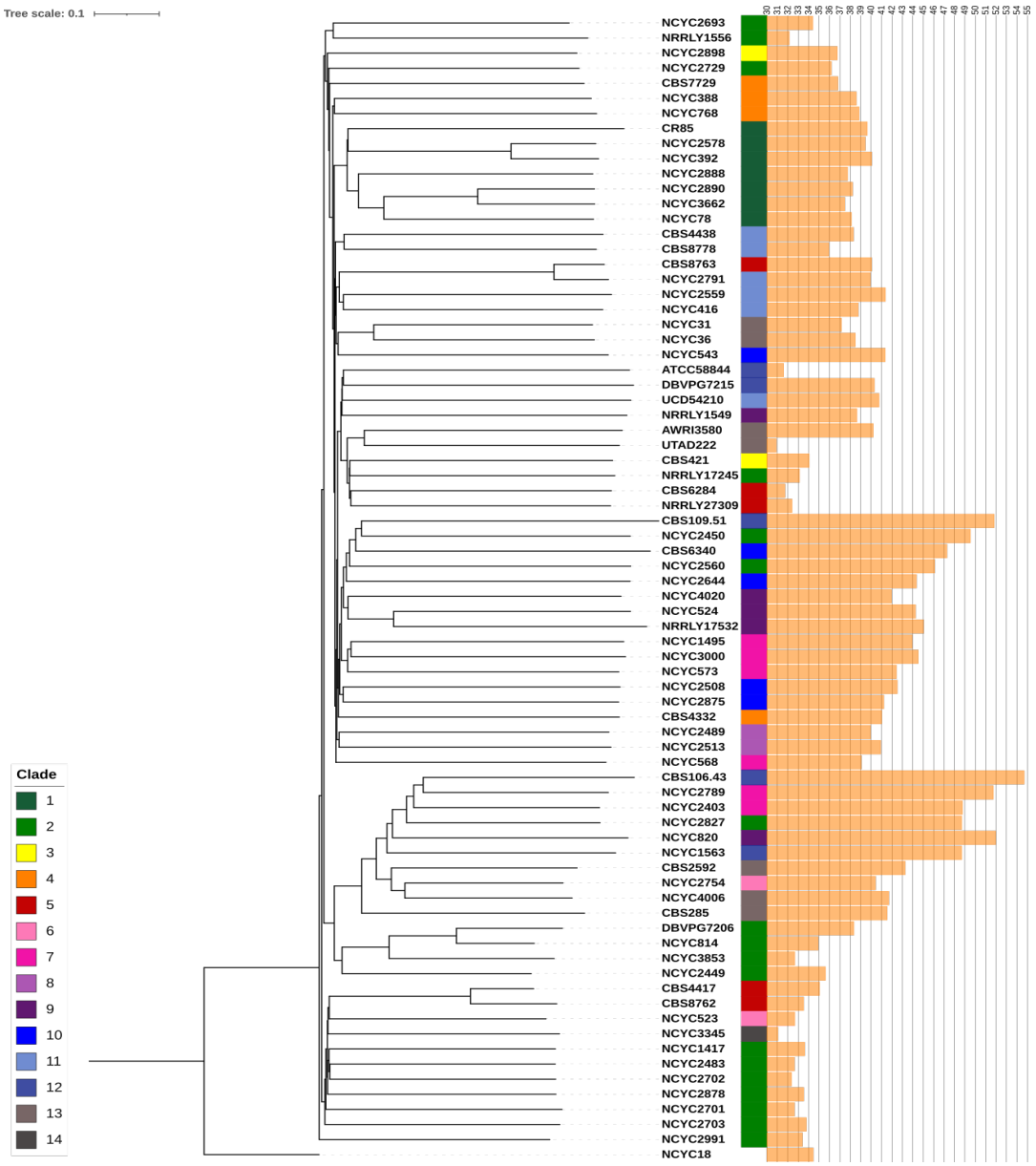


Figure 5.13: FFP tree using the four-letter DNA alphabet (ACGT) of seventy-five Saccharomyces complex species and outgroup *Wickerhamomyces anomalus* (NCYC18). Whole genome GC contents are shown alongside each species as orange bars.

quenced on the same plate. The FFP approaches in particular may be sensitive to contamination due to the broad and largely unfiltered dataset they use. Conversely, a strain found to be 'out of place' in the BUSCO tree was CBS8763. This strain, which was originally thought to be the species *Tetrapisispora nansiensis* has recently been identified as CBS8733 *Hanseniaspora opuntiae* (Carmen Nueno Palop, pers. comm.), the likely result of a strain mixup.

The MLST SNP tree was built from single nucleotide differences within the seven MLST regions of the 76 species. The resulting tree had clear similarities to the Kurtzman and Robnett tree but less than would be expected. One explanation for a greater than expected divergence could be missing genes. Of the 14 publicly available genomes included in the dataset, seven had no mitochondrial genomes. Of those seven strains, two mitochondrial genomes were successfully sourced from GenBank and merged with the corresponding nuclear genomes. Irregardless, with five species missing their mitochondrial genome, the phylogentic signal in this tree was almost certainly affected. Two out of the six MLST genes are mitochondrial (COXII and ssRNA) which will have biased the tree to some degree, with further potential effects from the misclassified CBS8763 sequence and the poorly assembled CBS106.43 genome.

The three FFP methods each resulted in trees with quite different taxon ordering, as seen in Figures 5.5, 5.6 and 5.7. The distance metrics also reflected this difference. The amino acid tree is the most similar FFP alphabet tree topologically to the clade ordering in the Kurtzman and Robnett tree (which used multiple conserved gene sequences). The observed difference between the FFP DNA and amino acid alphabet trees is perhaps because the amino acid code masks many differences observed at the DNA level, including all inter-genic regions, and therefore is more conservative. The FFP alphabet tree which was most dissimilar to the original tree was that derived using the RY alphabet. The large topology-only measures seen for the RY alphabet full genome tree may be partly explained when considering the study on accuracy (using the Robinson-Foulds unweighted metric as the measure of accuracy) and sequence length. Using this approach, topological accuracy clearly decreased as sequence length increased, which may at least partly explain why the topology of the RY full genome tree was divergent. The underlying reason for such a bias is yet to

be understood but could be a result of the smaller alphabet, which would likely lead to a weakening of the phylogenetic signal. Finally, the FFP 4-letter alphabet showed a genome clustering that appears to correlate with GC content.

5.7 Conclusions

The main aim of this project was to compare the accuracies of a number of state-of-the-art phylogenetic methodologies using a large, exemplar yeast NGS dataset. This aim was achieved here by comparing five different tree building approaches using a 76 species yeast dataset. All trees constructed using the same yeast dataset were found to be different. The BUSCO approach estimated the tree found to be the most similar to the expected topology and despite it being more complex and time consuming than the other approaches it would be the recommended choice for future yeast phylogenetic studies.

The study showed that there are clear algorithmic differences in how these methods generate trees, highlighting the need for comparative studies of different approaches. Comparing and testing the accuracy of different methods is very important as an incorrect taxonomy can greatly affect down-stream inferences. The causes for these differences may be the result of the vastly different amounts and types of data used. Datasets can range from whole genomes to a small number of conserved genes, whilst the type of data used could include coding, non-coding or amino acid sequence. Alignment free approaches have great potential when it come to dealing with large whole genome datasets which are becoming more common in phylogenetic studies. Using whole genomes can also harness a large amount of phylogenetic signal, which can aid in building the most accurate trees. However, issues have been identified in this study with the FFP software, which appears to have a sequence length bias with the two-letter RY alphabet and a GC bias with the four-letter DNA alphabet. As a result, neither of these FFP alphabet approaches can currently be recommended for use.

The next chapter discusses GC content within genomes and investigates this GC bias further with a simulation study. The early stages of a new alignment-free software tool for phylogenetic analysis, which will aim to overcome such biases, is also touched

upon.

Chapter 6

Testing for a GC bias in the FFP software

6.1 Summary

- GC bias seen in the FFP four-letter DNA alphabet tree building approach is investigated with a simulation study.
- GC content by codon position is assessed.
- A new piece of alignment-free software is developed.

6.2 Introduction

Guanine and Cytosine (G+C) are two of the four nucleic acids of DNA and are found at varying levels in the genomes of species across the tree of life. GC composition may be described at three levels: 1) **Overall GC content**, which in living organisms varies from 25% to 75% (Sueoka (1988)); 2) **Local GC composition**, which is mostly defined based on the positions within the genetic codon triplets - GC1, GC2 and GC3 denote the GC composition at the first, second and third site of codons, respectively - but with additional definitions for exonic fourfold-redundant sites (GC4) and intronic GC content (GCi); 3) The **ratio of G/C or A/T** within a single strand of DNA.

Complementary G+C pairs within the DNA double helix have three hydrogen bonds connecting them whereas A+T (Adenine and Thyamine) have two. This addi-

tional bond plays a role in the stability and secondary structure of DNA, as seen with the higher melting points needed for GC-rich sequences in PCR reactions. As a result, GC content was initially thought to be an indicator of the thermal environments in which an organism could survive, but bendability of the DNA structure rather than thermostability was found to correlate with higher GC content. This feature is thought to be related to active transcription in gene-rich genomic regions (Hurst and Merchant (2001), Vinogradov (2003)). Conversely, GC content in structural RNA was found to be correlated with optimal growth temperature. (Galtier and Lobry (1997)).

High GC content in yeast is often found in compact genomes as genes are GC-rich but it has also previously been linked to recombination (Bradnam et al. (1999), Birdsell (2002), Marsolier-Kergoat and Yeramian (2009), Lynch et al. (2010)). Recombination involves DNA repair, a process which is known to be biased toward GC-richness in mammals (Brown and Jiricny (1988)). Gene conversion is another process thought to be biased by GC content in yeast and in other eukaryotes (Pessia et al. (2012), Marais (2003)).

As shown in the previous chapter, species within the Saccharomyces complex appear to cluster together, within an FFP nucleotide-generated tree, by GC content rather than evolutionary relationship. This observation suggests a potential bias in the software. This "GC-attraction" is reminiscent of the well-known long-branch attraction (Carmean and Crespi (1995), Bergsten (2005)) phenomenon in phylogenetic studies. Long-branch attraction is the erroneous grouping of two or more long branches as sister groups due to methodological artifacts. However, it remains to be seen whether this observation is limited to the dataset in question or whether it is a more widespread methodological bias. The aim of this chapter was therefore to confirm or refute the presence of a GC bias in the FFP software by means of a controlled simulation study.

Clade	Strain	Species name	GC content
1	NCYC78	*Saccharomyces cerevisiae*	39.7%
2	NCYC2701	*Kazachstania viticola*	33.9%
3	NCYC2898	*Naumovozyma castellii*	37.6%
4	NCYC388	*Candida glabrata*	40.5%
4	NCYC768*	*Nakaseomyces delphensis*	41.8%
5	CBS4417	*Tetrapisispora phaffii*	37.3%
6	NCYC523	*Vanderwaltozyma polyspora*	34.5%
7	NCYC568	*Zygosaccharomyces rouxii*	40.1%
8	NCYC2489	*Zygotorulaspora mrakii*	40.6%
9	NCYC4020	*Torulaspora delbrueckii*	43.0%
10	NCYC2875	*Lachancea cidri*	42.4%
11	NCYC2791	*Kluyveromyces marxianus*	41.6%
12	ATCC58844	*Eremothecium sinecaudum*	41.5%
12	CBS109.51	*Eremothecium gossypii*	51.9%
13	NCYC31	*Hanseniaspora osmophila*	39.5%

Table 6.1: Clade information, Strain ID, Species name and GC content for 15 Saccharomyces complex species used in a GC simulation study (*= NCYC768, the GC-mutated strain).

6.3 Methods

6.3.1 Dataset

Previous chapters have introduced a 76 species dataset (one strain for each of 75 Saccharomyces complex species plus outgroup). A further assessment of the GC content of this dataset was made. A subset of this dataset was subsequently chosen for a GC-focused simulation study. This smaller dataset consisted of 15 Saccharomyces complex genomes chosen so that it contained a minimum of one strain from each genus and encapsulated a broad range of genome-wide GC contents (See Table 6.1). A custom bash script was used to extract genes and remove introns from the contigs of these strains. This script included the use of the BEDtools (Quinlan and Hall (2010)) v2.25 software for Gene Feature Format (GFF) file generation.

6.3.2 GC simulation

A custom Python script, co-written by Dr Jo Dicks, was used to change the GC content of the genes of NCYC768 without altering the encoded amino acid sequence. For simplicity, GC content was increased or decreased at random, exclusively at the third codon position only where it led to a neutral mutation. The original GC content of this strains' genic dataset was 41.8%. The script decreased this total GC content value to 40%, 35% and 30% and increased it to 45% and 50%. All mutation scenarios were run 10 separate times to ensure results were not artefactual. Gene sequences of the remaining 14 strains were left unaltered.

6.3.3 Tree building

FFP (Sims et al. (2009a)) 2v.3.0 (k=14, four-letter DNA alphabet) was used to generate distance matrices from the 50 simulated datasets (10 replicates for each of 5 GC contents). PHYLIP's (Felsenstein (1989)) v3.695 neighbor program was used to generate neighbor-joining trees from the distance matrices. The consense program was used to build consensus trees from the ten replicate trees resulting from each mutated GC percentage. iTol (Letunic and Bork (2006)) v5.7 was used to view and annotate the trees.

6.3.4 Software development

A custom version of the FFP approach was developed within this study, with the intention of developing an unbiased strand of the approach. The software, which was still in development by the end of this study, consisted of a bash script and a Python script calling the third-party Jellyfish software (Marçais and Kingsford (2011)) v2.0 for the k-mer counting step. All code can be found at https://github.com/aKeaneScientist/jellyphy.

6.4 Results

6.4.1 GC content by codon position

An assessment of GC content was undertaken on the full 76 species dataset. The total coding region GC content and GC content at each codon position was checked for correlation (See Figure 6.1). There is clearly a strong positive correlation (R^2= 0.758)

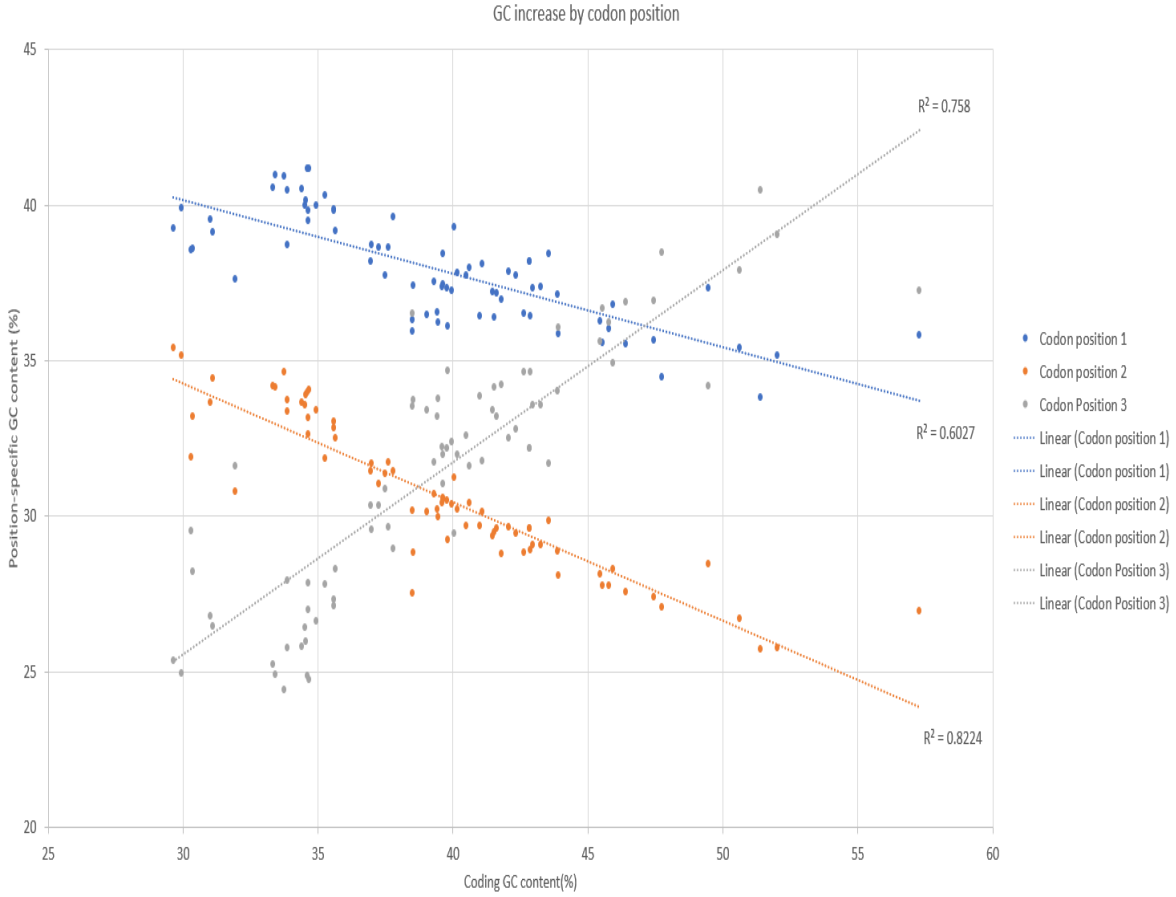


Figure 6.1: Relationships between GC content at first, second and third codon positions (blue, orange and grey points respectively) and overall coding GC content, for each of 75 Saccharomyces complex species and outgroup.

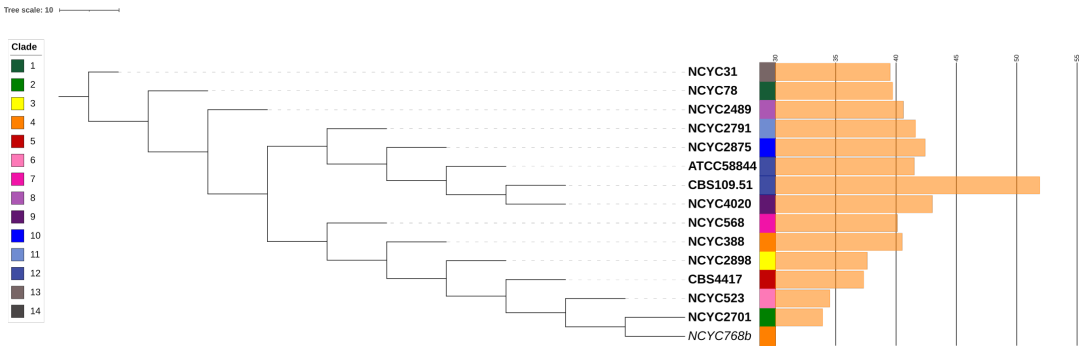
between GC at codon position 3 and coding region GC and negative correlations between GC at codon positions 2 (R^2= 0.822) and 3 (R^2= 0.603) and coding region GC. The relationships are likely to be the result, at least in part, of the redundancy in the amino acid code whereby changes at position 3 are often synonymous.

6.4.2 GC content and FFP trees

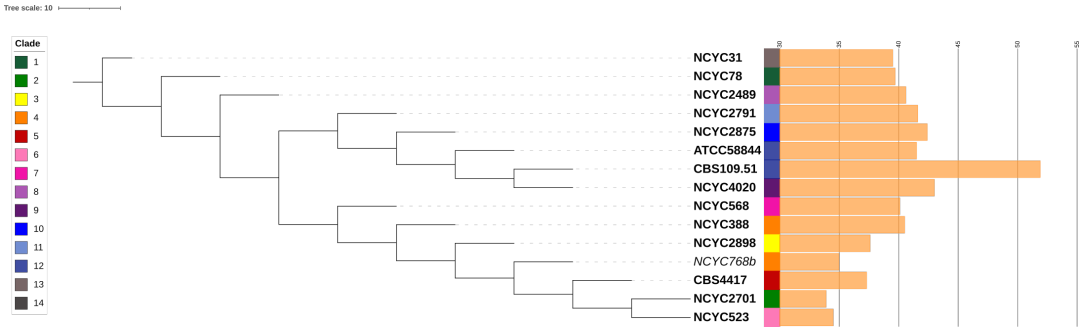
The GC content of the extracted genes (without introns) from each strain is shown in Table 6.1. The first tree (See Figure 6.2d) was built with this initial, unmutated dataset. As has been seen in the previous chapter, the branches are long and there appears to be a clustering of strains by GC content. However, phylogenetically, NCYC768 sits where expected, beside a strain of the same clade (NCYC388).

The other five trees seen in Figure 6.2 show the same datasets with decreased or increased GC content in NCYC768. Each tree is a consensus tree of 10 trees built from independently mutated NCYC768 gene sequences. While the trees all had the same topologies (See Figures E.1 to E.5 of the Appendix for all 50 trees), there were clear differences between most of the trees. First, the GC content of NCYC768 was reduced from 41.8% to 40% and as shown in Figure 6.2c, there was no topological difference between this tree and that seen in Figure 6.2d. The only difference in this tree to the original is branch length, which is an artefact of the 'mutant' trees being consensus trees, as mentioned above. The next tree includes NCYC768 with a GC content of 35% (See Figure 6.2b). Here the strain can be seen to move away from the original position, closer to lower GC content strains. The tree including the lowest GC version of NCYC768 (30%) can be seen in Figure 6.2a. The strain moved to the bottom of the tree, more distal than the lowest GC content strains.

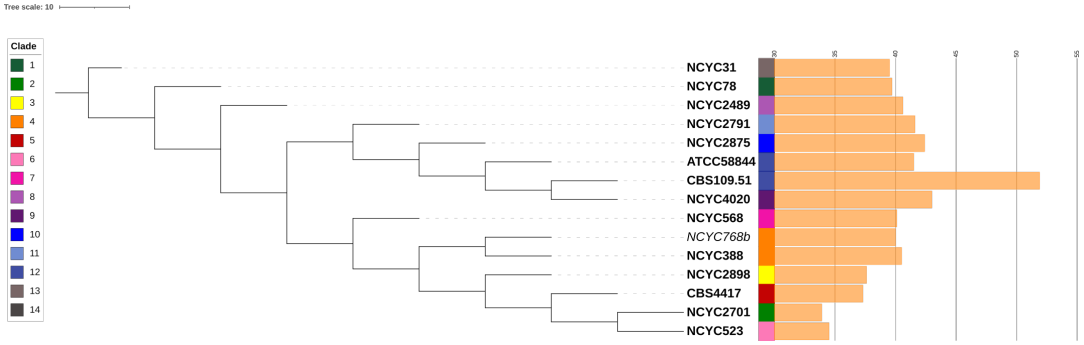
The GC content of NCYC768 was also increased to 45% and 50%. The tree with the mutated 45% GC content is shown in Figure 6.2e. Here the strain can be seen to move a large distance away from its original position adjacent to NCYC388, instead grouping with the two highest GC content genomes CBS109.51 (51.9%) and NCYC4020 (43%). Finally, the last tree (Figure 6.2f) shows the GC content of NCYC768 increased to 50%. The strain now sits as a sister strain to the highest GC strain in the set, CBS109.51.



(a) GC= 30%



(b) GC= 35%



(c) GC= 40%

6.5 Discussion

GC content varies within genomes across the tree of life and has been shown to vary significantly within the Saccharomyces complex. In the previous chapter, the GC content of each strain was annotated on the FFP four-letter DNA alphabet tree where an apparent GC-clustering was seen (See Figure 5.13). This unexpected finding shows species clustering together by GC content rather than evolutionary relationship, suggesting a significant bias in the methodological approach. This observation led to the undertaking of a simulation study to confirm or refute this finding. The study confirmed that a clear GC bias is present in phylogenetic tree building with the FFP

approach used with the four-letter nucleotide alphabet. As the GC content of a chosen strain was increased and decreased at random, but with its corresponding amino acid sequence fixed, the position of the strain (NCYC768) within the resulting tree moved in the direction of strains with similar GC content.

This is the first known study uncovering a GC bias in the FFP four-letter approach. The impact of this bias may lead to the building of inaccurate phylogenetic trees which can greatly affect downstream inferences. Although the alignment-free FFP approach to phylogenetic inference has great potential as it uses the full genomes of strains, and thus the maximum amount of information available to aid in accurate tree building, use of the nucleotide-based method would not be recommended at present. More thought is required regarding how to account for this variation in GC content when using k-mer based approaches. Potential ideas to account for GC variation in such an approach would be weighting of a genome's GC content depending on the average GC content of all genomes undergoing analysis.

A second issue with both the four-letter DNA and amino acid alphabet approaches, which was not investigated closely here, was the particularly long branches as shown in Chapter 5. It is hypothesised that this may be the result of the k-mer distance used. In the current FFP schema, two distinct k-mers are considered different to the same degree, irregardless of their similarity. This means, for example, that the k-mer **ACCTGATTGAAC** is considered as different from **ACCTGCTTGAAC** (one nucleotide difference) as it is from **CTAGCCAGTGTA** (twelve nucleotide differences), which is likely counter to their biological relationship and would potentially inflate pairwise distances. Using k-mer distance measures that did not require exact matching, is one possible way to account for this (e.g. binning of 'close' rather than exact k-mers). Generally speaking, the distance measure is currently too weak and cannot extract enough of the evolutionary signal within the data.

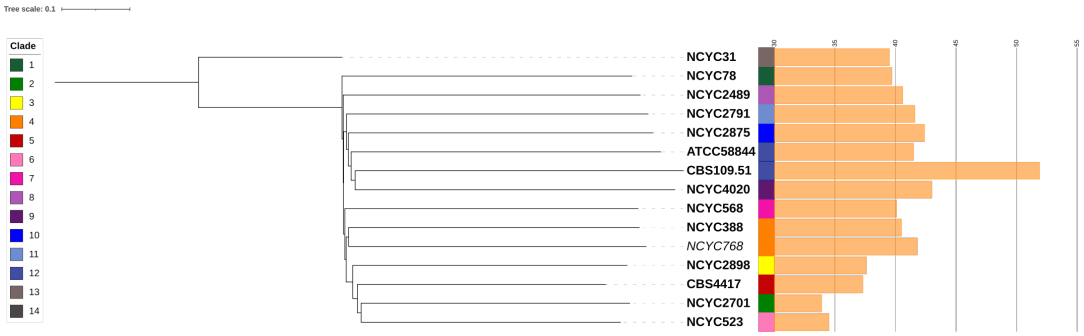
Aside from the long branching of the amino acid alphabet tree, the topology in the previous chapter (See Figure 5.7 in Chapter 5) was very similar to that expected. This appears to be the optimal FFP alphabet choice, but is it truly accurate? The accuracy of this alphabet approach was tested recently in a simulation study and shown to be less accurate than standard alignment-based approaches (Li et al. (2020)). Once

again, this may be the result of the degree of similarity between two k-mers which may not only affect the length of branches but also the phylogenetic relationship between two datasets.

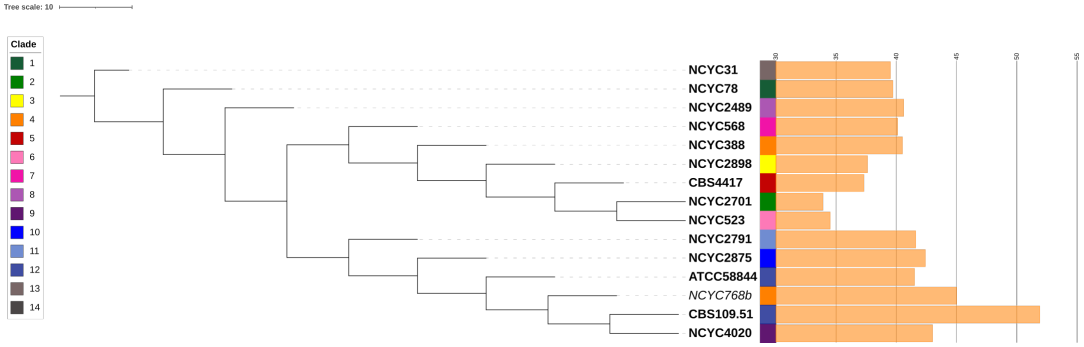
These results highlight the importance of further investigation into the accuracies of different phylogenetic approaches. In an effort to improve the current FFP approach, both with regards to the accuracy and computational efficiency of the algorithm, a new piece of alignment-free software written in Python and wrapping the Jellyfish software for efficient k-mer counting, was under construction towards the end of this project. The base program was successfully built and tested with small datasets, with resulting phylogenies then compared to FFP tree topologies. The total computation time taken to generate a distance matrix from 22 full yeast genomes was 3 hours on a single thread of the NCYC compute server. This is significantly faster than the original FFP software (v3.19) which would take days for the same dataset to run but slower than the new FFP software (2v3.0), which is multi-threaded. The next step in the process would be to improve the distance measure, currently the Jensen-Shannon divergence measure to account for both GC bias and long branches. The current version of the software can be found at https://github.com/aKeaneScientist/jellyphy. If time had permitted the new software would have been completed but more time was required on strain sequencing and quality control than had been anticipated. Further development of the new software will be a priority for future work.

6.6 Conclusions

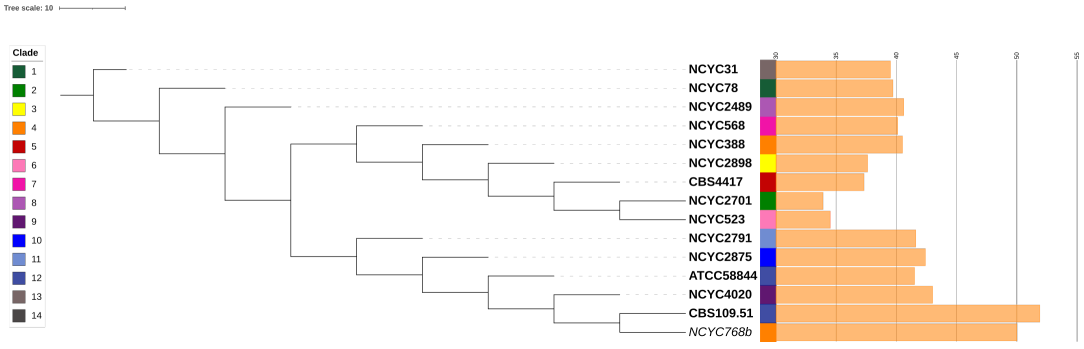
Alignment-free phylogenetic approaches have much potential in this era of big data. But as highlighted here, there are current issues with the FFP software. A key objective of this study was to investigate a previously hypothesised GC bias in the FFP four-letter DNA alphabet approach. This bias was clearly illustrated with a simulation study and as such this approach would not be recommended for future use. A further objective for this study which was successfully undertaken was to begin an effort to improve the current FFP approach by writing a new piece of alignment-free software. Finally, more investigations into alignment-free approaches and the development of unbiased and accurate trees built from these approaches is highly necessary.



(d) Original GC= 41.8%



(e) GC= 45%



(f) GC= 50%

Figure 6.2: FFP (four-letter DNA alphabet) phylogenetic trees of genes (without introns) from 15 Saccharomyces complex species annotated with GC content (orange bars). The GC content of the genic regions of NCYC768 was synonymously increased or decreased in all datasets except (d) (mutated sequence termed NCYC768b). All trees except (d) are consensus trees derived from 10 simulations. GC content of NCYC768 in all trees: (a) GC= 30%, (b) GC= 35%, (c) GC= 40%, (d) Original GC= 41.8%, (e) GC= 45% and (f) GC= 50%. Legend is Saccharomyces complex clade ordering.

Chapter 7

Discussion

7.1 Main goals

The ultimate goal of the project was to contribute to the computation of a yeast tree of life from whole genome sequences. The main aim of this project was to use a large, exemplar yeast dataset as a basis to compare state-of-the-art phylogenetic methodologies. Two further aims were to explore the genomic diversity within the dataset and to assess the composition of and phylogenetic signal within the core genome of a subset of the dataset. The key objectives of the project were: (1) to undertake stringent quality control of the 76 species draft genome assemblies for use in the project; (2) to assess the composition of and phylogenetic signal within the core genome of a subset of the dataset; (3) to assess the genomic similarities and differences within the dataset; (4) to compare the results of five different phylogenetic tree-building approaches on the full dataset; (5) to investigate a hypothesised GC bias in the FFP four-letter DNA alphabet approach through a simulation study and (6) to begin an effort to improve the current FFP approach by writing a new piece of alignment-free software, *jellyphy*.

7.2 Outcomes

Quality control

Generating the dataset for this project proved challenging and highlighted the importance of stringent quality control measures in genomic and phylogenetic studies. Many in-house sequenced and publicly available genome assemblies were, upon investigation, found to be misclassified or contaminated in the 76 species dataset. The

BLAST quality control step confirmed the identity of 75 out of the 76 chosen species, 28 of which were also confirmed with use of a custom Kraken database. Assembly quality was also assessed by statistics such as the N50 score, k-mer counts and BUSCO gene counts.

Core genome

The 41 species dataset in the core genome study was found to be highly diverse, as exemplified by the low proportions of sequence reads that mapped to a reference genome. Of the 15% of sequence reads of *Candida glabrata* strain CBS388 that mapped to the *Saccharomyces cerevisiae* S288c reference genome, a significant proportion were shown to map to the mitochondrial genome (40%) and to chromosome 12 (13.6%), likely a combination of a high proportion of conserved genes within these structures and multiple mitochondrial chromosomes within the sequenced sample. The mapped reads were assembled into contigs with two different assembly tools and the number of genes, gene names and chromosomal location were identified with BLAST, with only little variation observed between the results of the two assembly tools. The ortholog-finding BPGA pipeline identified a core set of genes from the amino acid sequences of 40 Saccharomyces complex species, at varying degrees of sequence identity, resulting in 591 genes at 50%, 82 genes at 75%, 38 genes at 80%, 19 genes at 85% and 5 genes at the 90% identity level. A BLAST process identified these genes and showed that the top 19 essential genes were involved in crucial processes such as protein synthesis, cell structure and metabolism. Phylogenetic trees were also built (FFP 20-letter amino acid alphabet) from these gene sets which showed clearly the effects different numbers of core genes have on phylogenetic accuracy. Perhaps surprisingly, only 82 core genes (75% sequence identity level) were required to produce the same tree topology as that achieved through use of the full proteome.

Comparative genomics of the Saccharomyces complex

For the goal of comparing the genomic information of strains from across the Saccharomyces complex, filtering out poor quality assemblies was crucial. The chosen filtering process resulted in a 58 species dataset which was nonetheless shown to be genomically diverse. A number of strains were found to be particularly unusual with larger or smaller than average genome sizes, gene counts, coding genome proportions and/or GC contents. In particular, global genome statistics were uncovered for ten

new draft genome assemblies (6 new species, 4 new strains) of relatively good quality.

Comparison of Phylogenetic methods

The main aim of this project was to compare phylogenetic methods using a key 76 species yeast dataset, which was achieved here. Whilst two strains unfortunately later proved to be misclassified and a number had poor quality assemblies, the study was still capable of showing the differences between the approaches. Of the five approaches tested, the BUSCO gene tree approach gave results most similar to the Kurtzman and Robnett tree topology, closely followed by the FFP 20-letter amino acid alphabet approach and more distantly by the SNP MLST, FFP 4-letter DNA alphabet and FFP 2-letter RY alphabet approaches. A sequence length bias was shown to affect the FFP 2-letter RY alphabet approach. Also, and most crucially, a potential GC bias was identified in the FFP 4-letter DNA alphabet approach.

GC simulation study

The goal of this study involved testing the hypothesis of a GC bias within the FFP 4-letter DNA alphabet approach, as identified in the previous chapter. A simulation study was conducted and entailed mutating the GC content of a single strain chosen from a 15-species dataset, followed by assessing the effects of the GC changes on the strain's true placement within the resulting phylogenetic tree. The hypothesised bias was clearly shown in this simulation. The study also showed a positive correlation between increasing coding genome GC content and GC content at the third codon position and a negative correlation at the other two codon positions. Finally, a new Python software tool named *jellyphy* was built in an effort to improve the current FFP approach, both in accuracy and computational efficiency, with the basic program completed at the end of this project.

7.3 Future directions

Quality control

Re-assembly of a number of poor quality genomes within the 76 species Saccharomyces complex dataset could be undertaken with different software and different k-mer options to find the best assembly for a genome. If this process still failed to produce

a good quality genome assembly, re-sequencing the strain or a different strain of the same species could be carried out. The k-mer Analysis Toolkit could also be used to take a closer look at the k-mer distributions of all sequence read datasets, with the potential to identify contaminated samples (Mapleson et al. (2017)). One interesting finding in this study was the high total number of k-mers found in the NCYC3345 genome relative to its genome assembly size. Further investigation into this genome could be made by generating a high quality sequence for the strain, perhaps using an alternative approach such as PacBio long-read sequencing. Species identification was key to dataset quality and depended on BLAST matching using species-specific 26S and 18S rDNA query templates. Assessing the ITS region could also be done, to add an additional layer of confidence to a species identification. The newly available Kraken fungal database could also be used to improve species identification and to look for the presence of dataset contamination.

Core genome

The percentage of sequence reads mapping to a reference genome was shown for six species only in this study. Assessing and comparing the percentages of sequence reads from the full species-confirmed 76 species dataset that mapped to a reference genome could show in more detail the diversity of the dataset. Adding more than one strain per species to the dataset could also be carried out to allow for the diversity within a species. A closer look at the putative gene names of the sequences identified by the different 'core' gene approaches (ABySS and Trinity RNA-Seq assembly or BPGA pipeline) could be taken to see if more genes were in common than recorded as a result of different gene aliases being found in the BLAST process. Repeating the BLAST annotation step to show more than the top hit may also show more similarities between the gene sets. One could also test different ortholog-identifying approaches to confirm or refute these findings. Finally, a closer look at non-coding core genes could be very interesting.

Comparative genomics of the Saccharomyces complex

This study could be repeated after re-sequencing a number of strains in the full 76 species dataset including those with poor assembly quality and replacing the two misclassified strains. This would give a clearer genomic picture of the Saccharomyces complex. Re-sequencing 'unusual' genomes could also be done to confirm or refute

these findings. One could also compare all strains present in this dataset and also in the Shen et al. (2018) dataset, not just the Eremothecium clade. Once again, the addition of more than one strain per species could give a more accurate description of species similarities and differences. Finer-scaled comparisons, such as the examination of genome organisation, patterns of heterozygosity and estimation of evolutionary rates, could also be made.

Comparison of Phylogenetic methods

Following the re-sequencing of all poor assemblies and replacement of misclassified strains, as mentioned previously, the five phylogenetic trees could be re-built using the new higher quality dataset. All good quality sequencing datasets used in a further recent study Shen et al. (2018) could also be added to the new dataset, increasing its scale and depth of diversity. To improve the SNP MLST tree approach, all publicly available genomes with no mitochondrial genome present could be removed from the dataset and the tree re-built. Another way to further confirm the differences between the five tree-building approaches is through a simulation study using a known tree topology, as was done for the FFP RY alphabet approach to show a sequence length bias. The sequence length bias could also be explored further by simulating more sequences to confirm the findings, as well as testing the different k-mer filtering options within the FFP software to see the effect of filtering out highly repetitive k-mers. The FFP amino acid alphabet approach was not investigated in detail here but, as mentioned in Chapter 6, was recently investigated by means of a simulation study (Li et al. (2020)). Repeating said simulation could be carried out to confirm the observed inaccuracy of this alphabet approach and to investigate further if this inaccuracy is the result of the distance measure used, as hypothesised in this study regarding the observed long branches. Testing other alignment-free approaches could also be carried out using resources such as the AF-Project, an online alignment-free benchmarking tool (Zielezinski et al. (2019)).

GC simulation study

The correlation shown between the coding GC content and GC at the different codon positions was made with the 76 species dataset. Ideally, this would be repeated with the high quality 58 species dataset or a new, higher quality 76 species dataset to give the most accurate indication of these relationships. Further development and testing

of the new *jellyphy* software built in this study could also be carried out, concentrating on identifying the most appropriate distance measure(s) to prevent long branches and GC bias.

7.4 Final conclusions

Working with biological datasets requires stringent quality control of sequencing data as shown in this study. Exploring a large, well-studied yeast species dataset, such as for the Saccharomyces complex, has shown how diverse these species really are. Higher quality datasets would show this in even further clarity. The identification of the composition of the core genome of a subset of this dataset can also contribute to our understanding regarding their evolutionary relationships to one other which in turn can also highlight interesting findings for both academia and industry. The affect that different numbers of core proteins had on phylogenetic tree topology was also shown in this study highlighting the need for deep consideration of which datasets to use for phylogenetic studies. The main aim of this project was successful in showing that different phylogenetic tree-building approaches do vary in their accuracy, computational intensity and ease of use. Whilst one widely-used alignment-free approach has been shown here to have biases which affect the accuracy of the phylogenetic trees estimated, the ability of these approaches to handle large NGS datasets with ease means they will undoubtedly continue to be used and hopefully improved upon. A comprehensive yeast tree of life will then be achievable.

Appendix

Appendix A

Strain choice and quality control

Clade	Strain ID	Species name	No. Contigs/ Scaffolds	N50	Unique	Distinct	Total	Max count
1	CR85	*Saccharomyces kudriavzevii*	17	862,320	9,111,325	10,218,185	11,849,084	1,097
1	NCYC2578	*Saccharomyces bayanus*	2,200	48,168	9,486,547	11,494,522	14,573,310	1,112
1	NCYC2888	*Saccharomyces mikatae*	368	157,774	8,893,876	10,126,202	12,038,882	752
1	NCYC2890	*Saccharomyces cariocanus*	448	141,541	8,975,553	10,152,787	11,997,839	1,063
1	NCYC3662	*Saccharomyces paradoxus*	858	66,736	8,997,244	10,161,065	11,892,750	998
1	NCYC392	*Saccharomyces pastorianus*	1,251	36,566	9,167,605	10,978,713	13,999,598	820
1	NCYC78	*Saccharomyces cerevisiae*	253	280,243	8,867,051	10,014,742	11,696,864	1,127
2	DBVPG7206	*Kazachstania turicensis*	17,386	3,678	9,503,995	11,419,665	15,213,463	730
2	NCYC1417	*Kazachstania lodderae*	1,417	31,719	8,071,402	9,656,957	12,238,488	1,217
2	NCYC2449	*Kazachstania telluris*	20,849	9,946	12,917,427	18,018,966	29,572,818	16,722
2	NCYC2450	*Candida humilis*	28,870	12,684	8,822,057	15,331,751	24,872,888	14,844
2	NCYC2483	*Kazachstania piceae*	4,386	43,101	8,131,888	9,946,609	13,564,493	1,710
2	NCYC2560	*Kazachstania sinensis*	2,182	47,906	9,269,435	10,271,284	11,975,833	523
2	NCYC2693	*Kazachstania servazzii*	2,126	38,946	6,305,597	9,424,674	16,241,391	2,498
2	NCYC2701	*Kazachstania viticola*	1,871	46,943	5,142,783	8,920,544	18,571,041	1,524
2	NCYC2702	*Kazachstania kunashirensis*	465	81,552	6,994,842	8,435,129	10,820,039	1,181
2	NCYC2703	*Kazachstania martiniae*	1,689	90,007	7,808,237	9,276,039	11,659,919	657
2	NCYC2729	*Kazachstania africana*	1,347	61,914	8,311,539	9,584,235	11,486,713	852
2	NCYC2827	*Kazachstania rosinii*	5,506	37,466	11,298,669	13,200,857	16,618,898	15,359
2	NCYC2878	*Kazachstania barnettii*	2,011	80,952	8,376,325	10,139,090	13,303,647	2,305
2	NCYC2991	*Kazachstania spencerorum*	2,032	89,048	7,606,866	9,298,107	12,142,093	837
2	NCYC3853	*Kazachstania bulderi*	13,228	5,335	7,158,278	11,653,113	21,275,729	3,163
2	NCYC814	*Kazachstania exigua*	8,120	10,883	12,050,896	16,834,754	26,320,355	5,099
2	NRRLY1556	*Kazachstania unispora*	585	159,570	7,358,715	9,154,918	12,315,571	1,785

Clade	Strain ID	Species name	No. Contigs/ Scaffolds	N50	Unique	Distinct	Total	Max count
2	NRRLY17245	*Kazachstania transvaalensis*	914	133,191	7,813,133	9,539,204	12,587,366	1,758
3	CBS421	*Naumovozyma dairenensis*	11	109,578	8,702,675	10,524,618	13,480,773	1,933
3	NCYC2898	*Naumovozyma castellii*	642	103,769	8,155,406	9,466,695	11,334,680	417
4	CBS4332	*Candida castellii*	1,242	13,774	8,593,626	9,367,101	10,444,390	513
4	CBS7729	*Nakaseomyces bacillisporus*	1,383	301,794	8,241,932	9,409,541	11,318,515	2,472
4	NCYC388	*Candida glabrata*	409	357,606	9,467,586	10,732,698	12,529,271	279
4	NCYC768	*Nakaseomyes delphensis*	217	211,289	8,500,727	9,503,255	11,037,963	1,976
5	CBS4417	*Tetrapisispora phaffii*	2,327	48,426	13,466,370	17,225,119	23,805,745	2,085
5	CBS6284	*Tetrapisispora blattae*	10	286,437	8,215,961	10,223,181	14,043,767	2,454
5	CBS8762	*Tetrapisispora arboricola*	878	141,586	8,526,775	9,757,303	11,701,420	2,321
5	CBS8763	*Tetrapisispora nanseiensis*	295	117,592	8,696,582	9,616,149	10,904,516	1,871
5	NRRLY27309	*Tetrapisispora iriomotensis*	576	317,713	7,061,776	8,982,184	12,048,828	2,180
6	NCYC2754	*Vanderwaltozyma yarrowii*	7,287	7,727	9,286,235	11,853,216	16,826,208	14,566
6	NCYC523	*Vanderwaltozyma polyspora*	1,169	186,992	8,575,129	10,693,737	14,566,801	1,997
7	NCYC1495	*Zygosaccharomyces bisporus*	1,330	18,213	8,830,598	9,556,744	10,559,358	1,479
7	NCYC2403	*Zygosaccharomyces mellis*	5,393	73,822	10,262,372	11,836,033	14,542,960	24,468
7	NCYC2789	*Zygosaccharomyces lentus*	9,002	97,590	11,887,068	13,699,305	16,818,280	14,303
7	NCYC3000	*Zygosaccharomyces kombuchaensis*	706	61,715	8,623,453	9,382,987	10,439,122	503
7	NCYC568	*Zygosaccharomyces rouxii*	310	1,059,696	7,852,437	8,750,058	10,028,051	1,252
7	NCYC573	*Zygosaccharomyces bailii*	729	44,422	8,824,799	9,556,220	10,483,053	315
8	NCYC2489	*Zygotorulaspora mrakii*	293	363,292	8,327,014	9,227,640	10,533,246	244
8	NCYC2513	*Zygotorulaspora florentinus*	914	196,060	8,327,014	9,227,640	10,533,246	244
9	NCYC4020	*Torulaspora delbrueckii*	523	230,122	7,825,051	8,473,609	9,297,279	145
9	NCYC524	*Torulaspora pretoriensis*	947	207,604	8,342,082	8,990,032	9,885,493	145

Clade	Strain ID	Species name	No. Contigs/ Scaffolds	N50	Unique	Distinct	Total	Max count
9	NCYC820	*Torulaspora globosa*	6,105	528,175	9,653,107	10,720,846	12,277,428	5,075
9	NRRLY1549	*Torulaspora microellipsoides*	302	300,364	7,867,837	9,224,995	10,974,686	520
9	NRRLY17532	*Torulaspora franciscae*	306	434,022	6,899,057	8,505,650	10,504,247	265
10	CBS6340	*Lachancea thermotolerans*	8	1,513,537	8,938,425	9,600,775	10,390,009	167
10	NCYC2508	*Lachancea fermentati*	911	136,750	8,571,575	9,318,150	10,334,712	107
10	NCYC2644	*Lachancea waltii*	669	113,908	8,867,425	9,560,785	10,493,537	215
10	NCYC2875	*Lachancea cidri*	2,521	23,555	5,100,431	10,005,345	21,252,786	430
10	NCYC543	*Lachancea kluyveri*	1,224	28,713	9,099,220	10,071,349	11,502,526	1,015
11	CBS4438	*Kluyveromyces aestuarii*	356	310,612	8,029,541	8,921,945	10,142,766	550
11	CBS8778	*Kluyveromyces nonfermentans*	345	193,381	7,357,314	8,273,479	9,580,508	1,773
11	NCYC2559	*Kluyveromyces dobzhanskii*	1,052	29,646	8,860,617	9,705,379	10,828,073	440
11	NCYC2791	*Kluyveromyces marxianus*	4,435	9,917	6,130,630	9,946,254	18,120,202	1,762
11	NCYC416	*Kluyveromyces lactis*	186	918,166	8,259,977	9,512,758	11,511,908	729
11	UCD54210	*Kluyveromyces wickerhamii*	510	3,6691	8,257,843	8,968,397	9,801,114	253
12	ATCC58844	*Eremothecium sinecaudum*	7	1,398,029	7,447,793	8,006,750	8,922,432	640
12	CBS106.43	*Eremothecium ashbyi*	14,369	2,328	7,743,331	8,813,181	10,961,273	646
12	CBS109.51	*Eremothecium gossypii*	148	303,511	7,739,299	8,210,476	8,778,285	164
12	DBVPG7215	*Eremothecium cymbalariae*	8	1,193,613	8,122,269	8,814,960	9,669,294	897
12	NCYC1563	*Eremothecium coryli*	5506	17,267	9,369,470	10,393,468	11,913,944	5,814
13	AWRI3580	*Hanseniaspora uvarum*	18	1,289,090	5,396,996	6,398,461	7,948,686	3,594
13	CBS2592	*Hanseniaspora occidentalis*	6,625	33,465	9,273,153	11,253,397	14,991,214	40,439
13	CBS285	*Hanseniaspora lindneri*	6,228	18,587	8,679,343	10,205,470	12,607,772	15,187
13	NCYC31	*Hanseniaspora osmophila*	486	239,914	7,813,034	9,142,747	11,434,085	8,046
13	NCYC36	*Hanseniaspora vineae*	951	199,177	7,774,202	9,298,170	11,995,032	4,313

Clade	Strain ID	Species name	No. Contigs/ Scaffolds	N50	Unique	Distinct	Total	Max count
13	NCYC4006	*Hanseniaspora valbyensis*	13,597	2,667	6,347,499	8,690,413	13,720,729	21,406
13	UTAD222	*Hanseniaspora guilliermondii*	208	91,417	5,793,935	6,981,604	9,029,085	18,806
14	NCYC3345	*Saccharomycodes ludwigii*	7,291	3,746	5,569,777	11,956,812	27,391,345	1,472
Outgroup	NCYC18	*Wickerhamomyces anomalus*	2,386	28,699	4,122,218	10,351,778	28,384,059	1,378

Table A.1: Quality control information of draft gene assemblies for 75 Saccharomyces complex species and outgroup. Number of contigs, N50 scores and k-mer statistics (Unique: the number of k-mers occuring exactly once; Distinct: the number of k-mers, ignoring their multiplicity; Total: the number of k-mers with multiplicity; Max count: the maximum of the number of occurrences).

Species name	26S Accession	18S Accession
Saccharomyces kudriavzevii	AB040995.1	NG_064874.1
Saccharomyces bayanus	AF113892.1	AY046227.1
Saccharomyces mikatae	AB040996.1	NG_064875.1
Saccharomyces cariocanus	AF399761	AY046224
Saccharomyces paradoxus	EU669466.1	NG_063106.1
Saccharomyces pastorianus	AF113893.1	X97805.1
Saccharomyces cerevisiae	EU884435.1	NR_132213.1
Kazachstania turicensis	NG_058312.1	AB086237.1
Kazachstania lodderae	AY048161	X83824
Kazachstania telluris	U72158.1	AY046236.1
Candida humilis	U69878.1	AB054678.1
Kazachstania piceae	AF399767	AY046233
Kazachstania sinensis	FJ527186	AY046238
Kazachstania servazzii	KM454442.1	NG_064888.1
Kazachstania viticola	AF398482	AY046234
Kazachstania kunashirensis	AF399769	AY046235
Kazachstania martiniae	AF399766	AY046231
Kazachstania africana	NG_055030.1	NG_063237.1
Kazachstania rosinii	KY107943.1	KY103665.1*
Kazachstania barnettii	AJ508590	AY046242
Kazachstania spencerorum	AY048162	X97807
Kazachstania bulderi	AF398486	AY046241
Kazachstania exigua	FJ153135.1	AY007905.1
Kazachstania unispora	NG_055027.1	NG_063236.1
Kazachstania transvaalensis	NG_055026.1	NG_063238.1
Naumovozyma dairenensis	Z75579.1	JQ689019.1
Naumovozyma castellii	AY007888.1	XR_002431960.1
Candida castellii	NG_055072.1	NG_063514.1
Nakaseomyces bacillisporus	NG_055071.1	NG_063246.1
Candida glabrata	JQ070154.1	AY083230.1
Nakaseomyes delphensis	U69576.1	X83823.1
Tetrapisispora phaffii	NG_055035.1	NG_063245.1
Tetrapisispora blattae	NG_055034.1	NG_063244.1
Tetrapisispora arboricola	NG_058391	NG_065576
Tetrapisispora nanseiensis	EF460662.1	AB016509.1
Tetrapisispora iriomotensis	AF399781.1	NG_064805.1
Vanderwaltozyma yarrowii	D83441.1	AB054674.1
Vanderwaltozyma polyspora	EF460663.1	JQ698890.1
Zygosaccharomyces bisporus	U72162	X91084
Zygosaccharomyces mellis	AB302837.1	AF339891.1
Zygosaccharomyces lentus	AF399792.1	Y16814.1

Zygosaccharomyces kombuchaensis	AF339904	AF339890
Zygosaccharomyces rouxii	KF002711.1	NG_065155.1
Zygosaccharomyces bailii	DQ872869.1	NG_065158.1
Zygotorulaspora mrakii	U72159	X90757
Zygotorulaspora florentinus	AF399774	X91086
Torulaspora delbrueckii	KM434245.1	NG_061300.1
Torulaspora pretoriensis	U72157	X84638
Torulaspora globosa	AF399782.1	X84639.1
Torulaspora microellipsoides	NG_055074.1	NG_062443.1
Torulaspora franciscae	U73604.1	NG_063354.1
Lachancea thermotolerans	NG_042626.1	NG_061071.1
Lachancea fermentati	NG_055076.1	NG_062442.1
Lachancea waltii	U69582.1	D83422.1
Lachancea cidri	U84236	X91085
Lachancea kluyveri	NG_055066.1	NG_062650.1
Kluyveromyces aestuarii	NG_055069.1	X89520.1
Kluyveromyces nonfermentans	NG_058314.1	AB012264.1
Kluyveromyces dobzhanskii	NG_055067.1	D83430.1
Kluyveromyces marxianus	NG_042627.1	NG_062653.1
Kluyveromyces lactis	U94922.1	AB054673.1
Kluyveromyces wickerhamii	NG_055068.1	NG_063255.1
Eremothecium sinecaudum	NG_055065.1	NG_063258.1
Eremothecium ashbyi	AB294409.1	AY046269.1
Eremothecium gossypii	NG_063967.1	AY046265.1
Eremothecium cymbalariae	NG_042628.1	NG_063259.1
Eremothecium coryli	NG_055064.1	NG_063257.1
Hanseniaspora uvarum	NG_055419.1	NG_063250.1
Hanseniaspora occidentalis	NG_055416.1	NG_063253.1
Hanseniaspora lindneri	NG_055417.1	NG_063248.1
Hanseniaspora osmophila	U84228.1	AY046259.1
Hanseniaspora vineae	NG_055415.1	NG_063251.1
Hanseniaspora valbyensis	U73596.1	NG_063247.1
Hanseniaspora guilliermondii	DQ872868.1	NG_063249.1
Saccharomycodes ludwigii	U73601.1	NG_063254.1
Wickerhamomyces anomalus	JN562717	DQ520880

Table A.3: Database accession IDs for all 26S/28S and 18S rRNA gene sequences for the 76 species analysed in this study. The sequences were used in a BLAST analysis to confirm species identities of draft genome assemblies of strains believed to derive from these species. (*==ITS-5.8 as 18S unavailable)

Kraken database species	Accession number	Strain ID	Kraken QC (%)
Candida castellii	GCA_001046935.1		
Candida glabrata	GCF_000002545.3	NCYC388	97.50
Eremothecium gossypii	GCF_000091025.4	CBS109.51	97.14
Hanseniaspora osmophila	GCA_001747045.1	NCYC31	94.77
Hanseniaspora uvarum	GCA_001747055.1		
Hanseniaspora vineae	GCA_002141495.1	NCYC36	95.68
Kazachstania africana	GCF_000304475.1	NCYC2729	85.67
Kazachstania servazzii	GCA_002214935.1	NCYC2693	96.84
Kluyveromyces aestuarii	GCA_000179355.1	CBS4438	94.79
Kluyveromyces dobzhanskii	GCA_000820885.1	NCYC2559	93.47
Kluyveromyces lactis	GCF_000002515.2	NCYC416	96.39
Kluyveromyces marxianus	GCA_001417885.1	NCYC2791	93.95
Lachancea fermentati	GCA_900074765.1	NCYC2508	95.17
Lachancea kluyveri	GCA_000149225.1	NCYC543	93.33
Lachancea thermotolerans	GCF_000142805.1		
Lachancea waltii	GCA_000167115.1	NCYC2644	95.06
Nakaseomyes delphensis	GCA_001039675.1	NCYC768	97.15
Naumovozyma castellii	GCA_000237345.1	NCYC2898	94.08
Saccharomyces bayanus	GCA_000167035.1		
Saccharomyces cerevisiae	GCA_000146045.2	NCYC78	86.38
Saccharomyces mikatae	GCA_000166975.1	NCYC2888	78.55
Torulaspora delbrueckii	GCA_000243375.1	NCYC4020	86.58
Torulaspora franciscae	GCA_003705175.2		
Torulaspora pretoriensis	GCA_003706005.1	NCYC524	73.50
Vanderwaltozyma polyspora	GCF_000150035.1	NCYC523	92.82
Wickerhamomyces anomalus	GCF_001661255.1	NCYC18	82.37
Zygosaccharomyces kombuchaensis	GCA_00370595.1	NCYC3000	97.00
Zygosaccharomyces rouxii	GCA_000026365.1	NCYC568	77.14
Zygotorulaspora florentinus	GCA_003671575.2	NCYC2513	97.17
Zygotorulaspora mrakii	GCA_00367156.1	NCYC2489	96.49

Table A.2: Kraken database of 30 publicly available genomes generated for this project. Species names and GenBank/RefSeq accessions are given for each genome used, along with accession IDs of the 25 strains (of the same species) that were identified using this database, along with the percentage of k-mers that matched to the database. The species designations of five genomes used within the database were not included in the 76 species dataset and therefore no matches were found.

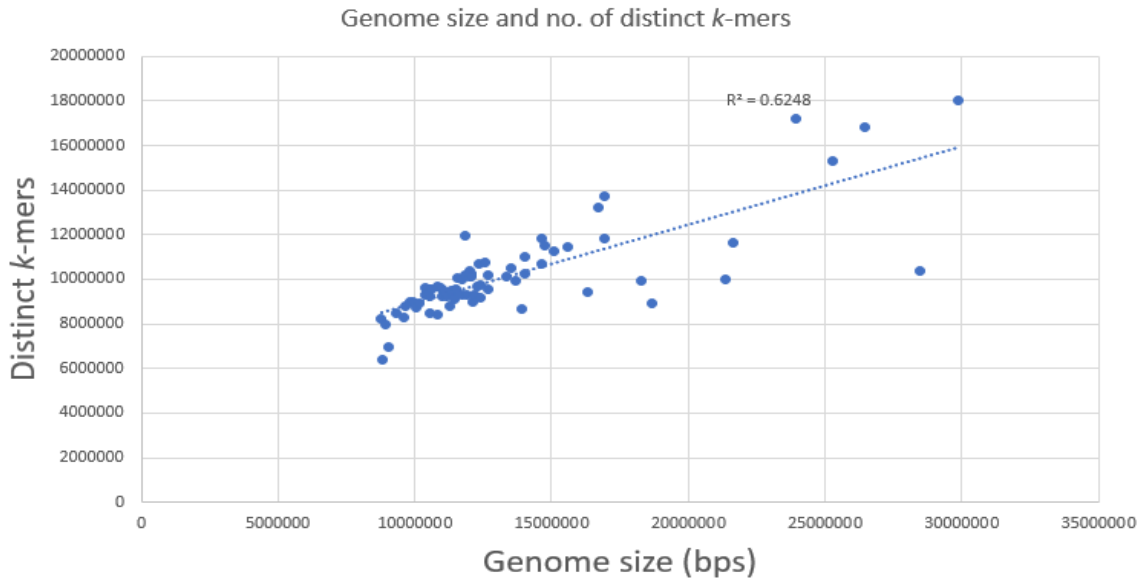


Figure A.1: The relationship between Jellyfish-derived distinct k-mers and genome size for 75 Saccharomyces complex genome assemblies plus an outlier species.

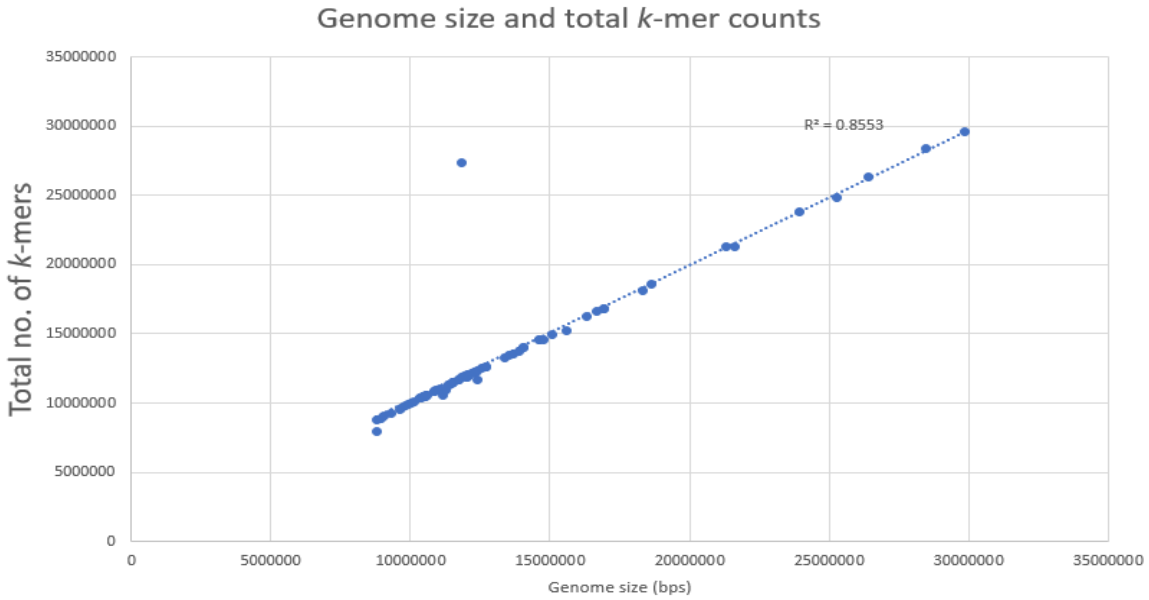


Figure A.2: The relationship between Jellyfish-derived total k-mers and genome size for 75 Saccharomyces complex genome assemblies plus an outlier species. The point lying far from the straight line is NCYC3345 (*Saccharomycodes ludwigii*).

Appendix B

Core genome strain choice

Clade	NCYC ID	Species name
	2600	*Saccharomyces paradoxus*
	2888	*Saccharomyces mikatae*
	2890	*Saccharomyces cariocanus*
	392	*Saccharomyces pastorianus*
	2578	*Saccharomyces bayanus*
1	78	*Saccharomyces cerevisiae*
	2560	*Kazachstania sinensis*
	2729	*Kazachstania africana*
	2701	*Kazachstania viticola*
	2703	*Kazachstania martiniae*
	2991	*Kazachstania spencerorum*
	1417	*Kazachstania lodderae*
	2483	*Kazachstania piceae*
	2702	*Kazachstania kunashirensis*
	3853	*Kazachstania bulderi*
	2878	*Kazachstania barnettii*
2	2693	*Kazachstania servazzii*
3	2898	*Naumovozyma castellii*
	768	*Nakaseomyes delphensis*
4	388	*Candida glabrata*
6	523	*Vanderwaltozyma polyspora*
	568	*Zygosaccharomyces rouxii*
	573	*Zygosaccharomyces bailii*
	3000	*Zygosaccharomyces kombuchaensis*
7	1495	*Zygosaccharomyces bisporus*
	2513	*Zygotorulaspora florentinus*
	2489	*Zygotorulaspora mrakii*
	524	*Torulaspora pretoriensis*
	4020	*Torulaspora delbrueckii*
	2875	*Lachancea cidri*
	2508	*Lachancea fermentati*
	2433	*Lachancea thermotolerans*
	543	*Lachancea kluyveri*
10	2644	*Lachancea waltii*
	2791	*Kluyveromyces marxianus*
	2559	*Kluyveromyces dobzhanskii*
11	416	*Kluyveromyces lactis*
	2739	*Hanseniaspora uvarum*
	36	*Hanseniaspora vineae*
13	31	*Hanseniaspora osmophila*
Outgroup	18	*Wickerhamomyces anomalus*

Table B.1: Forty NCYC Saccharomyces complex species and outgroup included in the core genome analysis.

Appendix C

Comparative Genomics

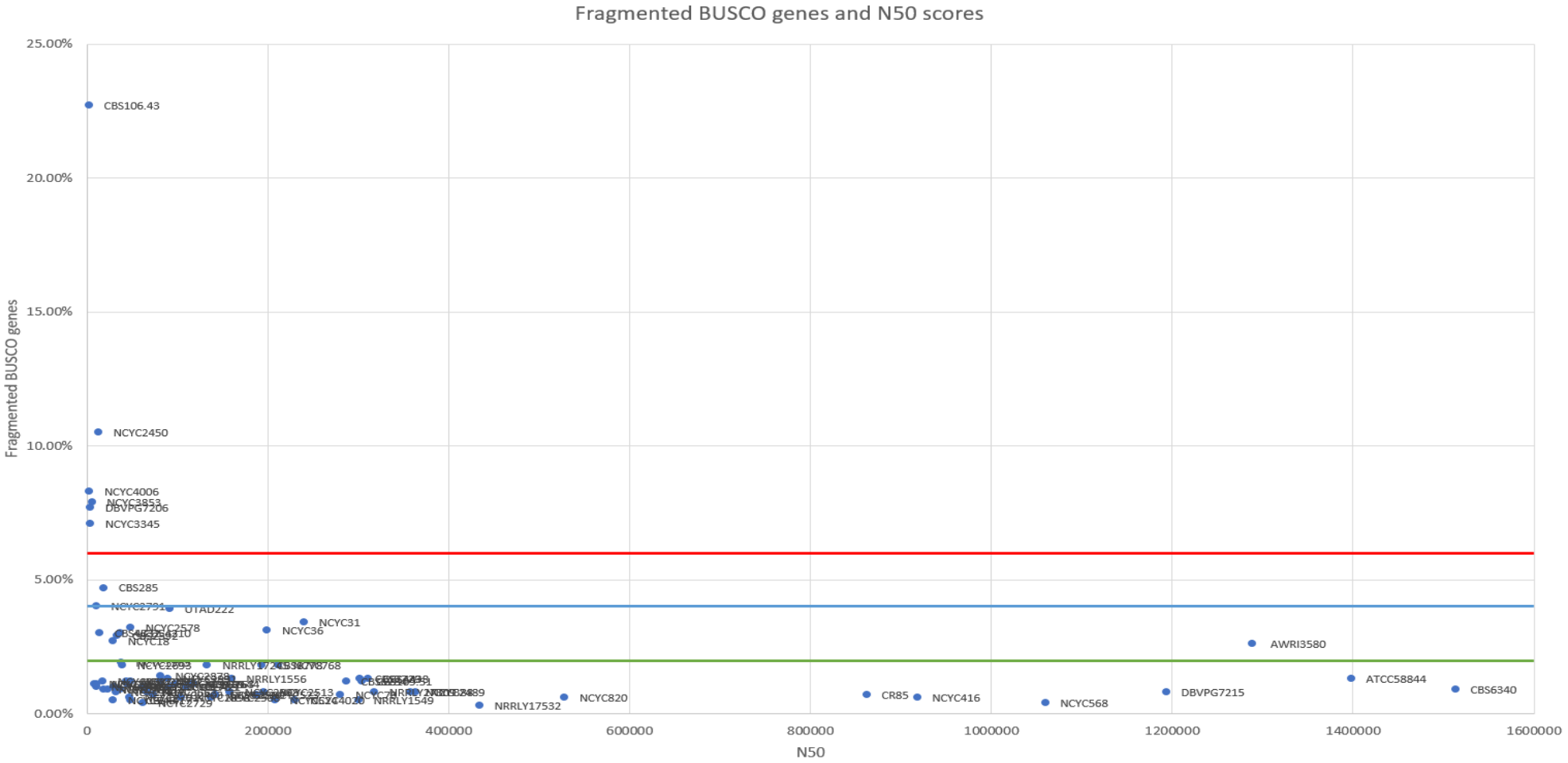


Figure C.1: N50 and percentage of fragmented BUSCO genes in the 75 Saccharomyces complex species and outgroup. The red, blue and green lines denote strains with more than 6%, 4% and 2% fragmeneted BUSCO genes respectively.

Clade	Strain ID	Species name	N50	Genome size(bp)	Coding (%)	Genes	GC(%)	GC3
1	CR85	*Saccharomyces kudriavzevii*	862,320	11,852,480	69.99	5,433	39.62	33.55
1	NCYC2888	*Saccharomyces mikatae*	157,774	12,057,322	69.39	5,540	37.74	31.73
1	NCYC2890	*Saccharomyces cariocanus*	141,541	12,017,915	69.58	5,526	38.22	32.17
1	NCYC3662	*Saccharomyces paradoxus*	66,736	12,051,421	66.76	4,986	37.48	33.21
1	NCYC392	*Saccharomyces pastorianus*	36,566	14,039,592	69.82	6,682	40.06	34.14
1	NCYC78	*Saccharomyces cerevisiae*	280,243	11,707,993	70.76	5,438	38.11	31.98
2	NCYC1417	*Kazachstania lodderae*	31,719	12,277,384	68.85	5,551	33.62	26.62
2	NCYC2449	*Kazachstania telluris*	9,946	29,850,499	57.34	11,243	35.61	25.25
2	NCYC2483	*Kazachstania piceae*	43101	13,673,105	61.59	5,529	32.64	26.99
2	NCYC2560	*Kazachstania sinensis*	47,906	12,035,436	70.17	5,576	46.10	38.47
2	NCYC2693	*Kazachstania servazzii*	38,946	16,306,775	68.36	7,563	34.40	28.30
2	NCYC2701	*Kazachstania viticola*	46,943	18,649,725	69.20	8,582	32.64	25.76
2	NCYC2702	*Kazachstania kunashirensis*	81,552	10,835,154	72.30	5,196	32.36	24.92
2	NCYC2703	*Kazachstania martiniae*	90,007	11,702,768	69.63	5,651	33.77	27.30
2	NCYC2729	*Kazachstania africana*	61,914	11,523,375	68.42	5,443	36.18	29.58
2	NCYC2827	*Kazachstania rosinii*	17,267	16,698,986	59.45	6,856	49.68	36.29
2	NCYC2878	*Kazachstania barnettii*	80,952	13,354,992	61.70	5,454	33.53	27.13
2	NCYC2991	*Kazachstania spencerorum*	89,048	12,192,244	67.91	5,654	33.41	25.82
2	NCYC814	*Kazachstania exigua*	10,883	26,437,421	57.47	10,380	34.92	24.43
2	NRRLY1556	*Kazachstania unispora*	159,570	12,384,023	67.21	5,485	32.14	24.95
2	NRRLY17245	*Kazachstania transvaalensis*	133,191	12,699,196	65.57	5,496	33.12	28.23
3	CBS421	*Naumovozyma dairenensis*	109,578	13,528,176	63.66	5,525	34.04	26.46
3	NCYC2898	*Naumovozyma castellii*	103,769	11,352,650	74.01	5,622	36.75	29.64
4	CBS7729	*Nakaseomyces bacillisporus*	301,794	11,358,607	69.03	5,181	36.81	30.89
4	NCYC388	*Candida glabrata*	357,606	12,545,987	63.20	5,191	38.58	32.59

Clade	Strain ID	Species name	N50	Genome size(bp)	Coding (%)	Genes	GC(%)	GC3
4	NCYC768	*Nakaseomyes delphensis*	211,289	11,050,768	69.25	4,986	38.84	34.23
5	CBS4417	*Tetrapisispora phaffii*	48,426	23,960,103	65.59	10,479	35.04	30.35
5	CBS6284	*Tetrapisispora blattae*	286,437	14,049,134	61.74	5,281	31.73	25.34
5	CBS8762	*Tetrapisispora arboricola*	141,586	12,385,752	64.74	5,248	33.52	27.85
5	NRRLY27309	*Tetrapisispora iriomotensis*	317,713	12,108,829	69.27	5,749	32.40	26.79
6	NCYC2754	*Vanderwaltozyma yarrowii*	7,727	16,928,921	54.50	6,716	40.43	24.85
6	NCYC523	*Vanderwaltozyma polyspora*	186,992	14,612,542	57.28	5,533	32.67	26.43
7	NCYC1495	*Zygosaccharomyces bisporus*	18,213	10,604,821	70.41	4,988	43.94	35.61
7	NCYC2403	*Zygosaccharomyces mellis*	73,822	14,623,258	59.62	6,208	48.77	31.71
7	NCYC2789	*Zygosaccharomyces lentus*	97,590	16,943,801	55.07	7,399	51.73	37.92
7	NCYC3000	*Zygosaccharomyces kombuchaensis*	61,715	10,460,019	71.49	5,010	44.52	36.87
7	NCYC568	*Zygosaccharomyces rouxii*	1,059,696	10,040,848	74.64	4,988	39.07	29.44
7	NCYC573	*Zygosaccharomyces bailii*	44,422	10,504,610	70.39	4,913	42.43	34.00
8	NCYC2489	*Zygotorulaspora mrakii*	363,292	10,542,908	71.57	4,954	39.94	31.59
8	NCYC2513	*Zygotorulaspora florentinus*	196,060	11,162,755	68.74	5,132	40.92	33.21
9	NCYC4020	*Torulaspora delbrueckii*	230,122	9,314,331	77.33	4,870	41.95	33.56
9	NCYC524	*Torulaspora pretoriensis*	207,604	9,913,881	73.62	4,916	44.26	36.23
9	NCYC820	*Torulaspora globosa*	528,175	12,360,492	62.51	5,662	51.96	36.92
9	NRRLY1549	*Torulaspora microellipsoides*	300,364	11,006,487	73.27	5,379	38.63	30.34
9	NRRLY17532	*Torulaspora franciscae*	434,022	10,534,808	76.86	5,457	45.04	36.07
10	CBS6340	*Lachancea thermotolerans*	1,513,537	10,393,318	71.32	4,829	47.28	36.68
10	NCYC2508	*Lachancea fermentati*	136,750	10,358,648	73.72	5,078	42.51	33.56
10	NCYC2644	*Lachancea waltii*	113,908	10,512,120	70.64	4,946	44.35	34.92
10	NCYC2875	*Lachancea cidri*	23,555	21,335,293	74.73	11,001	41.22	32.81
10	NCYC543	*Lachancea kluyveri*	28,713	11,543,969	69.53	5,396	41.35	34.63

Clade	Strain ID	Species name	N50	Genome size(bp)	Coding (%)	Genes	GC(%)	GC3
11	CBS4438	*Kluyveromyces aestuarii*	310,612	10,153,782	71.26	4,866	38.31	31.02
11	CBS8778	*Kluyveromyces nonfermentans*	193,381	9,595,169	71.77	4,603	35.94	28.95
11	NCYC2559	*Kluyveromyces dobzhanskii*	29,646	10,855,115	68.35	4,999	41.37	34.64
11	NCYC416	*Kluyveromyces lactis*	918,166	11,519,790	68.89	5,376	38.77	31.96
12	ATCC58844	*Eremothecium sinecaudum*	1,398,029	8,922,988	73.81	4,299	40.20	31.79
12	CBS109.51	*Eremothecium gossypii*	303,511	8,783,618	74.02	4,131	51.81	39.06
12	DBVPG7215	*Eremothecium cymbalariae*	1,193,613	9,669,912	71.31	4,531	40.31	32.22
12	NCYC1563	*Eremothecium coryli*	37,466	11,987,794	61.36	5,233	48.70	32.20
Averages:			**243,846**	**13,053,764**	**66.83**	**5,885**	**39.35**	**31.25**

Table C.1: Genome statistics of 58 Saccharomyces complex species.

Clade	Strain ID	Species name	N50	Complete	Single-copy	Duplicated	Fragmented	Missing
1	CR85	*Saccharomyces kudriavzevii*	862,320	98.0%	97.4%	0.6%	0.7%	1.3%
1	NCYC2888	*Saccharomyces mikatae*	157,774	98.2%	97.7%	0.5%	0.8%	1.0%
1	NCYC2890	*Saccharomyces cariocanus*	141,541	98.1%	97.5%	0.6%	0.7%	1.2%
1	NCYC3662	*Saccharomyces paradoxus*	66,736	97.7%	97.1%	0.6%	0.8%	1.5%
1	NCYC392	*Saccharomyces pastorianus*	36,566	98.1%	86.6%	11.5%	0.9%	1.0%
1	NCYC78	*Saccharomyces cerevisiae*	280,243	98.3%	97.7%	0.6%	0.7%	1.0%
2	NCYC1417	*Kazachstania lodderae*	31,719	97.9%	96.8%	1.1%	0.8%	1.3%
2	NCYC2449	*Kazachstania telluris*	9,946	97.9%	59.2%	38.7%	1.0%	1.1%
2	NCYC2483	*Kazachstania piceae*	43,101	97.4%	96.5%	0.9%	1.2%	1.4%
2	NCYC2560	*Kazachstania sinensis*	47,906	97.0%	96.2%	0.8%	1.2%	1.8%
2	NCYC2693	*Kazachstania servazzii*	38,946	95.7%	74.8%	20.9%	1.8%	2.5%
2	NCYC2701	*Kazachstania viticola*	46,943	96.9%	66.6%	30.3%	0.6%	2.5%
2	NCYC2702	*Kazachstania kunashirensis*	81,552	96.8%	95.4%	1.4%	1.2%	2.0%
2	NCYC2703	*Kazachstania martiniae*	90,007	97.4%	96.5%	0.9%	1.0%	1.6%
2	NCYC2729	*Kazachstania africana*	61,914	98.6%	98.1%	0.5%	0.4%	1.0%
2	NCYC2827	*Kazachstania rosinii*	17,267	97.3%	96.4%	0.9%	1.2%	1.5%
2	NCYC2878	*Kazachstania barnettii*	80,952	96.6%	95.5%	1.1%	1.4%	2.0%
2	NCYC2991	*Kazachstania spencerorum*	89,048	96.9%	95.8%	1.1%	1.3%	1.8%
2	NCYC814	*Kazachstania exigua*	10,883	97.2%	24.6%	72.6%	1.1%	1.7%
2	NRRLY1556	*Kazachstania unispora*	159,570	95.8%	95.1%	0.7%	1.3%	2.9%
2	NRRLY17245	*Kazachstania transvaalensis*	133,191	91.8%	91.4%	0.4%	1.8%	6.4%
3	CBS421	*Naumovozyma dairenensis*	109,578	97.8%	97.1%	0.7%	1.0%	1.2%
3	NCYC2898	*Naumovozyma castellii*	103,769	98.5%	96.3%	2.2%	0.6%	0.9%
4	CBS7729	*Nakaseomyces bacillisporus*	301,794	95.3%	94.6%	0.7%	1.3%	3.4%
4	NCYC388	*Candida glabrata*	357,606	97.9%	97.4%	0.5%	0.8%	1.3%

Clade	Strain ID	Species name	N50	Complete	Single-copy	Duplicated	Fragmented	Missing
4	NCYC768	*Nakaseomyes delphensis*	211,289	96.2%	95.4%	0.8%	1.8%	2.0%
5	CBS4417	*Tetrapisispora phaffii*	48,426	99.1%	22.0%	77.1%	0.5%	0.4%
5	CBS6284	*Tetrapisispora blattae*	286,437	96.2%	94.7%	1.5%	1.2%	2.6%
5	CBS8762	*Tetrapisispora arboricola*	141,586	98.0%	96.7%	1.3%	0.7%	1.3%
5	NRRLY27309	*Tetrapisispora iriomotensis*	317,713	97.6%	96.5%	1.1%	0.8%	1.6%
6	NCYC2754	*Vanderwaltozyma yarrowii*	7,727	96.7%	94.0%	2.7%	1.1%	2.2%
6	NCYC523	*Vanderwaltozyma polyspora*	186,992	98.4%	95.9%	2.5%	0.7%	0.9%
7	NCYC1495	*Zygosaccharomyces bisporus*	18,213	97.9%	97.7%	0.2%	0.9%	1.2%
7	NCYC2403	*Zygosaccharomyces mellis*	73,822	98.3%	98.1%	0.2%	0.7%	1.0%
7	NCYC2789	*Zygosaccharomyces lentus*	97,590	97.7%	97.3%	0.0%	1.1%	1.2%
7	NCYC3000	*Zygosaccharomyces kombuchaensis*	61,715	98.1%	97.7%	0.4%	1.0%	0.9%
7	NCYC568	*Zygosaccharomyces rouxii*	1,059,696	98.7%	98.5%	0.2%	0.4%	0.9%
7	NCYC573	*Zygosaccharomyces bailii*	44,422	97.7%	97.3%	0.4%	1.1%	1.2%
8	NCYC2489	*Zygotorulaspora mrakii*	363,292	97.2%	97.1%	0.1%	0.8%	2.0%
8	NCYC2513	*Zygotorulaspora florentinus*	196,060	97.9%	97.7%	0.2%	0.8%	1.3%
9	NCYC4020	*Torulaspora delbrueckii*	230,122	98.5%	98.3%	0.2%	0.5%	1.0%
9	NCYC524	*Torulaspora pretoriensis*	207,604	98.3%	98.2%	0.1%	0.5%	1.2%
9	NCYC820	*Torulaspora globosa*	528,175	98.0%	97.9%	0.1%	0.6%	1.4%
9	NRRLY1549	*Torulaspora microellipsoides*	300,364	98.2%	97.7%	0.5%	0.5%	1.3%
9	NRRLY17532	*Torulaspora franciscae*	434,022	98.2%	97.6%	0.6%	0.3%	1.5%
10	CBS6340	*Lachancea thermotolerans*	1,513,537	98.0%	97.7%	0.3%	0.9%	1.1%
10	NCYC2508	*Lachancea fermentati*	136,750	98.3%	98.1%	0.2%	0.6%	1.1%
10	NCYC2644	*Lachancea waltii*	113,908	97.4%	97.2%	0.2%	1.1%	1.5%
10	NCYC2875	*Lachancea cidri*	23,555	97.8%	48.4%	49.4%	0.9%	1.3%
10	NCYC543	*Lachancea kluyveri*	28,713	98.6%	98.5%	0.1%	0.5%	0.9%

Clade	Strain ID	Species name	N50	Complete	Single-copy	Duplicated	Fragmented	Missing
11	CBS4438	*Kluyveromyces aestuarii*	310,612	96.6%	96.4%	0.2%	1.3%	2.1%
11	CBS8778	*Kluyveromyces nonfermentans*	193,381	95.3%	95.0%	0.3%	1.8%	2.9%
11	NCYC2559	*Kluyveromyces dobzhanskii*	29,646	97.0%	96.8%	0.2%	1.0%	2.0%
11	NCYC416	*Kluyveromyces lactis*	918,166	97.7%	94.3%	3.4%	0.6%	1.7%
12	ATCC58844	*Eremothecium sinecaudum*	1,398,029	94.5%	94.3%	0.2%	1.3%	4.2%
12	CBS109.51	*Eremothecium gossypii*	303,511	96.4%	96.1%	0.3%	1.2%	2.4%
12	DBVPG7215	*Eremothecium cymbalariae*	1,193,613	96.9%	96.8%	0.1%	0.8%	2.3%
12	NCYC1563	*Eremothecium coryli*	37,466	95.0%	94.8%	0.2%	1.9%	3.1%
Averages:			**243,846**	**97.3%**	**91.1%**	**6.2%**	**1.0%**	**1.7%**

Table C.2: BUSCO and N50 statistics of 58 Saccharomyces complex species selected for a comparative study.

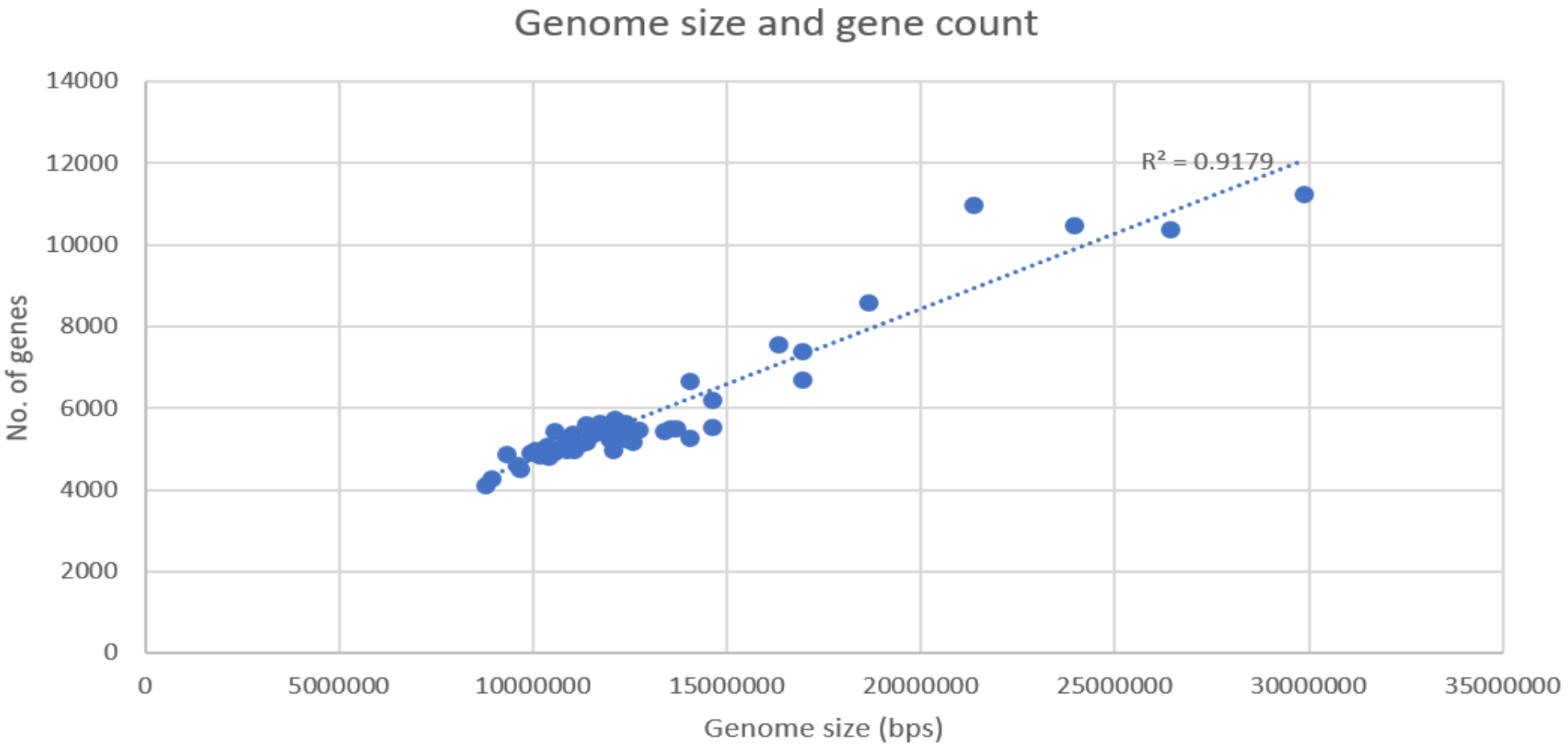


Figure C.2: Genome size and number of genes in 58 Saccharomyces complex species.

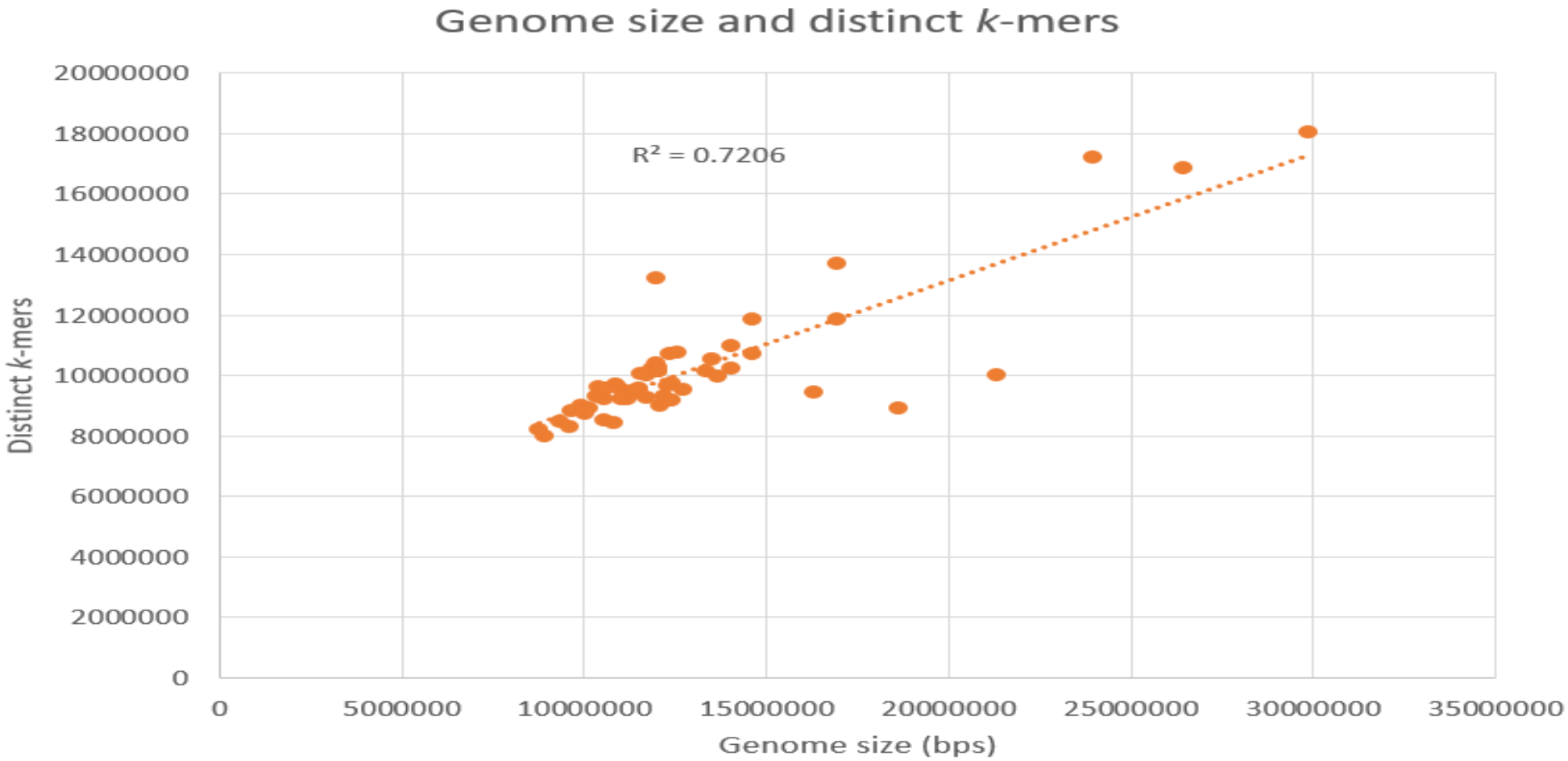


Figure C.3: Genome size and number of distinct k-mers found by Jellyfish (Marçais and Kingsford (2011)) in 58 Saccharomyces complex species.

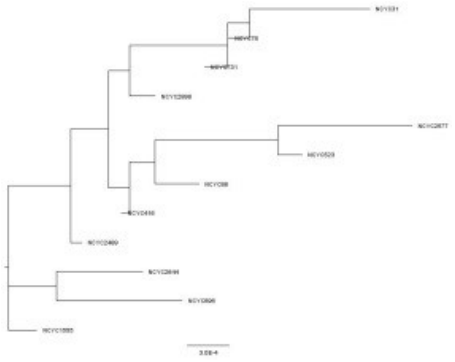
Clade	Strain ID	Species name	Assembly quality		Genome statistics			
			N50	Contigs	Size (bp)	Coding (%)	Genes	GC (%)
9	NCYC820	*Torulaspora globosa**	528,175	6,105	12,360,492	62.51	5,662	51.96
11	CBS8778	*Kluyveromyces nonfermentans**	193,381	345	9,595,169	71.77	4,603	35.94
1	NCYC2890	*Saccharomyces cariocanus***	141,541	448	12,017,915	69.58	,5526	38.22
7	NCYC2789	*Zygosaccharomyces lentus***	97,590	9,002	16,943,801	55.07	7,399	51.73
2	NCYC2878	*Kazachstania barnettii***	80,952	2,011	13,354,992	61.70	5,454	33.53
7	NCYC2403	*Zygosaccharomyces mellis**	73,822	5,393	14,623,258	59.62	6,208	48.77
1	NCYC3662	*Saccharomyces paradoxus**	66,736	858	12,051,421	66.76	4,986	37.48
2	NCYC2560	*Kazachstania sinensis***	47,906	2,182	12,035,436	70.17	5,576	46.10
2	NCYC2483	*Kazachstania piceae***	43,101	4,386	13,673,105	61.59	5,529	32.64
2	NCYC1417	*Kazachstania lodderae***	31,719	1,417	12,277,384	68.85	5,551	33.62

Clade	Strain ID	Species name	BUSCO statistics				
			Complete (%)	Single-copy (%)	Duplicated (%)	Fragmented (%)	Missing (%)
9	NCYC820	*Torulaspora globosa**	98.00	97.90	0.10	0.60	1.40
11	CBS8778	*Kluyveromyces nonfermentans**	95.30	95.00	0.30	1.80	2.90
1	NCYC2890	*Saccharomyces cariocanus***	98.10	97.50	0.60	0.70	1.20
7	NCYC2789	*Zygosaccharomyces lentus***	97.70	97.30	0.04	1.10	1.20
2	NCYC2878	*Kazachstania barnettii***	96.60	95.50	1.10	1.40	2.00
7	NCYC2403	*Zygosaccharomyces mellis**	98.30	98.10	0.20	0.70	1.00
1	NCYC3662	*Saccharomyces paradoxus**	97.70	97.10	0.60	0.80	1.50
2	NCYC2560	*Kazachstania sinensis***	97.00	96.20	0.80	1.20	1.80
2	NCYC2483	*Kazachstania piceae***	97.40	96.50	0.90	1.20	1.40
2	NCYC1417	*Kazachstania lodderae***	97.90	96.80	1.10	0.80	1.30

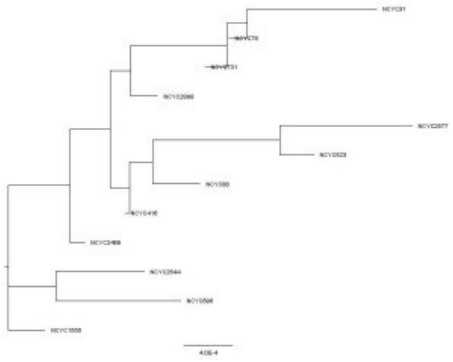
Table C.3: Key genome and BUSCO statistics of 10 strains with less than 2% Fragmented BUSCO genes and N50 >31,000, sequenced for the first time within this study and shown in order of N50 score. *= New strain, **= New species.

Appendix D

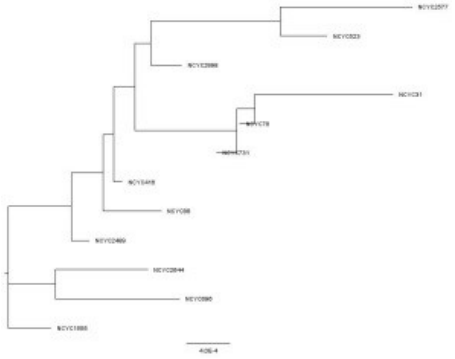
FFP phylogenetic analyses



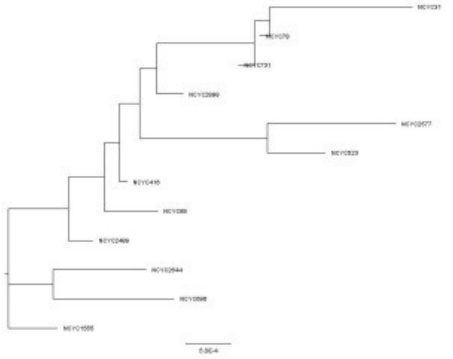
(a) *k-mer* length 10



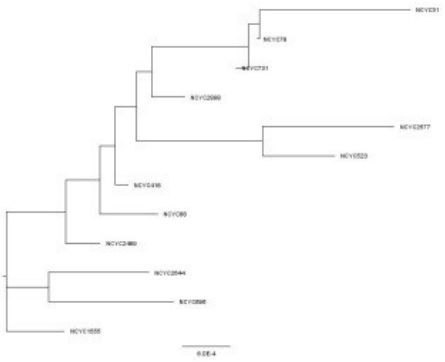
(b) *k-mer* length 11



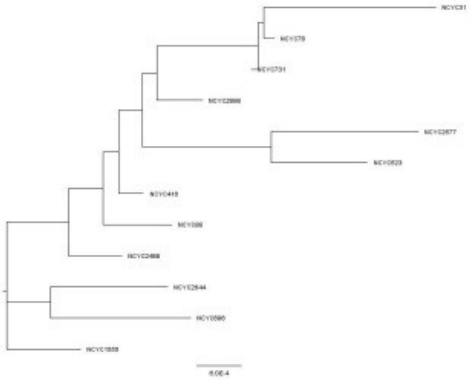
(c) *k-mer* length 12



(d) *k-mer* length 13



(e) *k-mer* length 14



(f) *k-mer* length 15

Figure D.1: Converging topologies of 14-species FFP trees with *k-mer* lengths ranging from 10 to 15. The topology of the tree remains identical for $k \geq 13$.

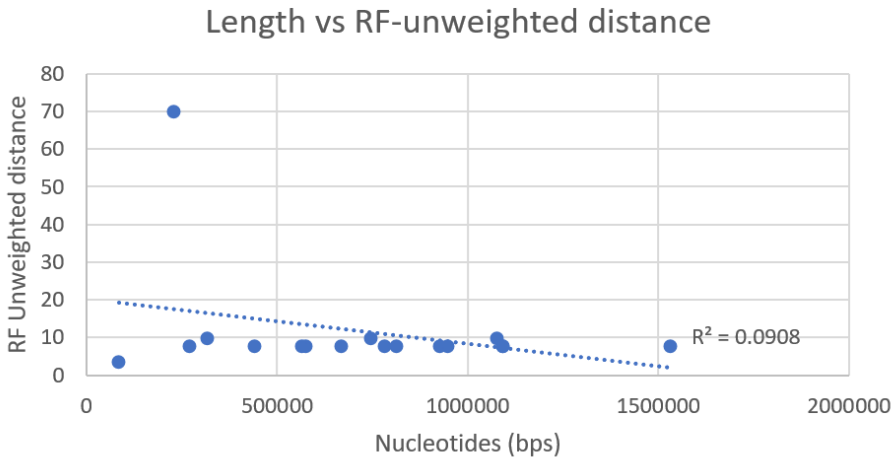


Figure D.2: A plot of sequence length vs Robinson-Foulds unweighted distance for seventeen FFP four-letter alphabet (ACGT) trees shows no significant correlation between these two factors.

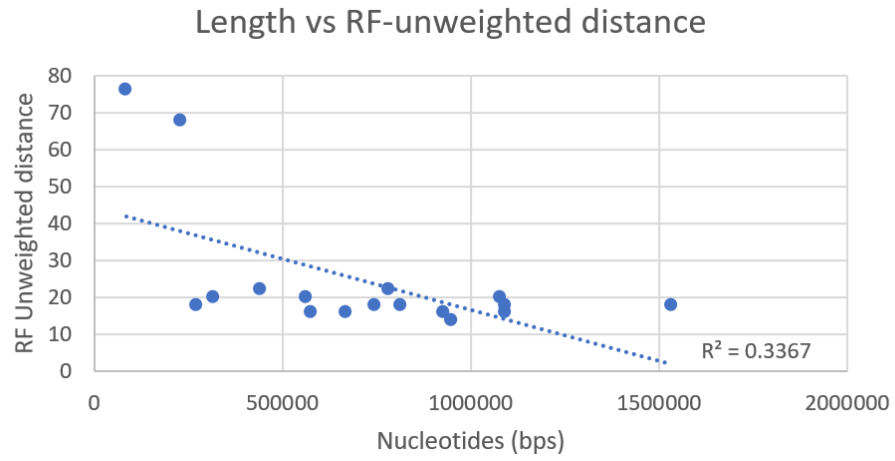
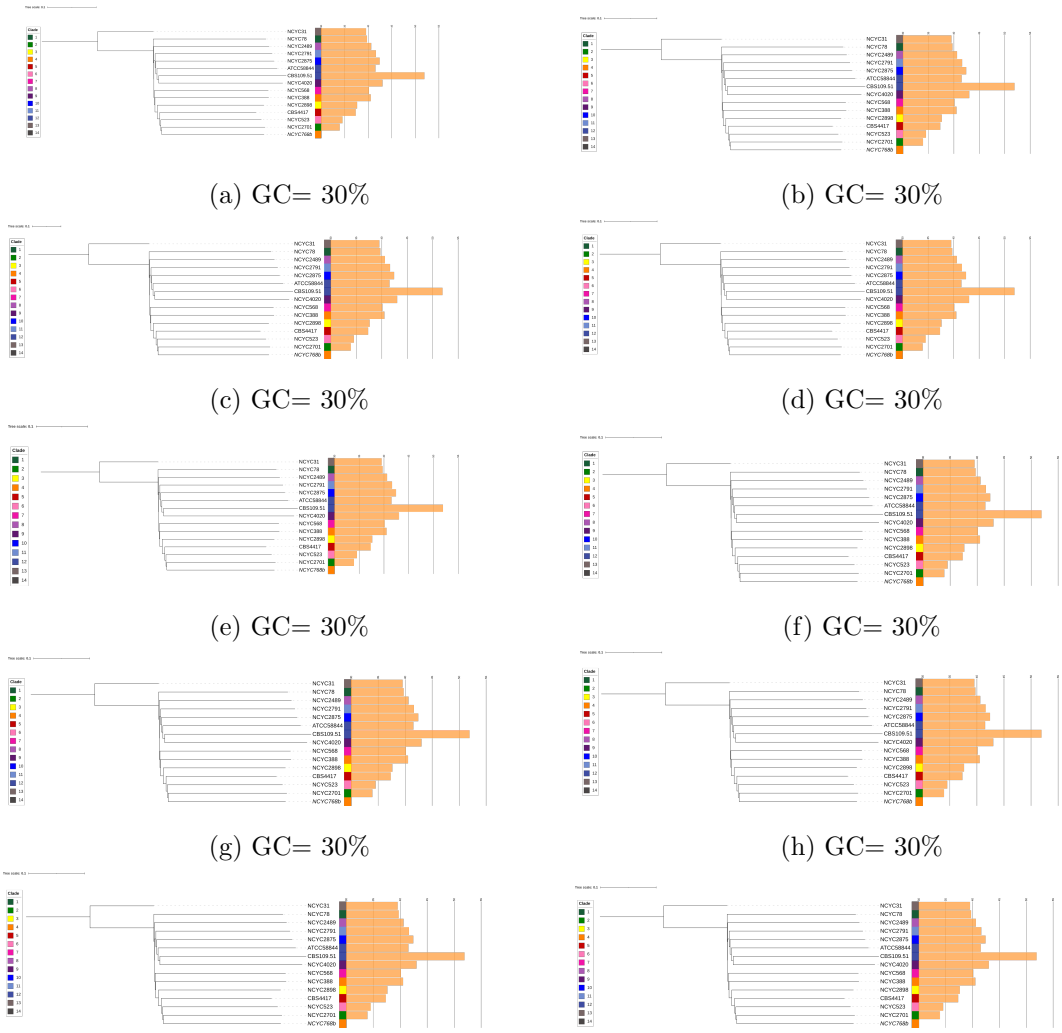


Figure D.3: A plot of sequence length vs Robinson-Foulds unweighted distance for seventeen FFP twenty-letter amino acid alphabet (AA) trees shows no significant correlation between these two factors.

Appendix E

GC simulations



(a) GC= 30% (b) GC= 30%

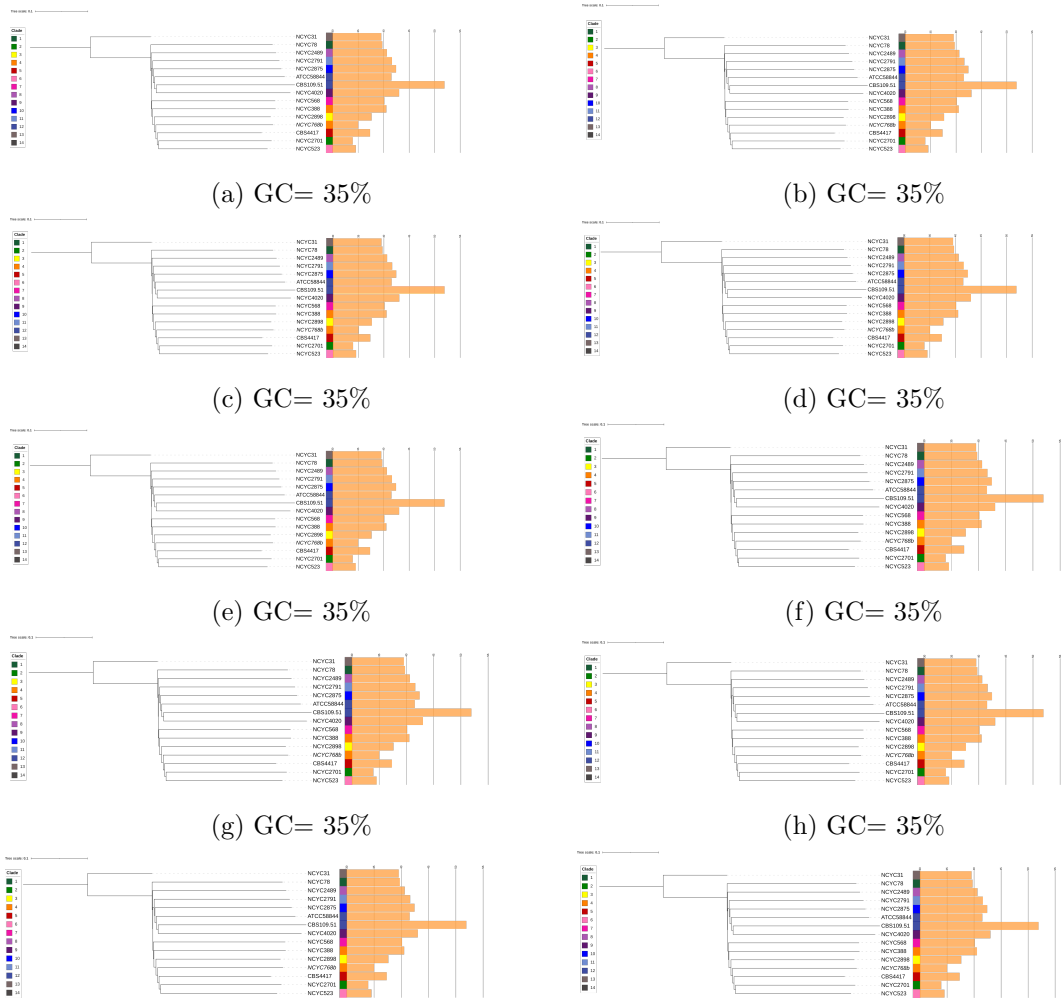
(c) GC= 30% (d) GC= 30%

(e) GC= 30% (f) GC= 30%

(g) GC= 30% (h) GC= 30%

(i) GC= 30% (j) GC= 30%

Figure E.1: FFP (four-letter DNA alphabet) phylogenetic trees of genes (without introns) from 15 Saccharomyces complex species annotated with GC content (orange bars). The GC content of the genic regions of NCYC768 was changed in each underlying dataset to GC = 30%.



(a) GC= 35%

(b) GC= 35%

(c) GC= 35%

(d) GC= 35%

(e) GC= 35%

(f) GC= 35%

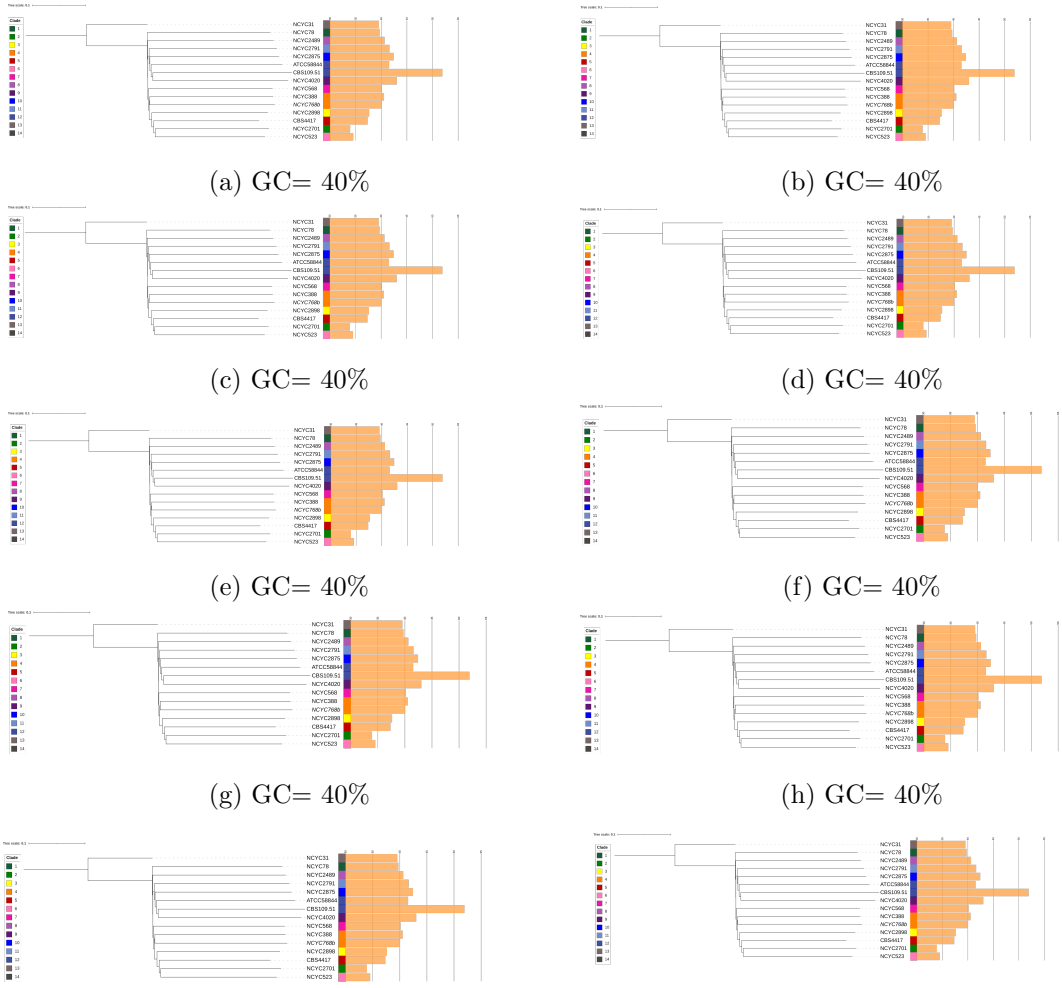
(g) GC= 35%

(h) GC= 35%

(i) GC= 35%

(j) GC= 35%

Figure E.2: FFP (four-letter DNA alphabet) phylogenetic trees of genes (without introns) from 15 Saccharomyces complex species annotated with GC content (orange bars). The GC content of the genic regions of NCYC768 was changed in each underlying dataset to GC = 35%.



(a) GC= 40%

(b) GC= 40%

(c) GC= 40%

(d) GC= 40%

(e) GC= 40%

(f) GC= 40%

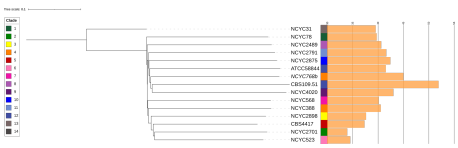
(g) GC= 40%

(h) GC= 40%

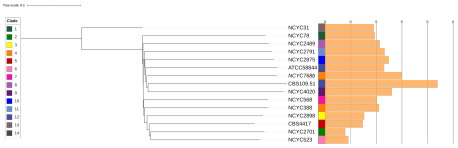
(i) GC= 40%

(j) GC= 40%

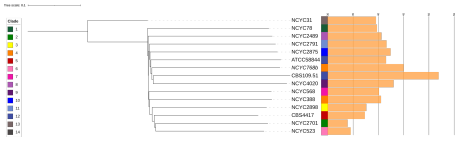
Figure E.3: FFP (four-letter DNA alphabet) phylogenetic trees of genes (without introns) from 15 Saccharomyces complex species annotated with GC content (orange bars). The GC content of the genic regions of NCYC768 was changed in each underlying dataset to GC = 40%.



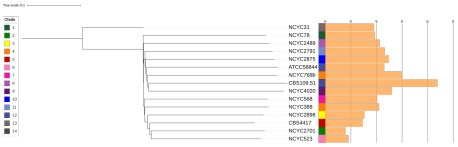
(a) GC= 45%



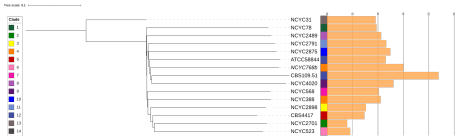
(b) GC= 45%



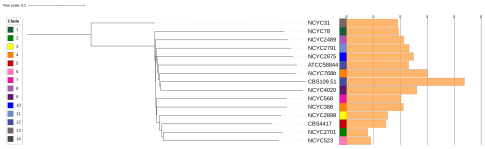
(c) GC= 45%



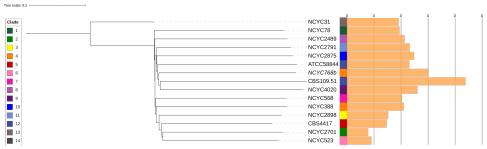
(d) GC=45%



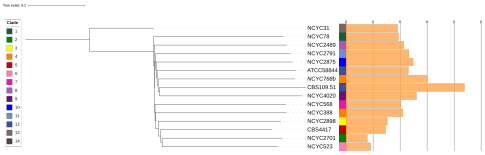
(e) GC= 45%



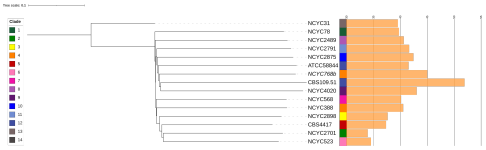
(f) GC= 45%



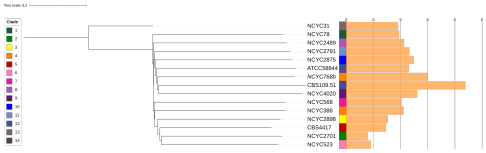
(g) GC= 45%



(h) GC= 45%

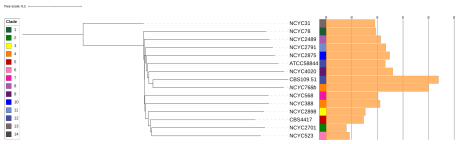


(i) GC= 45%

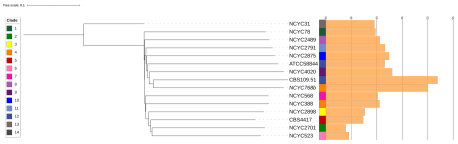


(j) GC= 45%

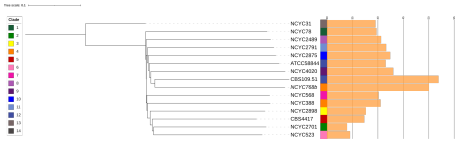
Figure E.4: FFP (four-letter DNA alphabet) phylogenetic trees of genes (without introns) from 15 Saccharomyces complex species annotated with GC content (orange bars). The GC content of the genic regions of NCYC768 was changed in each underlying dataset to GC = 45%.



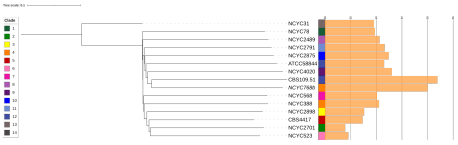
(a) GC= 50%



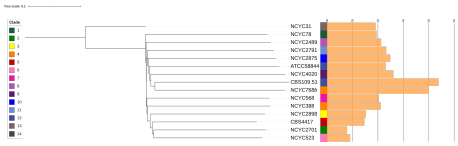
(b) GC= 50%



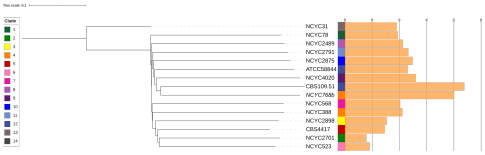
(c) GC= 50%



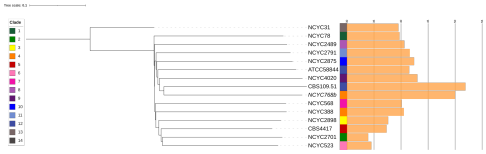
(d) GC= 50%



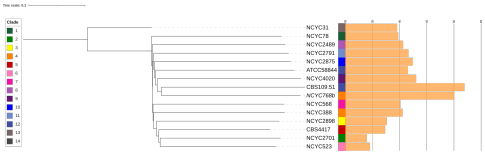
(e) GC= 50%



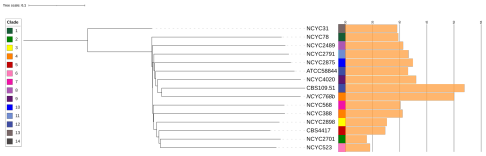
(f) GC= 50%



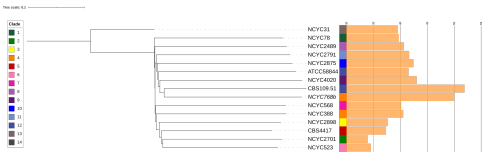
(g) GC= 50%



(h) GC= 50%



(i) GC= 50%



(j) GC= 50%

Figure E.5: FFP (four-letter DNA alphabet) phylogenetic trees of genes (without introns) from 15 Saccharomyces complex species annotated with GC content (orange bars). The GC content of the genic regions of NCYC768 was changed in each underlying dataset to GC = 50%.

Bibliography

Albertin, W., Setati, M. E., Miot-Sertier, C., Mostert, T. T., Colonna-Ceccaldi, B., Coulon, J., Girard, P., Moine, V., Pillet, M., Salin, F., et al. (2016). *Hanseniaspora uvarum* from winemaking environments show spatial and temporal genetic clustering. *Frontiers in Microbiology*, 6:1569.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410.

Andrews, S. et al. (2010). FastQC: a quality control tool for high throughput sequence data. https://github.com/s-andrews/FastQC.

Baldauf, S. L. (2003). Phylogeny for the faint of heart: a tutorial. *Trends in Genetics*, 19(6):345 – 351.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477.

Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, 21(2):163–193.

Birdsell, J. A. (2002). Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular Biology and Evolution*, 19(7):1181–1197.

Blackwell, M. (2011). The fungi: 1, 2, 3... 5.1 million species? *American journal of botany*, 98(3):426–438.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Bradnam, K. R., Seoighe, C., Sharp, P. M., and Wolfe, K. H. (1999). G + C content variation along and among *Saccharomyces cerevisiae* chromosomes. *Molecular Biology and Evolution*, 16(5):666–675.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., et al. (2010). The potential and challenges of nanopore sequencing. In *Nanoscience and technology: A collection of reviews from Nature Journals*, pages 261–268. World Scientific.

Broad-Institute (2018). Picard toolkit. `http://broadinstitute.github.io/picard/`.

Brown, T. C. and Jiricny, J. (1988). Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell*, 54(5):705–711.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAL: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.

Carmean, D. and Crespi, B. J. (1995). Do long branches attract flies? *Nature*, 373(6516):666.

Carreté, L., Ksiezopolska, E., Pegueroles, C., Gómez-Molero, E., Saus, E., Iraola-Guzmán, S., Loska, D., Bader, O., Fairhead, C., and Gabaldón, T. (2018). Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. *Current Biology*, 28(1):15–27.

Carroll, M. W., Matthews, D. A., and Hiscox, J. A. (2015). Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature Letters*, 524(7563):97–101.

Chaudhari, N. M., G. V. K. . D. C. (2016). BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports*, 6(24373).

Chaudhari, N. M., Gupta, V. K., and Dutta, C. (2016). Bpga-an ultra-fast pan-genome analysis pipeline. *Scientific Reports*, 6:24373.

Choi, J. and Kim, S.-H. (2017). A genome tree of life for the Fungi kingdom. *Proceedings of the National Academy of Sciences*, 114(35):9391–9396.

Choi, J. and Kim, S.-H. (2020). Whole-proteome tree of life suggests a deep burst of organism diversity. *Proceedings of the National Academy of Sciences*, 117(7):3678–3686.

Darwin, C. (1859). On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life. *London*, J. Murray.

Delsuc, F. and Brinkmann, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Review Genetics*, 6(5):361–375.

Dencker, T., Leimeister, C.-A., Gerth, M., Bleidorn, C., Snir, S., and Morgenstern, B. (2018). Multi-SpaM: a maximum-likelihood approach to phylogeny reconstruction using multiple spaced-word matches and quartet trees. In *RECOMB International conference on Comparative Genomics*, pages 227–241. Springer.

Dietrich, F. S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pöhlmann, R., Luedi, P., Choi, S., et al. (2004). The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, 304(5668):304–307.

Dujon, B. (2010). Yeast evolutionary genomics. *Nature Reviews Genetics*, 11(7):512–524.

Dujon, B. A. and Louis, E. J. (2017). Genome diversity and evolution in the budding yeasts (Saccharomycotina). *Genetics*, 206(2):717–750.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797.

Fabian, K. and Bernard, H. (2019). Phylonium- fast and accurate estimation of evolutionary distances. *GitHub*.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, pages 783–791.

Felsenstein, J. (1989). PHYLIP-phylogeny inference package (version 3.2). *Cladistics*, 5(163):6.

Flicek, P. and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, 6:S6–S12.

Foury, F., Roganti, T., Lecrenier, N., and Purnelle, B. (1998). The complete sequence of the mitochondrial genome of *Saccharomyces cerevisiae*. *FEBS letters*, 440(3):325–331.

Fouts, D. E., Brinkac, L., Beck, E., Inman, J., and Sutton, G. (2012). PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic acids research*, 40(22):e172–e172.

Gabaldón, T. (2020). Hybridization and the origin of new yeast lineages. *FEMS Yeast Research*, 20(5):foaa040.

Gabaldón, T. and Fairhead, C. (2019). Genomes shed light on the secret life of *Candida glabrata*: not so asexual, not so commensal. *Current genetics*, 65(1):93–98.

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., Teiling, C., Steffy, B., Taylor, M., Schwartz, A., Richardson, T., White, C., Baele, G., Maere, S., and Verstrepen, K. (2016). Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell*, 166(6):1397 – 1410.e16.

Galtier, N. and Lobry, J. (1997). Relationships between genomic G + C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of molecular evolution*, 44(6):632–636.

Gao, L.-z. and Innan, H. (2004). Very low gene duplication rate in the yeast genome. *Science*, 306(5700):1367–1370.

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

Gascuel, O. (1997). Bionj: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–695.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, 11(5):759–769.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. *Science*, 274(5287):546–567.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, 29(7):644.

Hall, B. G. (2005). Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Molecular Biology and Evolution*, 22(3):792–802.

Hamilton, W. L., Tonkin-Hill, G., Smith, E., Aggarwal, D., Houldcroft, C. J., Warne, B., Brown, C. S., Meredith, L. W., Hosmillo, M., Jahun, A. S., et al. (2020). Genomic epidemiology of COVID-19 in care homes in the east of england. *medRxiv*.

Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174.

Hendler, A., Medina, E. M., Kishkevich, A., Abu-Qarn, M., Klier, S., Buchler, N. E., de Bruin, R. A., and Aharoni, A. (2017). Gene duplication and co-evolution of G1/S transcription factor specificity in fungi are essential for optimizing cell fitness. *PLoS Genetics*, 13(5):e1006778.

Hittinger, C. T., Rokas, A., Bai, F.-Y., Boekhout, T., Goncalves, P., Jeffries, T. W., Kominek, J., Lachance, M.-A., Libkind, D., Rosa, C. A., et al. (2015). Genomics and the making of yeast biodiversity. *Current Opinion in Genetics & Development*, 35:100–109.

Holmes, E. C., Dudas, G., Rambaut, A., and Andersen, K. G. (2016). The evolution of Ebola virus: Insights from the 2013–2016 epidemic. *Nature*, 538(7624):193–200.

Hsiang, T. and Baillie, D. L. (2005). Comparison of the yeast proteome to other fungal genomes to find core fungal genes. *Journal of molecular evolution*, 60(4):475–483.

Hu, X., Yuan, J., Shi, Y., Lu, J., Liu, B., Li, Z., Chen, Y., Mu, D., Zhang, H., Li, N., et al. (2012). pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*, 28(11):1533–1535.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314.

Huffnagle, G. B. and Noverr, M. C. (2013). The emerging world of the fungal microbiome. *Trends in microbiology*, 21(7):334–341.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. *Nature microbiology*, 1(5):1–6.

Hurst, L. D. and Merchant, A. R. (2001). High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1466):493–497.

Illumina (2016). An introduction to next generation sequencing technology.

Johnston, M., Hillier, L., Riles, L., Albermann, K., André, B., Ansorge, W., Benes, V., Brückner, M., Delius, H., Dubois, E., et al. (1997). The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. *Nature*, 387(6632):87–90.

Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2015). Treescape: Statistical exploration of landscapes of phylogenetic trees. *https://cran.r-project.org/web/packages/treescape/index.html*.

Jordão, A., Vilela, A., and Cosme, F. (2015). From sugar of grape to alcohol of wine: Sensorial impact of alcohol in wine. *Beverages*, 1(4):292–310.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. *New York Academic Press*, page 21–132.

Jun, S.-R., Sims, G. E., Wu, G. A., and Kim, S.-H. (2010). Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proceedings of the National Academy of Sciences*, 107(1):133–138.

Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic acids research*, 30(14):3059–3066.

Kellis, M., Birren, B. W., and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983):617–624.

Kimura, M. (1987). Molecular evolutionary clock and the neutral theory. *Journal of Molecular Evolution*, 26(1):24–33.

Knop, M. (2011). Yeast cell morphology and sexual reproduction – a short overview and some considerations. *Comptes Rendus Biologies*, 334(8–9):599 – 606.

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., and Zdobnov, E. M. (2019). Orthodb v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic acids research*, 47(D1):D807–D811.

Kuhner, M. K. and Yamato, J. (2015). Practical performance of tree comparison metrics. *Systematic Biology*, 64(2):205–214.

Kurtzman, C., Fell, J. W., and Boekhout, T. (2011). *The yeasts: a taxonomic study*. Elsevier.

Kurtzman, C. P. (2003). Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the *Saccharomycetaceae*, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Research*, 4(3):233–245.

Kurtzman, C. P. and Robnett, C. J. (2003). Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses. *FEMS Yeast Research*, 3(4):417–432.

Kurtzman, C. P. and Robnett, C. J. (2013). Relationships among genera of the Saccharomycotina (Ascomycota) from multigene phylogenetic analysis of type species. *FEMS Yeast Research*, 13(1):23.

Kutty, S. N. and Philip, R. (2008). Marine yeasts — a review. *Yeast*, 25(7):465–483.

Lees, J., Kendall, M., Parkhill, J., Colijn, C., Bentley, S., and Harris, S. (2018). Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Research*, 3(33).

Leimeister, C.-A., Boden, M., Horwege, S., Lindner, S., and Morgenstern, B. (2014). Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, 30(14):1991–1999.

Leimeister, C.-A., Sohrabi-Jahromi, S., and Morgenstern, B. (2017). Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics*, 33(7):971–979.

Letunic, I. and Bork, P. (2006). Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–128.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079.

Li, Y., David, K. T., Shen, X.-X., Steenwyk, J. L., Halanych, K. M., and Rokas, A. (2020). Feature frequency profile-based phylogenies are inaccurate. *Proceedings of the National Academy of Sciences*, 117(50):31580–31581.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Liti, G., Barton, D. B., and Louis, E. J. (2006). Sequence diversity, reproductive isolation and species concepts in saccharomyces. *Genetics*, 174(2):839–850.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–341.

Llorente, B., Malpertuy, A., Neuvéglise, C., de Montigny, J., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 18. comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. *FEBS letters*, 487(1):101–112.

Lücking, R., Huhndorf, S., Pfister, D. H., Plata, E. R., and Lumbsch, H. T. (2009). Fungi evolved right on track. *Mycologia*, 101(6):810–822.

Lunter, G. and Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6):936–939.

Lynch, D. B., Logue, M. E., Butler, G., and Wolfe, K. H. (2010). Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. *Genome Biology and Evolution*, 2:572–583.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6):3140–3145.

Malpertuy, A., Tekaia, F., Casarégola, S., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., Bon, E., Brottier, P., de Montigny, J., et al. (2000). Genomic exploration of the hemiascomycetous yeasts: 19. ascomycetes-specific genes. *FEBS letters*, 487(1):113–121.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B. J. (2017). KAT: a k-mer analysis toolkit to quality control ngs datasets and genome assemblies. *Bioinformatics*, 33(4):574–576.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *TRENDS in Genetics*, 19(6):330–338.

Marsolier-Kergoat, M.-C. and Yeramian, E. (2009). GC content and recombination: reassessing the causal effects for the *Saccharomyces cerevisiae* genome. *Genetics*, 183(1):31–38.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764.

McCarthy, C. G. and Fitzpatrick, D. A. (2019). Pan-genome analyses of model fungal species. *Microbial genomics*, 5(2).

McGovern, P. E., Zhang, J., Tang, J., Zhang, Z., Hall, G. R., Moreau, R. A., Nuñez, A., Butrym, E. D., Richards, M. P., Wang, C.-s., Cheng, G., Zhao, Z., and Wang, C. (2004). Fermented beverages of pre- and proto-historic china. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17593–17598.

Merchant, S., Wood, D. E., and Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ*, 2:e675.

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46.

Michel, R. H. and McGovern, P. E. (1992). Chemical evidence for ancient beer. *Nature*, 360(6399):24–24.

Mirzoyan, Z., Sollazzo, M., Allocca, M., Valenza, A. M., Grifoni, D., and Bellosta, P. (2019). *Drosophila melanogaster*: A model organism to study cancer. *Frontiers in Genetics*, 10:51.

Montero, C. M., Dodero, M. R., Sánchez, D. G., and Barroso, C. (2004). Analysis of low molecular weight carbohydrates in food and beverages: a review. *Chromatographia*, 59(1-2):15–30.

Mora, M., Donati, C., Medini, D., Covacci, A., and Rappuoli, R. (2006). Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach. *Current opinion in microbiology*, 9(5):532–536.

Mueller, L. and Ayala, F. (1982). Estimation and interpretation of genetic distance in empirical studies. *Genetical Research*, 40:127–137.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., et al. (2000). A whole-genome assembly of Drosophila. *Science*, 287(5461):2196–2204.

Naseeb, S., Ames, R. M., Delneri, D., and Lovell, S. C. (2017). Rapid functional and evolutionary changes follow gene duplication in yeast. *Proceedings of the Royal Society B: Biological Sciences*, 284(1861):20171393.

Naumov, G. I., James, S. A., Naumova, E. S., Louis, E. J., and Roberts, I. N. (2000). Three new species in the saccharomyces sensu stricto complex: *Saccharomyces cariocanus*, *Saccharomyces kudriavzevii* and *Saccharomyces mikatae*. *International journal of systematic and evolutionary microbiology*, 50(5):1931–1942.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

Nye, T. M., Lio, P., and Gilks, W. R. (2006). A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics*, 22(1):117–119.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using minhash. *Genome Biology*, 17(1):132.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A., and Parkhill, J. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693.

Pasteur, L. (1857). Mémoire sur la fermentation alcoolique. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, 45:1032–1036.

Penny, D. and Hendy, M. (1986). Estimating the reliability of evolutionary trees. *Molecular Biology and Evolution*, 3(5):403–417.

Penny, D. and Hendy, M. D. (1985). Testing methods of evolutionary tree construction. *Cladistics*, 1(3):266–278.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L., and Marais, G. A. (2012). Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biology and Evolution*, 4(7):675–682.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., et al. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, 556(7701):339–344.

Phillips, M. J., Delsuc, F., and Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*, 21(7):1455.

Prevosti, F. J. and Chemisquy, M. A. (2010). The impact of missing data on real morphological phylogenies: influence of the number and distribution of missing entries. *Cladistics*, 26(3):326–339.

Quinlan, A. R. and Hall, I. M. (2010). BEDtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rambaut, A. and Drummond, A. (2009). Figtree v1.3.1.

Rambaut, A. and Grass, N. C. (1997). Seq-Gen: an application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238.

Rhind, N., Chen, Z., Yassour, M., Thompson, D. A., Haas, B. J., Habib, N., Wapinski, I., Roy, S., Lin, M. F., Heiman, D. I., et al. (2011). Comparative functional genomics of the fission yeasts. *Science*, 332(6032):930–936.

Rhodes, J., Beale, M. A., Vanhove, M., Jarvis, J. N., Kannambath, S., Simpson, J. A., Ryan, A., Meintjes, G., Harrison, T. S., Fisher, M. C., et al. (2017). A population genomics approach to assessing the genetic basis of within-host microevolution underlying recurrent cryptococcal meningitis infection. *G3: Genes, Genomes, Genetics*, 7(4):1165–1176.

Riley, R., Haridas, S., Wolfe, K. H., Lopes, M. R., Hittinger, C. T., Göker, M., Salamov, A. A., Wisecaver, J. H., Long, T. M., Calvey, C. H., et al. (2016). Comparative genomics of biotechnologically important yeasts. *Proceedings of the National Academy of Sciences*, 113(35):9882–9887.

Roberts, R. J., Carneiro, M. O., and Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biology*, 14(6):405.

Robinson, D. and Foulds, L. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131 – 147.

Roychowdhury, T., Vishnoi, A., and Bhattacharya, A. (2013). Next-generation anchor based phylogeny (Nex-ABP): Constructing phylogeny from next-generation sequencing data. *Scientific Reports*, 3.

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

Sarmashghi, S., Bohmann, K., Gilbert, M. T. P., Bafna, V., and Mirarab, S. (2019). Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology*, 20(1):34.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Bolchacova, E., Voigt, K., Crous, P. W., et al. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proceedings of the National Academy of Sciences*, 109(16):6241–6246.

Sedlazeck, F. J., Rescheneder, P., and Von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, 29(21):2790–2791.

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. In *Gene Prediction*, pages 227–245. Springer.

Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., Haase, M. A., Wisecaver, J. H., Wang, M., Doering, D. T., Boudouris, J. T., Schneider, R. M., Langdon, Q. K., Ohkuma, M., Endoh, R., Takashima, M., ichiroh Manabe, R., Čadež, N., Libkind, D., Rosa, C. A., DeVirgilio, J., Hulfachor, A. B., Groenewald, M., Kurtzman, C. P., Hittinger, C. T., and Rokas, A. (2018). Tempo and mode of genome evolution in the budding yeast subphylum. *Cell*, 175(6):1533 – 1545.e20.

Shen, X.-X., Zhou, X., Kominek, J., Kurtzman, C., Todd Hittinger, C., and Rokas, A. (2016). Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. *G3: Genes, Genomes, Genetics*, 6:3927–3939.

Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., and Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research*.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, 19(6):1117–1123.

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009a). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, 106(8):2677–2682.

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009b). Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proceedings of the National Academy of Sciences*, 106(40):17077–17082.

Sims, G. E. and Kim, S.-H. (2011). Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences*, 108(20):8329–8334.

Sober, E. (1983). Parsimony in systematics: Philosophical issues. *Annual Review of Ecology and Systematics*, 14:335–357.

Sokal, R. and Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Souciet, J.-L., Dujon, B., Gaillardin, C., Johnston, M., Baret, P. V., Cliften, P., Sherman, D. J., Weissenbach, J., Westhof, E., Wincker, P., et al. (2009). Comparative genomics of protoploid saccharomycetaceae. *Genome research*, 19(10):1696–1709.

Spencer, J. F. and Spencer, D. M. (2013). *Yeasts in natural and artificial habitats*. Springer Science & Business Media.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Steels, H., James, S., Roberts, I., and Stratford, M. (1999). *Zygosaccharomyces lentus*: a significant new osmophilic, preservative-resistant spoilage yeast, capable of growth at low temperature. *Journal of Applied Microbiology*, 87(4):520–527.

Steenwyk, J. L., Opulente, D., Kominek, J., Shen, X.-X., Zhou, X., LaBella, A. L., Bradley, N. P., Eichman, B. F., Cadez, N., Libkind, D., DeVirgilio, J., Hulfachor, A. B., Kurtzman, C. P., Hittinger, C. T., and Rokas, A. (2019). Extensive loss of cell cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts. *bioRxiv*.

Stratford, M., Bond, C. J., James, S. A., Roberts, N., and Steels, H. (2002). *Candida davenportii* sp. nov., a potential soft-drinks spoilage yeast isolated from a wasp. *International Journal of Systematic and Evolutionary Microbiology*, 52(4):1369–1375.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences*, 85(8):2653–2657.

Sukumaran, J. and Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569.

Swofford, D. L. (2001). Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0. b5.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, 278(5338):631–637.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17:57–86.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., et al. (2005). Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102(39):13950–13955.

Thompson, D. S., Carlisle, P. L., and Kadosh, D. (2011). Coevolution of morphology and virulence in candida species. *Eukaryotic Cell*, 10(9):1173–1182.

Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison—a review. *Bioinformatics*, 19(4):513.

Vinogradov, A. E. (2003). DNA helix: the importance of being GC-rich. *Nucleic acids research*, 31(7):1838–1844.

Vu, D., Groenewald, M., Szöke, S., Cardinali, G., Eberhardt, U., Stielow, B., de Vries, M., Verkleij, G., Crous, P., Boekhout, T., et al. (2016). DNA barcoding analysis of more than 9 000 yeast isolates contributes to quantitative thresholds for yeast species and genera delimitation. *Studies in Mycology*, 85:91–105.

Waddell, P. J. and Steel, M. (1997). General time-reversible distances with unequal rates across sites: Mixing and inverse gaussian distributions with invariant sites. *Molecular Phylogenetics and Evolution*, 8(3):398 – 414.

Wang, A. and Ash, G. J. (2015). Whole genome phylogeny of bacillus by feature frequency profiles (FFP). *Scientific Reports*, 5.

Waterhouse, R. M., Zdobnov, E. M., and Kriventseva, E. V. (2011). Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. *Genome Biology and Evolution*, 3:75–86.

Wendland, J. and Walther, A. (2005). *Ashbya gossypii*: a model for fungal developmental biology. *Nature Reviews Microbiology*, 3(5):421–429.

West, C., James, S. A., Davey, R. P., Dicks, J., and Roberts, I. N. (2014). Ribosomal DNA sequence heterogeneity reflects intraspecies phylogenies and predicts genome structure in two contrasting yeast species. *Systematic Biology*, 63(4):543.

Wetterstrand, K. (2016). DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). *www.genome.gov/sequencingcostsdata*.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*, 51(2):221.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579.

Wolfe, K. H., Armisén, D., Proux-Wera, E., OhEigeartaigh, S. S., Azam, H., Gordon, J. L., and Byrne, K. P. (2015). Clade-and species-specific features of genome evolution in the saccharomycetaceae. *FEMS Yeast Research*, 15(5).

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46.

Wu, G. A., Jun, S.-R., Sims, G. E., and Kim, S.-H. (2009). Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proceedings of the National Academy of Sciences*, 106(31):12826–12831.

Yi, H. and Jin, L. (2013). Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acids Research*, 41(7):e75–e75.

Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J., and Yu, J. (2012). PGAP: pan-genomes analysis pipeline. *Bioinformatics*, 28(3):416–418.

Zielezinski, A., Girgis, H. Z., Bernard, G., Leimeister, C.-A., Tang, K., Dencker, T., Lau, A. K., Roehling, S., Choi, J., Waterman, M. S., et al. (2019). Benchmarking of alignment-free sequence comparison methods. *BioRxiv*, page 611137.

Zielezinski, A., Vinga, S., Almeida, J., and Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*, 18(1):186.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. *Evolving genes and proteins*, 97:97–166.