Bioinformatic approaches to identify key genes involved in microbial metabolic control

PhD Thesis for the BIO school at the University Of East Anglia

Joseph Robert Shepherd

17 December, 2021

Abstract

This thesis combines biological experiments, computational analyses and software development in order to gain new knowledge of microorganisms that could be of benefit to industrial biotechnology processes. It has three linked components.

The main part involves the identification of genomic locations in a dataset of Whole Genome Sequenced (WGS) *Saccharomyces cerevisae* strains that correlate with furfural resistance, a chemical common in treated lignocellulosic waste biomass. The project comprises both experimental data gathering and computational analysis of the resulting datasets. Following an association analysis of the strains' phenotypes and genome-wide genotypes, directed evolution (DE) experiments are carried out to assess the impact on the strains' genomes. The sequence composition of the resulting strains is then compared to their states prior to the DE experiments in order to assess potential evolutionary paths, and to discover whether multi-strain resistance analysis is comparable to the directed evolution of select strains.

In the second part, diverse yeast strains are grown in YNB media, with subsequently obtained Nuclear Magnetic Resonance (NMR) spectra analysed computationally to quantitatively assess metabolite concentrations. Various *Saccharomyces cerevisiae* strains are also grown in malt extract media. The results of both analyses are examined and, where possible, compared in order to assess the relative potential of the strains in various industrial brewing processes. A Genome Wide Association Study on the malt datasets indicates genes potentially involved in metabolite quantity, that may taken forward in future research activities.

The final part of this thesis considers the computational prediction of specific cytochrome operons in all bacterial CDS genomes in the RefSeq database (2020). A new software program, ETMiner, is introduced and illustrated through its application to datasets with potentially interesting industrial profiles.

Github link for additional resources: https://github.com/Joenetics/PhD_Thesis.git

Acknowledgements

Thanks to my dad for pushing me onto this doctoral path; thanks to my mum for keeping me on it. I'd like to thank all my friends and family for all their encouragement.

Nothing in my PhD would have been possible without copious amounts of help from people who selflessly shared their knowledge and skills with me. From Ms. Carmen Nueno-Palop who instructed me in some basic yeast knowledge and protocols, to Dr. Adam Elliston who provided me with all the strains needed from the NCYC and his lab protocols to date- which my experiments follow.

Thank you to the entire lab for impromptu brainstorming sessions, sorely needed moral support and making long lab days fun. Thank you to Dr. Marcus Edwards for being ever helpful, available and patient with ETMiner from concept to completion and final analysis. I don't know how much the lab as a whole will function without you.

Without the help of the best supervisor, Prof. Tom Clarke, I would not have gotten half of my ideas, projects or experiments to pan out or, often, the ideas in the first place. Special thanks to Dr. Jo Dicks who enabled my projects' bioinformatics and statistics to miss the moon and reach for the stars.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The need for renewable metabolites

Human societies rely on numerous chemical compounds to function on a day-to-day level, with a total of 8,300 million metric tons of virgin plastic having been produced to date. As pressure increases to reduce fossil fuel use, any industrial chemicals currently relying on a supply of crude oil for their plastic production will undoubtedly face scrutiny. Therefore, biological systems that replace crude oil with simple sugars as a feedstock will become crucial to the future of chemical production in a range of industries [2, 3].

Yeasts, and *Saccharomyces cerevisae* (brewer's yeast) in particular, have been a platform for chemical production in industry for many years due to their well-established functional and safety profile. While other strains of yeast within genera such as *Candida*, *Endomycopsis*, and *Kluyveromyces* as well as other microorganisms, are important and might gain in popularity, S.*cerevisiae* strains will likely always be a major player due to their extensive safety profile in the food industry, their high fermentation rates and their superior ability to perform as a model organism for non-fungal eukaryotes [4, 5, 6, 7].

1.2 Mining the depths of microbial genomic data

For decades, a major focus of research was the accumulation and analysis of complete genomic data for species. This made sense, data were scarce and there was much left to learn. These early, species-level investigations were achieved through high-quality Sanger sequencing.

While Sanger sequencing remains the gold-standard for confirming sequence [8], it is prohibitively expensive. The original 'draft' human genome sequence was created through Sanger sequencing at a cost of $300 Million [9]. The cost is now, through Next Generation Sequencing (NGS) technologies, closer to $1000 [9].

Today's NGS reads are often of lower confidence than first generation reads, particularly in the 3' region of Illumina reads affected by phasing[8]. Phasing is an inherent risk in Illumina sequencing. It occurs after signal detection (for base calling) when the blocker is not correctly removed from the read and leads to the entire read being a base call behind and thus 'out of phase' with the other reads. Short reads, coupled with reduced quality as the read elongates, results in Illumina relying on reference genomes assembled through Sanger sequencing for completion (Nanopore[10] technologies and others promise to bridge this technical gap). Moreover, NGS can struggle in specific areas such as DNA regions with highly repetitive sequence motifs/structures (microsatellites, ...), high GC content or other idiosyncrasies [8, 9]

Despite its shortcomings, NGS has enabled the exploration of sub-species level genomes and population genomes. Now the focus is on analysing, or 'mining', large datasets consisting of tens or even thousands of genomes. This analysis encapsulates a variety of methodologies; from counting Copy Number Variations (CNVs) [11] and other repetitive regions, to assessing specific Single Nucleotide Polymorphisms (SNPs) [12] that may affect organism function, to predicting the structure [13] and function [14] of proteins produced by genomic regions.

This thesis will discuss some of these methodologies as systems to identify specific genes involved in key metabolic pathways with the ultimate goal of improved renewable sources of metabolite production. Using SNP-based Genome Wide Association Studies (GWAS) [12], we delve into the high-quality sequenced genomes of various yeast from the National Collection of Yeast Cultures (NCYC) [1]. Moreover, we use Gene Ontology (GO) [14] and manually curated experimental data to further predict resultant functional changes of said SNPs. Additionally, we use tools including BLAST [15], PRED-TMBB [16], Protsite [17] and the interactive Tree of Life (iTOL) [18] to predict and visualise cytochrome operons (including type and general structure) from bacterial genomic data.

Both of these methodologies utilise large databases of under-characterised genomic information and computational tools to infer the genetic basis of various

metabolic phenotypes. Whether raw sequencing data or curated CDS DNA data, all deal with raw DNA sequences that are analysed by unbiased algorithms. This ensures that the current analysis is unaffected by previous characterisations.

All these tools and analyses are of vital importance if we are to begin to make sense of the enormous quantities of genomic data publicly available. The ability to conduct preliminary categorisation and assessment of genomic elements based entirely on computational methodologies is vital to reduce the targets of potential experimental interest.

With accurate predictions, it is possible to both identify genomic regions of interest and prepare appropriate experimental methodologies in advance. The potential cost-saving effects of such predictions can be invaluable. They also provide an avenue for direct mutational experiments- including the use of plasmids with regions of genomic interest.

1.3 Yeast as a platform chemical production system

1.3.1 Yeast diversity and utility

As this body of work will discuss the usage of various yeast strains, it is first necessary to provide a definition of yeast. Yeast are single-celled eukaryotic organisms within the kingdom Fungi that derive their name from the foamy product of their anaerobic digestion of sugars. They are divided into two phyla; Ascomycota and Basidiomycota (see Figure 1.3.1), with Ascomycota - the 'true yeasts' - containing the *Saccharomyces* genus, some of whose ten species are widely used in industrial processes. While yeast genomes are more complex than those of single-celled prokaryotes, sharing genomic features and phenomena such as introns and polyploidy with other eukaryotes, in terms of genome size they are similar to prokaryotes- with only moderate quantities of intergenic DNA [19]. Found in almost every environment in the wild, various yeast have been found to digest a wide array of feedstocks, while capable of producing the metabolic chemicals characteristic of eukaryotes [20].

Approximately 1,500 yeast species are currently defined, estimated to be $\sim 10\%$ of those yet to be discovered, and while 'yeast' is a non-scientific name without a specific definition, its holders usually share similarities. For example, species

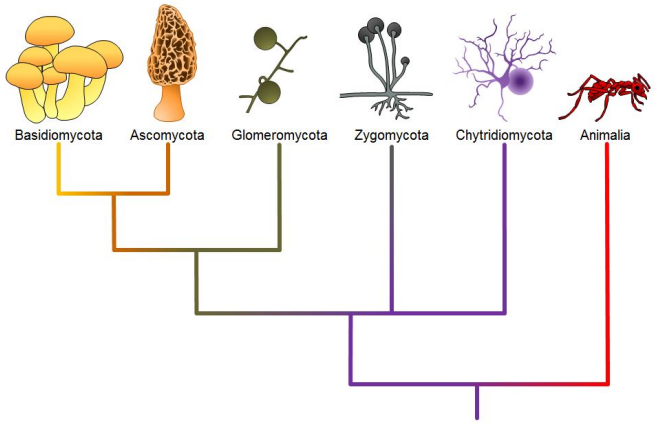


Figure 1.3.1: Overview of fungal phyla. Yeasts are found in the Ascomycota and Basidiomycota. Taken from https://courses.lumenlearning.com/suny-osbiology2e/chapter/classifications-of-fungi/.

defined as yeast most often undergo mitosis for cell division, but rely on meiosis during times of stress to acquire novel genetic variants and increase cell adaptability [21].

Within this project, yeast strains from over 200 yeast species - along with their genome sequences - were potentially available for analysis, with each species dataset ranging from one to 200 strains. However, with the aim of minimising background genetic 'noise', it was deemed beneficial to select a sizeable set of strains with a high degree of genomic similarity between them, not least because yeast are known for a high frequency of aneuploid/polyploid genomes. The final dataset chosen for analysis comprised approximately 200 *Saccharomyces cerevisae* strains. Firstly, this was the largest within-species group of strains available to the project. Secondly, the strains are known to share a common out-of-China ancestral origin [22]. As such, they can be relied upon to be highly genetically similar and for a larger share of the phenotypic differences between strains to be attributable to discrete Single Nucleotide Polymorphisms (SNPs) rather than Copy Number Variations (CNVs) or other genomic variations (e.g, inversions, translocations, ploidy). Thirdly, and potentially most importantly, this species has been used for millennia as a production platform for alcoholic beverages. Furthermore, yeasts, and specifically *Saccharomyces cerevisae*, have an extensive safety profile as metabolic platform chemical producers [4]. Used widely in industry to produce everything from food additives such as acidity regulators (Succinate [23]) and ethanol to plastics [24] and antibodies [25], they are reliable, safe and generally non-pathogenic. Succinate holds particular focus for research as a highly versatile molecule that

is generally considered to be safe. However, it has also been implicated in gut inflammation and tumorgenesis- such as when produced by gut microflora in dietary fibre digestion [23, 26]. Regardless of succinate's eventual regulatory fate, the final methodology for SNP elucidation would remain the same as for any other metabolite or phenotype [27, 28].

Due to ecological and environmental concerns, yeast have stepped in as a prime production platform for chemical production for industry [29]. Unfortunately, in the early days of large-scale biofuel development (largely ethanol), in order to produce the quantities necessary to satisfy human requirements, a sizeable portion of farmland was devoted to providing the feedstocks for industrial fermenters instead of humans[30]. To combat this, and any conflicts of interest between consumers and industrial giants, an alternative source of feedstock was required; lignocellulosic waste biomass.

The use of lignocellulosic waste biomass in biofuel and biochemical production ensures that arable land suitable for farming is utilised in growing crops for human consumption. The by-products that would otherwise be wasted are then repurposed as a yeast feedstock for industry. This creates value for farmers, which can then be invested in further productivity gains in feeding the human population.

Unfortunately, the inedible sections of plants (stalks/leaves/roots) are usually composed of harder-to-digest sugars in complex, tight compositions. As such, pre-treatments to the feedstock are necessary to release the simpler sugars needed for yeast growth. This pre-treatment often releases growth inhibiting chemicals that limit yeast growth. Notable among these chemicals is furfural, released from the heated acid-pretreatment of lignocellulosic waste biomass [31]. This thesis will largely discuss how to evaluate genetic variants linked with phenotypic resistance to furfurals, which act as broad-spectrum growth inhibitors [32].

1.3.2 Yeast metabolism

Yeast is almost unique in that it serves as a model of metabolism for eukaryotes while also being single-celled. This allows for the relatively easy modelling of a eukaryotic metabolic system using a rapidly evolving/adapting organism. Yeasts produce highly similar secondary metabolites to other eukaryotes - including humans - for example through the shared TCA cycle (see Figure 1.3.2, [33]).

Furthermore, yeasts are widely used as model production vectors amenable to protein over-expression, highly useful in research [34].

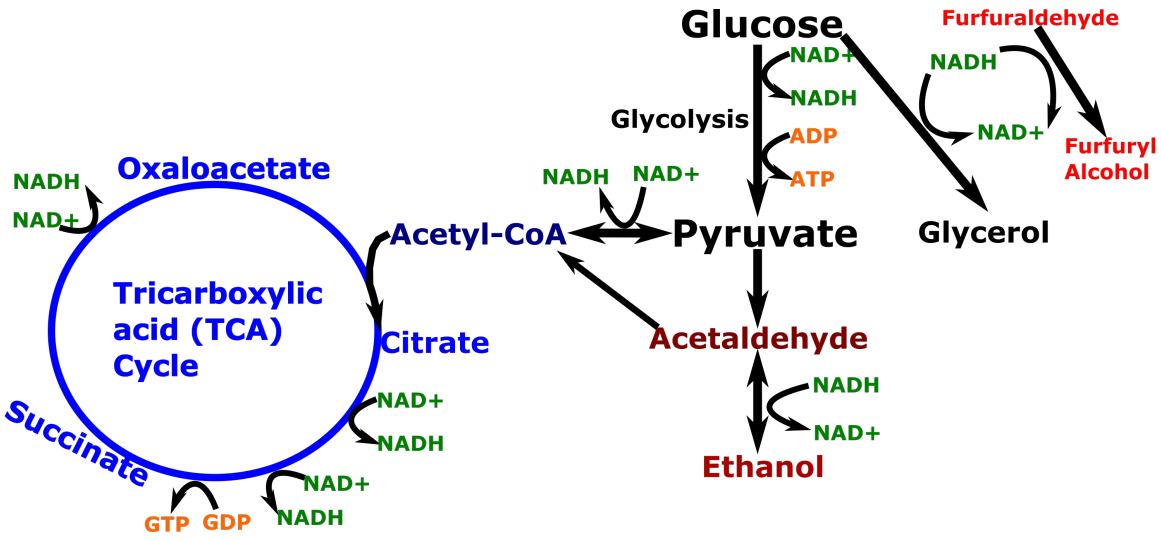


Figure 1.3.2: ***TCA cycle, glycolysis and their metabolites***
The citric acid/tricaboxylic acid cycle (TCA cycle) common to most eukaryotes on left (blue). Top to bottom shows the fermentative pathway (red). Succinate, with energy molecules ADP/ATP and GDP/GTP (orange) and NADH generation (green) illustrated on the TCA cycle which is used in hydrogenation to detoxify furfuraldehyde to furfuryl alcohol.

As can be seen in figure 1.3.2, NADH is generated through glycolysis to form pyruvate. In anaerobic, or Crabtree conditions, Ethanol formation is used as an electron donor to allow oxidation of NADH back to NAD+, which allows further energy production (ATP, orange, figure 1.3.2) from glycolysis. However, Furfuraldehyde detoxification requires NADH and is therefore a competitor for NADH. The result is that furfuraldehyde is used as an alternative sink for NADH in place of glycerol production [35, 36]. Furfural also acts as a replication limiter, but does not limit cell activity and consumes less glucose which is proportional to growth [35, 36].

In high-glucose environments, yeast behave as if they were under anaerobic conditions. Referred to as the Crabtree effect, it causes lower yields of desired metabolic chemicals due to excess carbon being shuttled to ethanol production pathways. Recent metabolomics approaches have attempted to harness yeast biology to avoid the Crabtree effect and maintain high metabolite productions in elevated glucose media to fully utilise high-energy environments. Raised levels of fumaric and malic acid appear to be the trigger that inhibits the Crabtree effect [37].

This approach is of interest to this thesis for a variety of reasons. For one, it could validate the metabolomics approach to microbiological production platforms.

Additionally, part of this thesis focusses on low-ethanol DNA variants as a field of commercial interest. Lastly, any production platform that can reduce its ethanol production and more efficiently utilise its carbon source for metabolite production is of high commercial value.

As a point of interest, could a strain be engineered to have high fumaric and malic acid production (to avoid the Crabtree effect [37]) and lowered glycerol production (to maximise furfural detoxification [35, 36])? As a point of metabolomic interest, could such a strain hold the 'key' to high furfural resistance, high metabolite production in high carbon, low oxygen environments?

While the answer is not known yet, fortunately the yeast proteome is understood to a relatively high degree of quantitative accuracy [20] which could aid any such metabolomic efforts. In regions where this well-characterised proteome's protein sequences translate directly into DNA, sequence reads derived from shotgun sequencing experiments can be mapped with high fidelity. Focussing on these genomic regions, we can zero in on mutations directly affecting protein AA composition and, therefore, likely function (assuming the mutation is not silent nor frame-shifting).

Using such as strategy, sequence reads generated or obtained within this study were mapped to a coding sequence (CDS) pan-genome constructed from 1,011 yeast strains collected across the globe [22]. In this CDS-centred pangenome, we exclude intergenic regions which may contribute more noise than signal in their mutations. This high-fidelity reference allows the accurate identification of variants in newly sequenced genomes and the prediction of phenotypic variations that might arise as a result of them, due to reliable Gene Ontology (GO) data for each yeast CDS gene. For example, a free-floating protein might have a strong sequence similarity to a known membrane-bound enzyme. Using this knowledge, the free-floating protein's function may be predicted by comparing its sequence to that of the enzyme, as depicted in Figure 1.3.3. With three bp per codon and 4 possible nucleotides in each position, there are 64 (4_3) possible nucleotide combinations. With only 21 base amino acids, less than a third of mutations are likely to affect protein sequence and, therefore, function.

This approach has both strengths and weaknesses. A common weakness is the ignoring of genetic variations outside of the coding regions. It also cannot capture ploidy or CNVs, which often lead to expression changes or phenotypic novelty. Yet

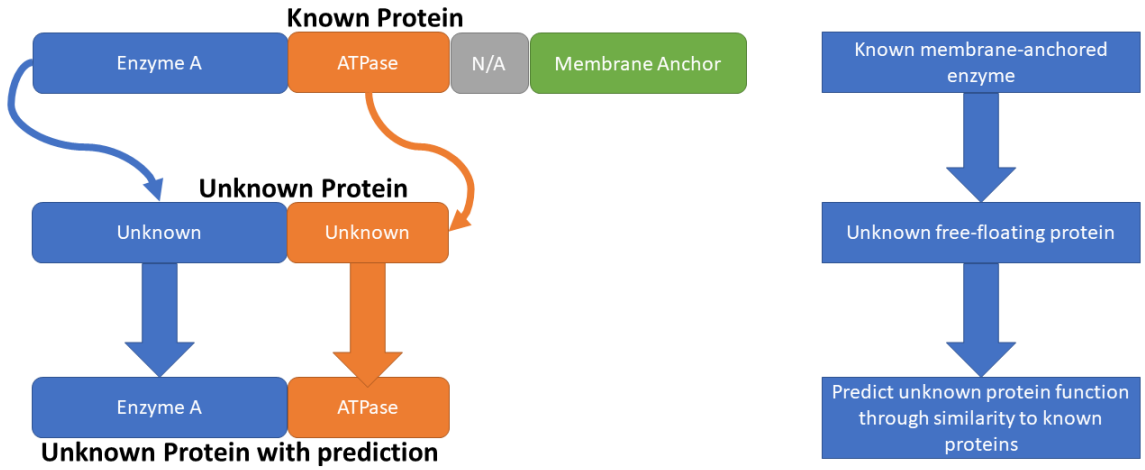


Figure 1.3.3: Process for predicting the function of a newly identified protein sequence, using the Gene Ontology (GO) of a protein with a similar underlying sequence

the technique is very effective in identifying proteins, usually enzymes involved in metabolic processes, that lead to varied phenotypes. These mutations can then be identified, effects predicted and protein variants tested *in vitro* with relative ease, particularly when compared to expression levels, for example, whose effects can be hard to measure outside the cell and its system.

1.3.3 Respiration and photosynthesis

In order to achieve renewable metabolite sources, researchers subvert respiration and photosynthesis processes. Respiration is within microbial systems a series of energy-transferring redox reactions that can be aerobic or anaerobic. In aerobic conditions, di-oxygen is reduced as a terminal electron acceptor to form water. In anaerobic respiration, the electrons are passed to other molecules or iconic elements [38].

These forms of respiration have relevance throughout this body of work. Using anaerobic respiration, yeast strains produce ethanol in a bid to rid themselves of excess electrons from respiration. Using glucose as a carbon source is simple, yet economically wasteful. Research is shifting to utilising pre-treated lignocellulosic waste biomass as the energy source of these metabolite-producing yeasts (Chapter 3). However, this shift often leads to the release of growth-limiting chemicals. Resisting these compounds that inhibit respiration is vital for higher ethanol yields

[2, 39, 40, 41, 42].

Moreover, yeast are far from the only anaerobic microbes. Many bacteria and archaea use anaerobic respiration to survive in oxygen-depleted environments. However, some employ novel electron acceptors in their respiration. For example, electrogenic bacteria use extracellular compounds and metals to accept excess electrons (Chapter 5).

Some bacteria even accept electrons as energy from their environment to form organic carbon compounds [43]; others still use photosynthesis to extract electrons from water by leveraging the energy of the sun.

1.3.4 Yeast genetics

The yeast genome is highly varied and complex, with a large accessory gene complement [44] and some even claiming the presence of entire accessory chromosomes [45]. Some yeast DNA also shows clear signs of a lateral gene transfer origin, including from bacteria. These factors result in a genomic composition that is highly 'plastic', with both SNPs and gene content variation common. Moreover, yeast genetic diversity is further complicated by chromosomal phenomena.

Polyploidy is common in yeast, especially resulting from cross-species hybridisation, and aneuploidy is frequently seen in some yeast species [45]. This broad genetic diversity is likely utilised by yeast as an evolutionary strategy in adapting to new environments [46, 47, 48]. There is also evidence that pure strains in stable environments tend towards diploidy or other more stable chromosomal arrangements [49]. For example, the *Saccharomyces cerevisiae* haploid genome is approximately 12.1 Mbp in length, with 6,611 open reading frames (OFRS; Saccharomyces cerevisiae Genome Overview | SGD (yeastgenome.org)). However, across the global population of this species, huge variation in ploidy is observed, from haploid, diploid and aneuploid, to tetraploid (common to ale yeasts) and even higher. This high degree of variation is then a great resource for both the yeast and those who wish to research it [50]. Higher ploidy levels have also been linked to faster environmental adaptations [51, 52].

At the chromosomal level, the age of a between-species hybridisation event can be estimated via ploidy levels. Meanwhile, DNA-level variations can provide evidence of closer evolutionary relationships, such as recent divergent evolution

where two sister populations colonised and adapted to different environments.

In order to reduce the influence of between-species hybridisation events from this study, only *Saccharomyces cerevisae* strains have been selected for analysis. Remaining polyploidy and aneuploidy should be mainly the result of within-species or within-strain events, which can be dealt with to a certain extant by analytical procedures. Furthermore, the focus on such a dataset then becomes DNA-level variations that have arisen from recent evolution and adaptation within a set of closely related strains. This level of evolution and inter-connectedness may be evaluated through the estimation of a Q-Matrix representing the strains' descent from a small number of founder populations (see section 2.2.4).

1.3.5 The National Collection of Yeast Cultures

The strains used in this thesis were all obtained through a collaboration with the National Collection of Yeast Cultures (NCYC) based at Quadram Institute Bioscience in Norwich. The NCYC was established in 1951 in order to provide pure, authenticated yeast strains to the biological community. An example of NCYC characterisation work can be seen in Figure 1.3.4. In the 70 years since its establishment, the collection has grown to encompass approximately 4,000 strains from over 500 yeast species, including strains of taxonomic, scientific, industrial and environmental importance. A recent effort to sequence the genomes of the NCYC strains has led to almost 1,000 strain-specific datasets, all of which were made available to this project. The largest NCYC species group, at over one-quarter of the collection were *Saccharomyces cerevisiae* strains, further reflected in the composition of the sequencing dataset. This species bias represents to a large degree the collection's origins as a brewing collection and the general utility of this species in industry.

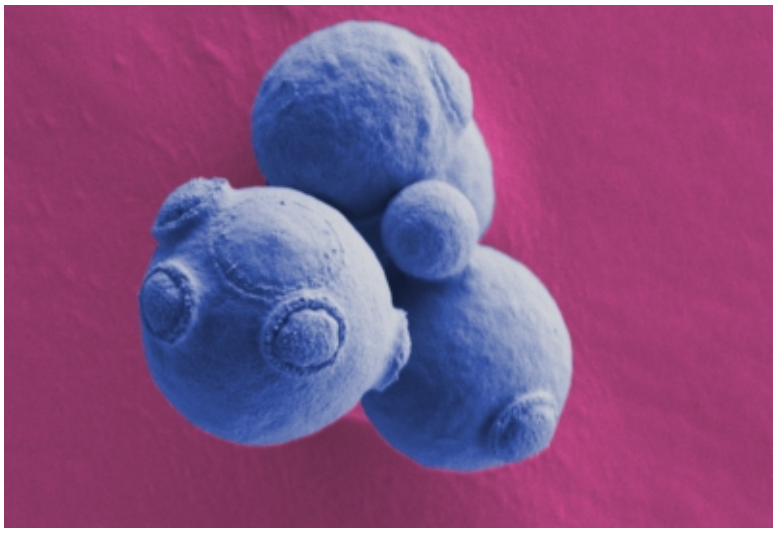


Figure 1.3.4: False-coloured scanning electron microscope image of Saccharomyces cerevisiae by Kathryn Cross and Carmen Nueno Palop, Quadram Institute Bioscience [1]. The scars left by daughter cell production by this budding yeast are easily visible.

1.4 Alternative respiration in bacteria

While the previous section discussed Yeasts' respiration and metabolism, this section will discuss bacterial respiratory systems. This is to contrast one system with another, show similarities and demonstrate where these similarities end.

Including the previously mentioned, there are many forms of respiration. These range from the classic oxygen-dependent aerobic to anaerobic and even Crabtree fermentative systems. These respirations enable yeast to produce valuable metabolites and develop techniques for resistance to growth inhibitors. Other forms of respiration involve photosynthesis and the carbon cycle. However, there exist many more forms of respiration found solely in bacteria. Many of these 'alternative' forms of respiration are undergone by specialised bacteria (or archaea) in highly specific, and often toxic, environmental niches [53].

Bacteria are a hugely varied taxonomic group and can occupy extreme ecological niches that would have been thought of as unlikely previously. The list of bacteria with unusual respiratory systems include those that colonise underwater thermal vents by reducing mercury to its elemental metal [54], those oxidising sulfur in caves [55], those which respire metals [56, 57] and many more [53]. It is almost certain that listing environments in which they do not grow is easier; bacteria are likely to grow anywhere where energy potentials exist (electrical, chemical, light, radioactive,...) [53].

This thesis will attempt to identify metal-respiring (electrogenic) bacteria as a

model for identification of alternative respiration in bacteria. The insights gained by a more comprehensive elucidation of one form of respiration in bacteria could cast a light on other forms. Moreover, the commercial possibilities of any unusual microbial respiration are not to be ignored.

1.4.1 Bacterial Genomic Properties

Bacterial genomes are diverse; there are over 50 bacterial phyla with genomes currently sequenced [58]. Bacterial genomes lack the intergenic 'junk' DNA of eukaryotes and have much more compact (and smaller) genomes ranging from a low of 112kbp to over 14Mbp, but represent a huge diversity with over 89,000 different gene families [58]. Bacterial genomes are generally smaller than their eukaryotic counterparts as genome size increases and decreases based on acquisition and loss of functional accessory gene regions [59], while eukaryotes may increase in size due to Short Tandem Repeats (STRs), the expansion of intergenic regions, or other non-coding regions or transposable elements [59]. With the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas system, bacteria possesses some resistance to viral infection that could otherwise expand and modify their genomes [58].

Bacterial genomes are also not as well-defined in structure as eukaryotes; they possess no nuclear envelope and may be composed of more than one DNA fragment [60]. Highly adaptive, they rapidly undergo purifying selection in nutrient-poor environments [59].

The bacterial genes of entire bacterial biochemical pathways are often organised into clusters, or 'operons', which work together to perform a function. These bacterial gene clusters a useful resource for researchers, as they are often self-contained biochemical pathway units.

1.4.2 Bacterial Gene Clusters

Bacterial genes are not randomly dispersed along their parent genome. Usually, genes encoding proteins in a common biochemical pathway are found clustered together as a single 'gene cluster' along one section of a bacterial chromosome. There are many possible theories proposed to answer this mystery [61].

The most common theory for the occurrence of bacterial gene clusters is that of

efficiency of co-regulation; if the genes of a single pathway are clustered, they can more easily be activated together when needed. This process is more efficient than transcribing genes from multiple genomic locations at once. This energy saving is particularly helpful if the genes interact directly [61].

A second theory states that clustering genes is for the benefit of the selfish operon; inheriting a single section of a biochemical pathway is unhelpful. By having the genes clustered closely together, it ensures that any horizontal gene transfer is more likely to include the entire functional pathway of the cluster. The increased likelihood of the entire cluster being transferred increases the adaptability of the receiving bacteria and decreases the chance of the received genes being lost to purifying selection. Therefore, this 'selfish operon model' is a close contender to the co-regulation model, if controversial [61].

The final theory states that clustered genes are less likely to be disrupted by a single mutation along the genome and are therefore more robust [61]. Whichever the true reason, and it may well be a mix of any of the three theories [62], it is clear that gene pathways tend to cluster into functional operons within bacterial genomes. The ability to identify an entire functional operon is therefore essential to understanding biological functionality. Once identified, an operon must be extracted whole; a partial operon might not function at all or could even be toxic to a cell by interfering with other biochemical pathways. Therefore, finding where the operon begins and ends is crucial to fully characterising the resultant proteins of the operon.

This whole-pathway-in-one system makes it theoretically easy to isolate an entire bacterial pathway, transfer it to another bacteria, and to subsequently mutate it for experimental analyses [63]. With this in mind, identification of bacterial gene clusters is vital to microbial research and forms a third experimental section to this thesis.

1.4.3 Gene clusters Mining Within Bacterial Genomes

Gene clusters are an excellent resource for experimental studies. The gene cluster, or operon, is a single co-regulated self-contained unit for an entire pathway. In this way, it is possible to isolate and transfer the operon as a whole entity without needing to identify disparate genes across the bacterial genome. In such an

approach, the whole gene cluster is isolated and cloned into a plasmid. Occasionally with the help of a marker gene, the presence and absence of the plasmid can be explored phenotypically in competent bacteria. Further, once the phenotype is fully characterised, mutations within the gene cluster can be explored. Identical bacteria with variants of the same plasmid can be assessed for changing phenotypic characteristics.

With this knowledge in mind, and by examining available genome sequences of a range of bacteria, this project aims to elucidate a series of operons involved in alternative respiration in bacteria (i.e, electrogenic respiration). The methodology employed could potentially be applied to other gene families to identify novel operons across the current and future bacterial genomic databases.

1.5 Project Overview

1.5.1 Project Aims

The aims of the project were varied, yet utilised similar bases of knowledge and understanding. The first aim was to discover the basis for resistance to lignocellulosic biomass pretreatment products (namely Furfuraldehyde, in our study). This was be conducted through evaluation and parsing of the wealth of genetic data available from the sequenced genomes at the NCYC. The results indicated either the genetic basis of broad-based resistance or indicate regions of interest within the genome.

Similar tools were used for the identification of the genetic basis of metabolite production in the yeast strains included in the study. For this study, Succinate was chosen as a candidate molecule of high interest to industry. Succinate, involved in ATP production within the TCA cycle, is a useful chemical used in plastics such as 1,4-butanediol, in acidity regulation in soft drinks, flavouring and even as a precursor to many pharmaceutical molecules [23].

Lastly, the project utilised similar genetic techniques to elucidate the structure of as-yet unknown cytochromes operons hidden within widely available public bacterial databases (RefSeq[64]). It was hoped that the project would uncover novel and interesting operons and aid in the understanding of cytochromes. Additionally, the project described a useful methodology applicable to a wider array of future

operons and aided in the discovery of many operons of interest.

1.5.2 Research Themes

Genome Wide Association Studies (GWAS) attempt to discover the hidden correlations between a phenotype and its polygenic features through the usage of many closely related genomes. GWAS are powerful tools for uncovering the risk scores for complex diseases such as cancer and diabetes to genes involved in increased crop yields and much more. In these complex diseases, or phenotypes in the case of this thesis, the traits being investigated are not binary and have many contributing factors. The resultant strength of the phenotype (metabolite production, furfural resistance, cancer growth,...) is based on many genomic locations and therefore must be statistically teased out of the collective organisms' genomes as a whole. To do this, GWAS correlate genomic features (CNVs, SNPs, ...) to complex phenotypes through various statistical measures such as Linear Mixed Models (LMMs). While other statistical methods exist, our focus shall be on LMMs due to both good characterisation and simple integration of random effects models(section 2.2.4).

LMMs are extended Linear Regression models which attempt to account for relationships between data points. In our study, the relationships between an independent variable (here the phenotypic data) and a dependent variable (here a given genetic variant) while also accounting for random effects (here, population effects are modelled the form of a Q-Matrix[**?**]). Given the focus of the study on furfural resistance, a single 'resistance score' was needed to quantify a strain's utility in this area.

The GWAS of this thesis makes use of vast quantities of data inaccessible to human comprehension without computational analysis. As described in Section 3, over 100,000 SNPs are correlated to nearly 200 *Saccharomyces cerevisae* strains' phenotypes. Even the phenotypes themselves are computational constructs aiming to account for the broad-spectrum effects of a lag-phase extending yeast growth inhibitor. To eliminate subjectivity in resistance phenotype assessments, and to accomplish the laborious calculations of correlating them to SNPs, computers were essential (further in section 2.2).

This same method was applied to elucidating the genetic basis of specific

metabolites produced by the yeast strains (Section 4). The technique's versatility is illustrated in its broad applications and utility.

In Section 5, computational models were developed to predict cytochrome operons hidden within the entire RefSeq database [64] of bacterial genomes. This necessitated obtaining the raw data, reformatting it to accommodate limited computer resources, curating it and then finally building the operon predictions.

All sections of this thesis made use of a range of industry-standard software [15, 65, 66, 67] and bespoke software created through two programming languages (Python3 [68] and R [69]).

1.6 Summary of thesis

This thesis will focus on many separate parts in turn. Chapter 2 will provide an overview of terms, and will introduce tools and techniques applied across the thesis. For example, it will describe which statistical methods were employed, how whole genome sequencing was performed, how the data were analysed and why specific analytical models were selected.

Chapter 3 focuses on elucidating the genetic basis of furfuraldehyde resistance in *S. cerevisae* strains. These strains were expected to have comparably high resistances, without the high variability of non-*S. cerevisae* strains. Furfuraldehyde is a by-product of the pretreatments necessary to break down the complex sugars of lignocellulosic waste biomass and is a common growth-limiting chemical in renewable platform chemical production systems [70, 71, 72, 73]. Gaining resistance to this inhibitor is crucial to attain the higher carbon conversion efficiencies necessary to produce desired metabolites from industrial microbial systems at profitable cost levels [74]. Directed Evolution (DE) experiments were also carried out to investigate the fates of SNPs with putative furfural associated phenotypic effects in an evolving population.

Chapter 4 deals with the genetic basis of various metabolite expression levels in malt media with *S. cerevisae* strains, and other metabolites in mostly non-*S. cerevisae* strains. The results will permit future researchers to select yeast stock strains based on desired metabolic profiles. The methodologies employed were custom-made to best suit the data in question and refined from previous sections of the thesis. Using the same data pipelines, this chapter correlated yeast metabolites (phenotype) to the

SNPs (genotype) of each strain.

Chapter 5 deals with the prediction of cytochrome operons in bacterial species through computational analysis of the public RefSeq CDS database. The pipeline can be applied to other operons requiring investigation. This chapter created a repository of cytochrome operons across the bacterial genome, allowing future researchers to gain an insight into which bacteria would be interesting targets for future experimental research. Many novel operons were discovered; some in completely unexpected species.

Finally, Chapter 6 discusses the work carried out and suggests future avenues of research. Drawing together the disparate elements of this thesis, it creates a holistic view of the research and elucidates the relevance to many fields of human interest, such as those of human health and medicine.

Chapter 2

Methods

There exist in this body of work many sections that call upon the same methods. These methods could either be used within the experimental or bioinformatics analyses presented. In both cases, several protocols, analytical ideas and approaches have been used many times, sometimes shared between distinct chapters, and so have been described here to retain a single point of reference. This arrangement also works to separate the theory and results of the work from the practical methods employed to explore them. This chapter may be referred to throughout the thesis.

2.1 Statistical Approaches

2.1.1 P-values for Genome Wide Association Studies

Correlation does not necessarily mean causation. This mantra is known widely even beyond the scientific community. While true, it is not the whole picture. Mathematicians have crafted statistical models that specialise in identifying the proportion of determination (R-squared value) of each variable onto an outcome (or dependent variable). There are likely innumerable specialised models utilised in various fields in science, from mouse research to climate modelling, to gather as much predictive value as possible.

While many statistical values of confidence exist to evaluate models, arguably the most well-known is the p-value. Simply put, the p-value is *the probability of obtaining a given result, or one more extreme, under a specified null hypothesis*. A very small p-value suggests that the result is very unlikely under that null hypothesis.

Unfortunately, spurious correlations are the norm when considering any dataset.

As the dataset grows in size, so too does the likelihood of a obtaining a false positive with a low p-value. This can give a false sense of significance to specific outcomes. While there are methods to reduce this false positive rate, it is always a concern that cannot be cleared until 'wet lab' research validates any correlations. Despite these shortcomings, the practice is nonetheless common within the GWAS community.

Due to the widespread use of the p-value within the GWAS community, we have decided to select it as our statistical measure of confidence, particularly as we apply it to datasets where downstream validation experiments are possible.

Some final considerations include spurious correlations and the false discovery rate (FDR), which are important to consider when using p-values for large correlation studies. Q-matrices (section 2.2.4) are employed to reduce the FDR associated with kinship relationships between species. A simple method employed here is to divide the p-value threshold (0.05) by the size of the dataset (SNP number) in a Bonferri correction. However, this can still be confounded by linkage disequilibrium [75, 76]. In conclusion, even with good data and extensive analysis, p-values simply a good method to test the Null Hypothesis (that no correlation exists) but should never be confused as concrete proof of correlation and must be experimentally validated before any conclusions can be drawn.

2.1.2 E-value for DNA Sequence Comparison

Similar to the p-value, the E-value is often misunderstood or confused. The E-value is used in pattern-recognition software such as BLAST [15]. The value represents the number of 'hits' you would have been likely to receive based on the query sequence length (m), the size of the database used (n) and the bit score (match score) 'S'. The bit score (S) increases as sequence similarity between query and target DNA increases, and is independent of total query sequence length or database size. The bit score is dependent on the Gumbel distribution (λ), the raw alignment score (s) and the scoring matrix constant (K, equation 2.1).

The equation for the E-value illustrates its constituent components in a very intuitive manner (equation 2.1);

$$E = \frac{mn}{2^S} \tag{2.1}$$

$$where$$

$$S = \frac{\lambda \times s - ln(K)}{ln(2)}$$

For example, an e-value of 1 indicates your 'hit' is likely to have come by chance as one such hit would be expected by chance from a query of the input size when matched to the size of the database used. The E-value then reduces exponentially as the query matches more closely to a sequence in the database. This is expected, as a random query sequence would not be expected to have a perfect match within a database. Any matching is likely to be relevant, with sufficient matching making it significant. Often, when selecting a 'hit', an e-value of at most 0.01 is preferred. This indicates a less than 1% chance that the 'hit' has been found by random chance matching.

2.1.3 PCA plots

A Principal Component Analysis (PCA) plot attempts to illustrate complex multidimensional data into fewer combinations of the original variables while preserving as much variation as possible. The first component displays the highest proportion of variance to assist in parsing the data and displaying it effectively, with most users analysing the first two or three components.

For example, in figure 2.1.1, a made-up scenario is created with individuals to try and identify the differences between two year-groups in a school. After many variables are measured in the PCA, two main components emerge. One component could be a combination of height and weight, while the other is of muscle strength and reaction speed. In the future, we would know that the best way to differentiate year-groups would be based on these two Principal Components (PCs). The PCA attempts to illustrate the main components to the variance in the data and thereby to group the data.

That is to say, the first component attempts to explain the most amount of variance. The second PC then explains the maximal amount of variance in an orthogonal direction to the first component (X and Y directions on a 2D plot). Every

subsequent component attempts to repeat this but for all pre-existing PCs. After the X and Y comes the Z direction in a 3D map, until we enter more dimensions (4+) than can be illustrated in a simple 3D plot.

PCA plots, in reducing dimensionality, therefore display the Proportion of Variance attributable to each component of the analysis. This is helpful, as it indicates which is the most useful component in parsing the data. It illustrates, for example, if most of the variance is shown within a single component or if the total data variation is due to many components efficiently.

Limitations with the use of PCA include hiding the importance of secondary components, while also struggling with data that has not been standardised. The benefit of reducing dimensions can also lead to the drawback of information loss if handled inappropriately. Nonetheless, PCA is a powerful tool that is widely used in research due to both speed and reliability [11, 77, 78, 79, 80, 81].

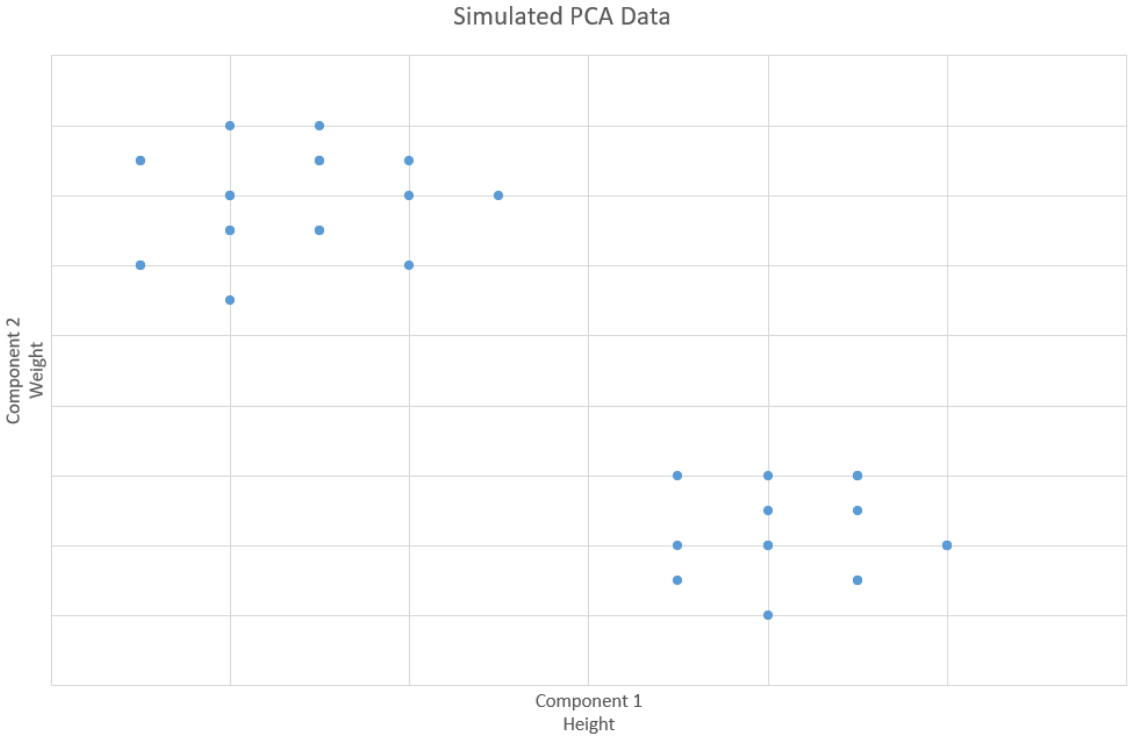


Figure 2.1.1: ***Example PCA***
A PCA plot to illustrate a multi-dimensional dataset that has been reduced to two main components. Messy data on individual samples has been reduced to 2 clear groupings on the plot which can then be investigated as a way to differentiate the population of individuals into two groups based on the components extracted.
Note: The plot is for illustrative purposes only, and does not have real-world data.

2.1.4 K-means Clustering for Phenotypic Datasets

K-means clustering is a method of partitioning *n* observations into K clusters- where each observation belongs to the cluster with the nearest mean value. In this way, it is possible to sort the observations into clusters that best reduces the within-cluster variance. With the identification of the optimal K value (i.e, the number of clusters), it becomes possible to sort the observations according to the best cluster points that already exist hidden within the data.

This technique allows researchers to parse characteristics into intuitive values on a simple linear scale. That is, a broad range of values can be grouped into a few clusters- which are easier to understand.

In some broad-spectrum phenotypes, such as resistance to a growth inhibitor, strains can exhibit highly varied resistance phenotypes and a set of scales relative to different aspects of resistance (length of lag phase, maximal OD in a time-frame,...). Using K-means it becomes possible to sum up the individual K-means cluster scores of these disparate variables into a single holistic 'resistance score' for each strain.

Figure 2.1.2 gives an example of a hypothetical phenotype that is variable among specific microorganism strains. The phenotype can then be clustered into groups that represent its variability. This low-complexity discrete variable is much easier to sum into a larger score, and tells us immediately of the strain's phenotype intensity relative to other strains. I.e, A strain that clusters in the bottom cluster is on the low end of the phenotype expression.

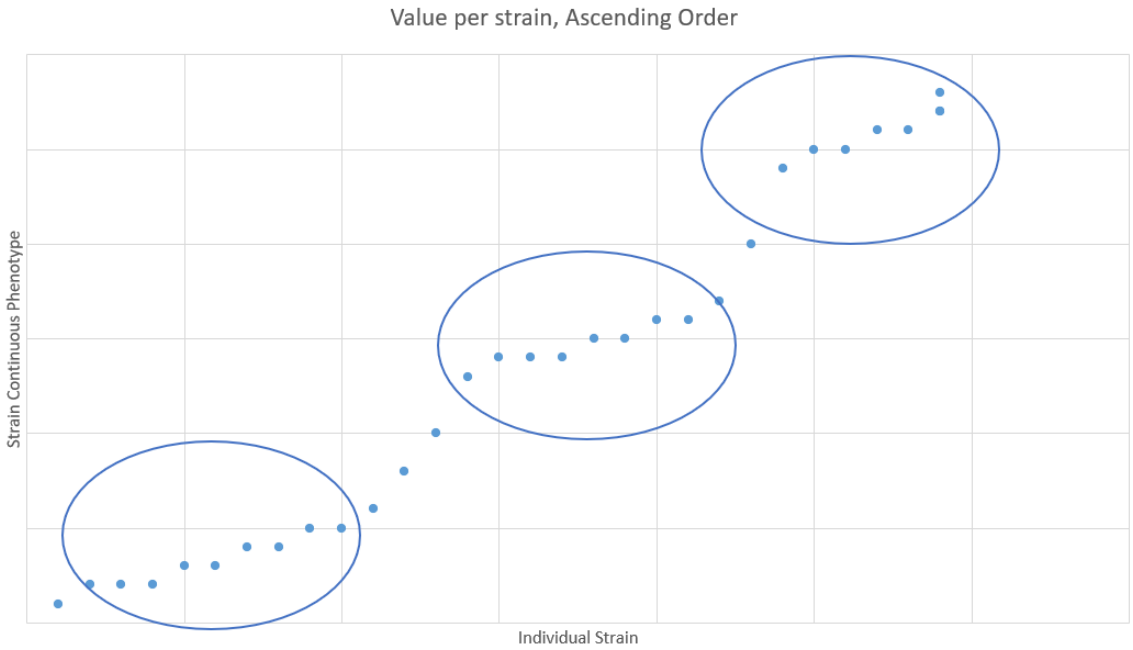


Figure 2.1.2: ***Example K-means***
An example plot to illustrate an example phenotype measured from the strains (Y-axis) per strain (X-axis). The phenotype can then be roughly divided into three clusters (optimal K=3). Whichever cluster best explains a strain's phenotype value thenceforth becomes its value (1,2,3). This is also a useful way to turn continuous variables into low-complexity discrete integers which are much easier for analysis.
Note: The plot is for illustrative purposes only, and does not have real-world data.

2.2 Genomic methods

2.2.1 GWAS and high throughput genomic assembly

Genome Wide Association Studies (GWAS) use statistical methodologies to search a moderate to large set of entire genomes for features of DNA that correlate to known phenotypes and while many rely on SNP data [12], others explore alternative genetic features such as Copy Number Variations (CNVs) [11]. Studies usually focus on SNPs within coding regions but may use other features such as gene content or non-coding SNPs [12]. Thereby, each individual genome's unique SNP fingerprint can potentially be correlated to their known phenotype(s). In humans, such a process can be used to calculate polygenic risk scores for health conditions such as diabetes and cancer. In agriculture and industrial chemical productions, association analyses can be carried out to examine traits such as livestock milk yields or crop yields [82] while genome-scale flux analysis becomes a possibility [83].

It is important to keep in mind that the study size is highly dependent on the phenotype(s) being investigated. The less clearly grouped phenotypes are between

conditions, the greater the number of genomes necessary to obtain the required statistical power to fully isolate the genomic variants responsible for the phenotype. For example, the genomic variants responsible for dwarfism are easier to isolate than the genomic variants causing a height difference of a few centimetres.

In the experiments carried out here, such a broad outlook was taken. The genomes of the species undergoing analysis were converted to SNP matrices which could subsequently be correlated to phenotypic datasets such as the production of specific metabolites and resistance to growth inhibitors. These results could, hypothetically, be carried forwards in genetic modification experiments with the strains used, in order to evaluate the functional effects of SNPs highly correlated to phenotypes of interest.

Such a high throughput genomic pipeline could be used to create models to predict the phenotypes of sequenced but as yet uncharacterised yeast strains or species. This would enable researchers to prioritise the experimental testing of strains that carry a higher likelihood of possessing desirable phenotypes.

2.2.2 SNP genome

The SNP genomes for each strain were assembled using a highly conservative variant calling pipeline (section 2.4.3) coupled with other usual genomic tools. This ensured a resultant SNP genome of very high fidelity and low false positive rates- increasing the likelihood of relevance of any top hits.

The genotypes of the strains are thereby transformed to a list of SNPs. The SNPs were selected if they were present as alternative alleles in at least 5% of the strains, known as the Minor Allele Frequency (MAF) percentage, in the study. The SNPs were then assembled into a linear genome that matched loci (ORFs) on the artificial yeast pan-genome assembled from 1,011 yeast strains' collected worldwide [22].

2.2.3 Linear regression and Linear Mixed Models

P-values were obtained through linear regression models known as Linear Mixed Models (LMMs) that incorporate relationships between phenotypic and genotypic data, while also accounting for relationships between yeast strains that could otherwise produce spurious correlations. In this way, we account for evolutionary relationships between strains through Q-Matrix models. Once the SNP genome of

each strain has been determined, LMMs are used to then correlated the phenotypes of the strains to their SNPs.

Our analyses use a Linear Mixed Effects Regression (LMER) model. This is a Linear Regression that takes into account both the fixed and random effects of the system phenotypes from the dependant variable of the model, with SNPs modelled as fixed effects and inter-strain relationships as random effects. This latter part is crucial to achieving reliable results.

Even in using 'only' *Saccharomyces cerevisiae* strains in the input dataset, there are many micro (SNP/gene level) and macro (chromosomal level) genetic variations between strains. It is therefore crucial to attempt to account for this by quantifying the strains' evolutionary relationships. This is accomplished through incorporating the Q-Matrix (section 2.2.4) into the LMER model.

This Q-Matrix might raise or lower p-values of specific SNPs when compared to a LR model. If a variant is found highly correlated to a phenotype in a LR, its p-value may be significantly reduced in cases where the Q-Matrix shows a high correlation between sub-population and a given phenotype. This is because the sub-population may share variants that other sub-populations do not, thereby making it less likely that a specific variant was causative of the phenotype in question. However, if the variant is found across all sub-populations (via the Q-Matrix analysis) but always correlates well with the desired phenotype, its p-value will be raised significantly as its effect transcends the sub-population level.

2.2.4 Q-Matrix Automation

Within any GWAS analysis (or any correlational study), there exists the issue of false positives. This is exacerbated in studies with highly related genomes that are likely to share a greater proportion of SNPs by virtue of genetic relatedness, obscuring the genomic causes of an investigated phenotype. Without taking relatedness of genomes into account, a straight forward correlation of genotypes to phenotypes would present many erroneous results that drew on kinship relationships instead of the desired phenotypic groupings.

The Q-Matrix is a mathematical structure that attempts to measure the relatedness of strains within a SNP dataset. This is accomplished by using the diversity within the SNP dataset to predict the number of founder populations

necessary to give rise to the distribution of variation seen in the dataset. The Q-Matrix scores relatedness by attributing proportions of each strain's genome to a founder genome. By using the Q-Matrix within an LMER analysis, this allows the GWAS to exclude many correlations erroneously formed through the Founder Effect.

The Q-Matrix in this analysis is of three predicted founder populations and the fraction of each strain's genome attributable to each founder. In a simplistic example, a group of strains might be predicted to have evolved from three ancestral populations (table 2.1). A strain in that group might incorporate all parts of this ancestral history in its genome and owe 20% of its genome to one ancestor, 60% to a second and 20% to a third (strain 9, table 2.1). A strain could also have its entire genome (i.e. 100%) attributable to a single founder (strain 2).

Strain Number	Founder 1	Founder 2	Founder 3
1	0.4	0.4	0.2
2	1	0	0
3	0.8	0	0.2
4	0.3	0.4	0.3
5	0.1	0.9	0
6	0.1	0.1	0.8
7	0.5	0	0.5
8	0	0.1	0.9
9	0.2	0.6	0.2
10	0.6	0.3	0.1

Table 2.1: ***Q-Matrix Example***
An example of a Q-Matrix, with predicted founder populations (top) and strain numbers (left). This shows us the predicted distribution of each strain's genome into the 3 predicted founders.

Any strains that share a similar distribution with respect to their ancestry are less likely to have shared SNPs counted as significant when correlated to a phenotype. This is because it becomes difficult to disentangle chance correlation between highly related species and actual causative SNPs. In our study, the SNP data for the Q-Matrix underwent two rounds of curation; The first involved reducing any false positive rates by using very conservative CIGAR strings with the FAT_CIGAR tool (section 2.4.3). Secondly, any 'unknown' genes annotated in the reference pangenome were removed. This ensured any 'accessory' genes seen in only a few species were not counted and did not skew our analysis. Further, any genes listed as part of the 'core' genome in the reference pangenome were kept while other genes

were discarded. This was to permit the measurement of evolutionary history in only highly-conserved genomic regions. All these measures attempted to best quantify the 'relatedness' between strains through a single Q-Matrix set of values with a hypothetical number of founder populations.

2.2.5 PSIKO Q-Matrix

The PSIKO [84] Q-Matrix is a validated algorithm [85, 86] for the prediction of strain relatedness through a Q-Matrix. PSIKO allows for the inference of admixture coeffecients through a combination of linear kernel-PCA and least-squares optimisation. Much faster than most alternative Q-Matrix prediction software, it is ideal for the analysis of very large next-generation datasets [84].

2.3 Yeast Experimental Methods

The following section concerns the specifics of the methodology used in this thesis for chapters 3 and 4.

2.3.1 Strain acquisition, preparation and storage

The yeast strains used in the study were provided by the National Collection of Yeast Cultures (NCYC) within the Quadram Institute Bioscience (QIB) in the Norwich Research Park. Each strain has both a corresponding accession number and taxonomic designation in the NCYC system [1]. Two separate growth medias were used within experiments using the NCYC yeast strains, as stated.

Unless otherwise stated, and in all furfural experiments in chapter 3, a minimal Yeast Nitrogen Base (YNB) media was used. This consisted of 6.9g/L YNB media with 10g/L glucose as a carbon source. Furfural concentrations were varied, as stated in each experimental section. When maltose media was utilised, the media contained 100g maltose extract per litre of purified water. Used to replicate conditions in breweries, it was used for metabolomic analysis of the yeasts' products in chapter 4.

2.4 Yeast Whole Genome Sequencing

Following the directed evolution experiment in chapter 3, whole genome short-read sequencing of isolated single-strains was carried out with an Illumina NextSeq [8] sequencer to determine the level of evolution and adaptation to the sequential furfural conditions. This was to continue a project already underway to fully sequence all the NCYC strains with the same short-read Illumina technology and to ensure comparability. Strains were plated on YNB agar with 1mg/mL furfural. Single colonies were extracted and then grown in sterile YNB media.

Whole genomic DNA of each strain was sent for sequencing at the Quadram Institute. DNA extractions were performed according to a NCYC yeast DNA extraction protocol (personal communication, Adam Elliston). The main kit used for the extraction was the Masterpure Yeast DNA purification Kit that contained TE buffer, Cell Lysis Solution and MPC protein precipitation reagent. Zymolase and RNAase A were purchased separately.

2.4.1 NCYC Yeast DNA extraction protocol

Figure 2.4.1 describes the NCYC Yeast DNA extraction protocol modified cosmetically where appropriate to reflect the specific experiment.

1. Strains were grown at room temperature in their YNB media for 3-5 days.
2. 1 mL was taken from each of the 8 wells in the first plate with 7 (excluding the repeated control H) for a total of 15 samples. This was duplicated into 2 technical replicates for a total of 30 samples.
3. Cells were pelleted in 14K rpm for 5 minutes.
4. The supernatant was discarded and the pellets frozen at -20°C.
5. 100 μL of zymolyase (10 mg/mL) was added to each pellet.
6. Samples were incubated for 30 minutes at 37 °C.
7. Samples were centrifuged for 5 minutes at 14K rpm.
8. The supernatant was discarded again and 300 μL of Cell Lysis Solution added with 5 μL of RNAse A.
9. Cells were resuspended by gentle vortexing.
10. Cells were incubated at 65°C for 15 minutes.
11. Samples were cooled on ice until the next step for 5 minutes (maximum 1 hour).
12. 150 µL of MPC protein precipitation reagent was added, then the samples were vortexed again.
13. Cell debris was pelleted away by centrifugation at 14K rpm for 5 minutes.
14. Supernatant was transferred to a clean Eppendorf tube and 500 µL of ice cold isopropanol was added.
15. Samples were centrifuged at 14K rpm for 10 minutes to obtain DNA pellets.
16. Supernatant was discarded and 500 µL of ice cold 70% ethanol added.
17. Samples were centrifuged at 14K rpm for 5 minutes.
18. Supernatant was discarded and tubes aired for 5 minutes on the bench to allow any residual ethanol to evaporate.
19. 35 µL TE buffer was added once dry and samples were left in the fridge overnight for DNA to dilute.

Figure 2.4.1: Protocol for growth and DNA extraction of yeast samples.

2.4.2 Raw Read cleaning and trimming

Data was cleaned and trimmed with trimmomatic-0.32. In the DE experiments, this step was preceded by the concatenation of two technical replicates. Trimmomatic [87] is a tool used to clean up Next Generation Sequencing (NGS) reads by trimming NGS (usually Illumina) adapter sequences in a read preprocessing step and removes low-quality regions, particularly of the 3' region of Illumina reads [87]. In essence, the tool attempts to remove artefacts of current NGS sequencing technologies. The removal of duplicate reads, that can occur during the library preparation process, was achieved through BBTools [88].

Once the raw reads data have been cleaned, they are ready to be processed in the next step, mapping to a reference genome. Below (figure 2.4.2), we see the steps involved in trimming the data appropriately. In step 1, we can concatenate technical replicates into a single file. In 2, we remove duplicate reads. In 3, we use trimmomatic to trim the reads.

1. cat Raw-reads/S1-R1.fastq Raw-reads/S2-R1.fastq > Raw-reads/S1-R1.fastq
   cat Raw-reads/S1-R2.fastq Raw-reads/S2-R2.fastq > Raw-reads/S1-R2.fastq

2. clumpify.sh in1=Raw-reads/S1-R1.fastq in2=Raw-reads/S1-R2.fastq out1=Raw-reads/S1-dedup-R1.fastq out2=Raw-reads/S1-dedup-R2.fastq dedupe subs=0

3. java -jar Trimmomatic-0.32/trimmomatic-0.32.jar PE Raw-reads/S1-dedup-R1.fastq Raw-reads/S1-dedup-R2.fastq Deduplicated-and-trimmed-reads/S1/S1-forward-paired.fastq Deduplicated-and-trimmed-reads/S1/S1-forward-unpaired.fastq Deduplicated-and-trimmed-reads/S1/S1-reverse-paired.fastq Deduplicated-and-trimmed-reads/S1/S1-reverse-unpaired.fastq ILLUMINACLIP:Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

Figure 2.4.2: Commands used to Trim raw reads from sequenced DNA; clumpify.sh comes from BBTools.
subs = 0 sets 0 substitutions
LEADING:3 - Removes any bp on 5' end with less than 3 phred score
TRAILING:3 - Removes any bp on 3' end with less than 3 phred score
SLIDINGWINDOW:5:15- removes the 5 bp on the 5' end of read if the average phred score is lower than 15 (low quality read segment). This often is used in place of LEADING/TRAILING options
MINLEN:36 sets the minimum read length to 36bp

2.4.3 Mapping cleaned reads to artificial genome

Once the data has been preprocessed by Trimmomatic [87], it is ready for the reads to be aligned to the reference genome. For this step, Burrows-Wheeler Alignment (BWA) [67] was used (BWA mem). This was to make them more directly comparable with other existing data pipelines. This step allowed the creation of SAM (Sequence Alignment/Map) files which were duly converted to Binary Alignment/Map (BAM) files. The BAM files were then sorted and indexed. The pipeline was designed to be adapted to many various analyses; if files were no longer needed, they were deleted to save space.

The resulting files (BAM, Sorted BAM and Index BAM) were subsequently utilised by FreeBayes [66] to create Variant Call Format (VCF) files by mapping each genome onto the reference (in our case artificial) genome and identifying differences from it using a statistical model.

The resultant file was then ready for analysis through custom scripts and software (figure 2.4.3). In step 1, a SAM file is created, showing the alignment of the reads to the reference, and is converted to binary format in step 2. In step 3, index files are created. In step 4, the FAT-CIGAR bash script removes reads from the sorted BAM file that do not perfectly map to the reference genome at both ends (for a user-specified length, 20bp in this case), thereby reducing false positive variant calls in downstream analyses. Finally, in step 5, FreeBayes is used to call variants from the reference genome in VCF format.

1. bwa mem allORFs-pangenome.fasta Deduplicated-and-trimmed-reads/S1/S1-forward-paired.fastq Deduplicated-and-trimmed-reads/S1/S1-reverse-paired.fastq > BWA-alignments/S1.sam
2. samtools view -S -b BWA-alignments/S1.sam > Samtools-BAM-files/S1/S1.bam
3. samtools index Samtools-BAM-files/S1/S1.sorted.bam
4. FAT_CIGAR script
5. freebayes -f allORFs-pangenome.fasta Samtools-BAM-files/S1/S1.sorted.bam > freebayes-output/S1-freebayes-SNP-genome.vcf

Figure 2.4.3: ***Bash Commands for VCF creation***
Commands for creating VCF files from Trimmed and deduplicated DNA read files. FAT-CIGAR script used was a pre-publication version (Prithika Sritharan, personal communication); please see https://github.com/prithikasritharan/FAT-CIGAR for the most recent version.

In a final data cleaning step, three R packages were used (snpStats, VariantAnnotation, GenomicFeatures in a bespoke script CreatingGWAS_Data.r) to further filter the data in RStudio. This filtered the data based on quality, and mutation type by only selecting for high-quality, binary SNPs. This resulted in a high-quality SNP genome for each original Illumina dataset.

2.4.4 SNP Matrix

The data of each SNP 'genome' in a study was joined together into large matrices using custom Python3 scripts SNP_and_MAF_finding.py followed by MatrixMaker.py. This allowed for easier future data accession and for correct indexing of all mutations with a Minor Allele Frequency of 5% or more (i.e, if an SNP is present in more than 5% of genomes it was added to the matrix, with any genome not possessing the SNP assigned the reference allele). This dataset was then accessed rapidly by any correlation pipelines designed with custom scripts using RStudio packages.

To construct the data frames necessary for RStudio [69], Bioconductor and Python3 [68] were used (scripts above). This data from Python3 was also used to create evolutionary kinship relations (Q-Matrix) based on genetic distance using PSIKO [84] (section 2.2.4).

The basic pipeline is; CreatingGWAS_data.r, followed by SNP_and_MAF_finding.py, then the second segment in CreatingGWAS_data.r and lastly MatrixMaker.py. The work was split into segments to help compartmentalise processes.

2.4.5 Correlating reads to phenotype

Final SNP genome matrix matching to phenotype was accomplished through a custom RStudio script ManhattanPlotGeneration.r which allowed allowed each SNP within the matrix of many genomes to be matched to the phenotype of each strain. This script allowed for the output of data in Comma Separated Values (CSVs) format as well as easy-to-understand Manhattan plots for ease of conceptual comprehension.

2.5 Metabolic growth protocol

For metabolomic data, each 20µL aliquot of the strains were grown in 96-deepwell plates with a total volume of 1mL as previously. The plates were sealed with breathable seals in anaerobic conditions and incubated at 25°C for five days. This was performed to achieve a final growth spectrum, with complete utilisation of glucose without entering cell death due to age and/or lack of nutrients.

The supernatant media was then analysed through NMR to quantitatively measure the metabolic products of the yeast. This allowed us to compare the output of each strain; any error in the buffer was removed by using a single master mix.

2.5.1 NMR Preparation and data analysis

The strains, once in their final state after 5 days of anerobic growth, were spun down at 3000rpm for 15 minutes. The pellets were discarded, with 400µL of supernatant added to a new 1mL deepwell plate. The 400µL of supernatant was subsequently mixed thoroughly with 400µL of Nuclear Magnetic Resonance (NMR) buffer (section 2.5.2). When thoroughly mixed, the plate was re-spun at 3000rpm for 15 minutes. Finally, 600µL of supernatant from each well was taken and placed in NMR tubes.

Following 500mHz proton-NMR, the data was curated with TopSpin (command: apk0.noe) and analysed quantitatively with CHENOMX Profiler software. This allowed, using TSP as a reference peak of known concentration, the quantification of many metabolites (figures 4.2, 4.3) in the sample. Additionally, NMR allowed us to return to the raw data at will to add to our list of analysed metabolites.

The strengths of NMR include high reproducibility, a broad range of molecule detection, the sample being unaffected by the analysis, with sensitivity (mmol) that can be increased with a higher field strength and highly accurate for smaller inorganic molecules (as in this study with ^{1}H resonance).[89].

2.5.2 NMR buffer

The Nuclear Magnetic Resonance (NMR) buffer contained:

- NaH2PO4.H20 42g/L
- K2HPO4 16.5g/L

- 0.5mM TSP 86mg/L
- Sodium Azide 200mg/L
- 100mM EDTA 1mL/L for final 100nM solution (100 mM EDTA-0.372g in 10mL D20)
- Made up to 1L with D2O

The Sodium Phosphate Hydrous and Di-Potassium Phosphate Anhydrous were pH buffers to balance the pH to roughly 7pH. The EDTA ensured free metal ions were bound and removed in the centrifuging step to limit future NMR interference/noise. Sodium Azide acted as an anti-microbial agent to ensure growth had stopped in the media. D2O acted as the 0-point of the spectra. Finally, the TSP (Trimethylsilylpropanoic acid) was used as a reference peak for NMR analysis which would allow the quantification of other metabolites based on the TSP peak intensity with a known concentration.

2.5.3 YNB Media

When YNB is referred to within the context of this thesis, the following recipe shown in table 2.2 is to be kept in mind. The YNB product code was discontinued from Formedium, but any YNB without glucose (or other carbon source) but containing vitamins, amino acids and minerals should be sufficient to replicate the media.

FORMULA	FINAL CONTENT (mg/L)
Histidine HCl	10
Methionine	20
Tryptophane	20
Biotin	0.002
Ca-Panthotenate	0.4
Folic acid	0.002
Inositol 2 Nicotinic Acid, (Niacin)	0.4
p-Aminobenzoic Acid	0.2
Pyridoxine HCl	0.4
Riboflavin	0.2
Thiamine HCl	0.4
Boric Acid	0.5
Copper Sulfate	0.04
Potassium Iodide	0.1
Ferric Chloride	0.2
Manganese Sulfate	0.4
Sodium Molybdate	0.2
Zinc Sulfate	0.4
Potassium Phosphate, monobasic	1000
Magnesium Sulphate. anh	500
Sodium Chloride	100
Calcium Chloride.anh	100
Ammonium Sulphate	5000

Table 2.2: Yeast Nitrogen Base media components- Formedium product code CYN02.

Chapter 3

Predicting and Increasing Furfural Resistance using GWAS Approaches

3.1 Furfurals explanation- a lignocellulosic metabolic hurdle

3.1.1 The microbial challenge

As finite oil reserves deplete and governments mandate the usage of fuels and chemicals derived from bio-sustainable sources, the case for microbial platforms for metabolite production is strong [90]. Recently, the focus of these platforms has broadened from the initial replacement of fossil fuels to include fossil-fuel derivatives (e.g. plastics, succinate, tar), such that a wide range of metabolic products are becoming increasingly viable [91]. Oil-derived chemicals can be highly long-lasting and polluting, particularly in the case of plastics. The development of biodegradable plastics, derived from non-fossil sources, is therefore of high scientific interest. For example, microbially-derived and biodegradable polymers such as polyhydroxyalkanoates (PHAs) have been identified as potential alternatives to oil-based plastics [91] and are the focus of a growing number of initiatives.

A key aim of sustainable bio-production efforts is to use plant biomass as a biochemical production vehicle. Such a system draws together farmers, microbiologists, brewers, and many other disciplines, to enable the conversion of live biomass into biofuels and other platform chemicals. First generation biofuel

programs used edible crop seeds as feedstocks. Easily digestible by microorganisms and rich in sugars, seeds provide an ideal environment for microbial growth. Unfortunately, while such a microbial system efficiently converts simple sugars into the desired platform chemicals, it was also found to have deleterious societal impacts. When farmland is dedicated to producing feedstocks for industrial processes, the land (and crops) cannot simultaneously provide sustenance for the human population. As a consequence, farmland for food becomes scarcer and food prices increase. Additionally, farmers in poorer countries encounter the dilemma of feeding their populace or earning more money by selling their cash crops to overseas biotech companies. As farmland is occupied by bio 'cash' crops, a hungry populace would need to plant new acreage; deforestation might then become the temporarily optimal path to food security [2, 40, 92, 93, 94].

These factors, among others, made it difficult to permit the usage of foodstuffs as feedstocks for biochemical production. Other methods were then required that utilise renewable, sustainable sugars without impacting current food supply-chains. As such, the focus of biofuel and biochemical production efforts shifted towards the use of plant biowaste, often the inedible parts of plants produced for food. Lignocellulosic waste, lignin- and cellulose-rich biomass from inedible parts of plants such as the stalks, has subsequently become of high interest to industry as a carbon source for industrial fermentation [40, 95].

3.1.2 Furfural Production Through Pretreatments

When attempting to utilise lignocellulosic waste biomass in bioproduction processes, it is necessary to extract the full range of sugars found therein. The stalks, leaves and other inedible sections of plants are composed of cellulose, hemi-celluloses and lignin arranged in rigid structures. Within these structures, the sugars are bound tightly together and are therefore inaccessible for most microbes. Consequently, it is necessary to break down these strong bonds and release the carbon necessary for the microbes to grow and produce their platform chemicals [96].

To reach the more-accessible cellulose, it is first necessary to solubilise the lignin and hemi-celluloses. This is often done through heat and/or acid pre-treatment, efficient but expensive techniques [40, 97]. The acid, and elevated temperatures,

hydrolyse the release of sugars such as xylose and pentose. Other well-known sugars released by such pretreatments include monosaccharides such as glucose and galactose, along with their respective disaccharides and oligosaccharides. These sugars are then metabolised by the microbial production platforms to produce the desired platform chemical metabolite [97].

However, pre-treatments can also release growth-inhibiting chemicals that slow microbial growth and, as a consequence, decrease metabolic outputs [31]. In particular, lignin degradation releases antimicrobial phenylic compounds. While microbes are known to digest some of these compounds, such as benzoic acid [98], others such as cinnamic acid are digestible only by a limited range of microbes, for example non-brewing yeasts [99]. Other phenolic compounds (coniferyl aldehyde, ferulic acid and 4-hydroxybenzoic acid) are well known microbial inhibitors [100].

Taken as a whole, lignin degradation is highly inhibitory to microbes such as yeast. In addition, the xylose and pentose sugars released in the acid catalysis of the hemicelluloses and cellulose can be dehydrated to form furan compounds, including furfuraldehyde, hydroxymethylfurfural (HMF) and furoic acid [40]. These last compounds are of particular concern due to their broad-spectrum inhibitory effects.

In summation, heat and acid are necessary for the hydrolysis of complex bonds in the inedible parts of plants into simple sugars. However, this pretreatment also releases various inhibitory factors, ranging from phenylic compounds from the plants' lignin, to furans from the same sugars it was designed to release [97, 101]. As such, the identification or development of microbial strains capable of both utilising the extracted sugars while also being resistant to the wide range of inhibitory factors is crucial. This section of the thesis focusses on the discovery of yeast strains resistant to the inhibitor furfuraldehyde [97, 101, 102].

3.1.3 Lignocellulosic Pretreatments and Furfural Origins

As previously mentioned, heat and acid pre-treatments are often used for the catalysis of complex starches into simple sugars. The sugars are subsequently converted by microorganism production platforms into various metabolite chemicals of interest. However, this pre-treatment also causes the release of growth inhibitors, such as furfuraldehyde.

As we proceed it is important to fully understand the function of furfuraldehyde (and other furfurals). Furfurals are lag phase extenders for yeast growth, inhibiting growth through a broad spectrum of effects. They are known to damage DNA, proteins and membranes through an increase in Reactive Oxygen Species (ROS) as well as increasing the yeast cells' sensitivity to osmotic stresses [103, 104]. As such, they cause great stresses on the yeast cells. This is exacerbated by the fact that the yeast cells expend NADH as an electron donor to reduce the furfural to the less-toxic furfural alcohol anaerobically in fermentation reactions, as shown in figure 3.1.1 [101, 105].

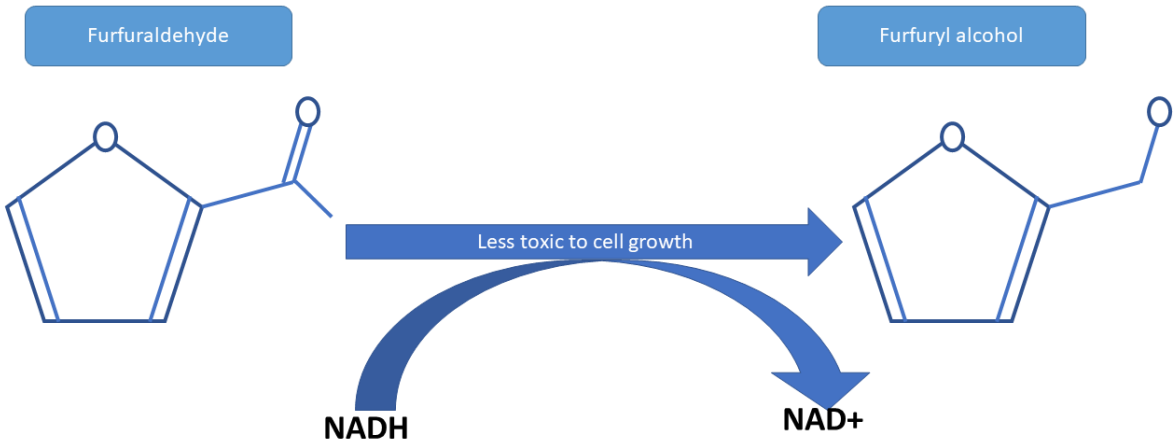


Figure 3.1.1: Furfuraldehyde detoxification through NADH-mediated hydrogenation of furfuraldehyde.

This NADH sink means that a smaller pool of NADH is available for growth and replication and therefore the stress caused by furfurals is increased. A resistance to the effects of furfurals might therefore involve a host of methods such as; DNA protective measures, increased NADH availability for detoxification of furfurals, membrane changes to shuttle more furfurals out/ protect from osmotic stresses as well as other broad-spectrum ROS-mediating proteins [101, 106].

To produce the aforementioned sugar-rich feedstocks necessary for industry, lignocellulosic waste is often treated in a myriad of ways [96]. One common method is to provide sufficient heat and acid to break up the long inaccessible carbohydrate chains into free-floating sugars for use by the microbial production platforms. Unfortunately, due to the elevated temperature and acidity, factors antagonistic to

yeast growth are produced from the hemicelluloses, glucoses and xyloses. In the conditions necessary to release the sugars, such as diluted sulphuric acid coupled with high temperatures, the xylose and glucose released can be dehydrated to hydroxymethylfurfural (HMF) and other furfurals. This makes furfuraldehyde a useful model inhibitor of yeast growth where furfuraldehyde represents an exogenous inhibitor to which yeast will have adaptated little [7, 32, 101, 107].

A resistance to the toxic and lag-phase extending effects of furfural, which is present in much treated lignocellulosic waste, is therefore of significant importance for industry. Production of any chemicals using yeast as a production vehicle and pre-treated waste biomass as a feedstock is greatly reduced as the cells struggle to resist the damage from furfural even as they sink significant amounts of NADH to detoxify it [101]. Often, to find a mechanism for resistance to chemicals, researchers focus on a single, promising biomolecular pathway or enzyme. Due to furfuraldehyde's unspecific wide-ranging effects [106], it is difficult to accomplish this within an individual study of distinct genes. This difficulty might be attributed to furfuraldehyde not being present in 'natural' environments where yeast are found, so there is no specific naturally evolved stress response.

3.1.4 Investigating natural and forced resistance to furfurals in yeast strains

In this chapter, experiments to understand the genetic basis of furfural resistance in Baker's yeast *Saccharomyces cerevisiae* will be carried out. First, a set of over one hundred *S. cerevisiae* strains obtained from the NCYC are tested for their ability to grow in the presence of furfuraldehyde. Then, using various statistical models, the contributing factor of each single SNP is measured against the strain's overall resistance to furfural. In the manner of table 3.3, each SNP's variation profile across the entire strain set is matched to that of the computed resistance score. The higher the correlation between the two, the lower the p-value of the SNP and the likelier it is to be causative for the phenotype. In the second experiment, the results of the first study are used to design and conduct a directed evolution experiment where both single strains and strain mixes are challenged with furfural doses in the expectation that evolved strains will show high resistance to this growth inhibitor. The prevalence of key SNPs will be compared between the two datasets. Prior to

describing the experiments, two new computational methods developed and used within this work will be introduced, for both phenotype classification and yeast genomic structure analysis.

3.2 A Novel, Automated Method for Quantifying Resistance Phenotypes from Yeast Growth Curves

3.2.1 Preparing strains for OD growth curve analysis

For OD growth analysis, a 20µL aliquot of each strain was grown in 96-deepwell plate with a total volume of 1mL. Plates were sealed with breathable seals in anaerobic conditions and incubated at room temperature for three days. This allowed for a suitable stock for future aliquot to be produced, while also providing sufficient supernatant for metabolite profiling via NMR analysis.

Once three days of growth had been achieved, a 20µL aliquot was removed from each well and placed in a new 96-well micro-plate with a final volume of 200µL. The strains were then grown for 24 hours at 25°C in a FLUOstar Omega plate reader (BMG Labtech) with a 600nm OD reading taken every 30 minutes and used as a biomarker for cell growth and density. OD values were used as a measure of cell biomass/number and therefore a high OD value would be expected to correlate with higher resistance.

3.2.2 Measuring resistance phenotypes

To acquire the resistance score/growth phenotype for each yeast strain from the described OD readings, a consistent, automated way to measure resistance was necessary. Notably, the method developed was general and therefore could be used for any growth inhibitor, rather than being limited to furfuraldehyde.

When graphs of the 49 half-hourly OD readings are compared visually, it can be easy to discern a highly resistant strain (fast growth leading to high cell biomass) to a highly sensitive strain (very little or no growth). However, it can be difficult to assign values to resistance on a discrete linear scale; this is especially true when comparing medium-resistance strains. This task becomes increasingly subjective, such that different people could classify the same strain as differently resistant.

For example, looking at the 96-well plate in figure 3.2.1, we can see how difficult it would be to discern levels of 'resistance' between strains. To solve this issue, computational analysis of the OD growth curves is necessary.

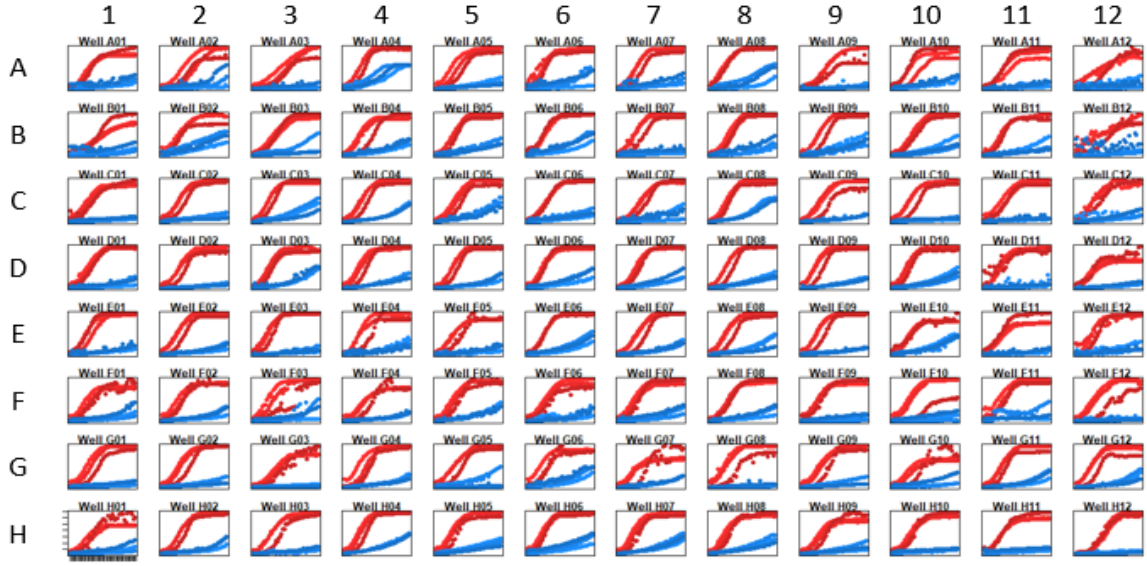


Figure 3.2.1: 96-well plate example, with growth in furfural media (blue) and growth in control YNB media (red) shown with full biological triplicates superimposed onto each relevant well. Displays plate '1' from figure 2 in the appendix. Control well without strains (H12) replaced with Plate 1 well H12 from separate excel growth file. The graphs show, for example, that the yeast strain in well A4 grows well in the presence of furfural whereas the strain in well C1 does not.

3.2.3 OD growth curve feature selection

To analyse the growth curves computationally, it was first necessary to determine which features of growth could be calculated algorithmically. The Maximal OD value of the growth curve (MaxOD) was the obvious first characteristic and easiest to measure. As the name indicates, it is simply the highest Optical Density reading of a given growth curve. Next, the timepoint along the curve at which its slope μ is highest was calculated. To find the various values for μ along the curve, several 'windows' of 9 time points (representing a four-hour period) were taken across the growth curve. For each window, a linear regression of the 9 OD values was conducted, and the slope μ attributed to the 5th (middle) timepoint. The 37 different values of μ were then compared, with the timepoint at which the highest slope (μ_{max}) was observed denoted as $T_{\mu_{max}}$ (Equation 3.1a).

The Inflection Point of the curve was also measured. First, the average of the 2nd, 3rd and 4th OD values (i.e. 0.5hr, 1hr, 1.5hr timepoints) was calculated, creating a 'baseline' minimal OD value (OD_0) that ignored the initial, highly variable first

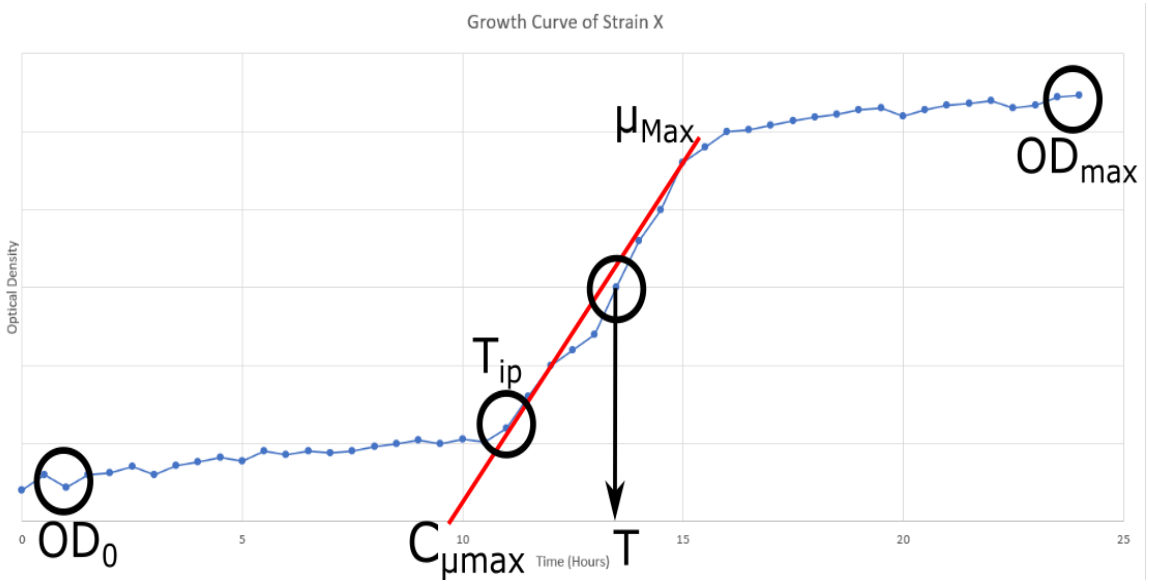


Figure 3.2.2: ***Parameters from equation 3.2.2 illustrated on an example of a growth curve.*** *MaxOD = Maximal OD of graph, OD_0 = OD baseline (i.e. average OD value of 2nd, 3rd and 4th timepoints), μ_{max} = highest slope (i.e. growth rate), $C_{\mu_{max}}$ = intercept with y-axis of the regression line of the maximum slope, T = timepoint of maximal μ_{Max}, μ_{Max} = maximal growth rate (red line), T_{ip} = predicted end to lag phase.*

value. A horizontal line was then established with this value at all timepoints (equation 3.1b). Then, by equating this baseline with the line of maximum slope (equation 3.1a), the inflection point T_{IP} could be found (equation 3.1c).

$$y_1 = \mu_{max} \times T + c_{\mu_{max}} \quad (3.1a)$$

$$y_2 = OD_0 \quad (3.1b)$$

$$\mu_{max} \times T_{IP} + c_{\mu_{max}} = OD_0 \quad (3.1c)$$

3.1: Deriving the Inflection Point of the growth curve shown in Figure 3.2.2 by equating the regression line of the maximum slope (equation 3.1a) with the horizontal OD baseline (equation 3.1b). μ_{max} = highest slope (i.e. growth rate), $c_{\mu_{max}}$ = intercept with y-axis of the regression line of the maximum slope, OD_0 = OD baseline (i.e. average OD value of 2nd, 3rd and 4th timepoints), T_{IP} = Inflection Point

As a final characteristic, the C-value (i.e. $c_{\mu_{max}}$) was extracted from the growth curves. The is simply the y-axis intercept of the regression line for the maximum slope μ_{max} (equation 3.1a). These values, and various ratios between control media growth and growth under an inhibitor, were selected as they are easily recognisable to biologists, visually identifiable on a graph, and represented core parameters of a growth curve. Figure 3.2.2 illustrates how these different features account for much of a growth curve's characteristics and how they inter-relate. Collectively, they draw together much of the information stored on a growth curve to inform us

of the overall resistance of a strain.

3.2.4 Growth curve feature distributions

To determine which features were useful discriminators of the growth phenotype, their distributions across an experiment are measured (see Section 3.4 for the description of an experiment leading to the growth curve data seen in figure 3.2.3). Unfortunately, these features are difficult to analyse collectively; they have widely varying means, variance and units. To rank them on a single scale of relative resistance, it is necessary to relate all the features and strains to each other. This is done using the k-means [108] algorithm. The k-means algorithm allows us to cluster each feature individually on a linear scale. Those features which display a uniform distribution across the strain set would be particularly suitable for resistance score binning using such an approach and could be carried forward for further analysis.

From the graphs in figure 3.2.3, we see a range of very different distribution shapes. The graph of $T_{\mu_{max}}$ in figure 3.2.3a, the timepoint at which the slope μ is at its highest for a given strain, displays a highly uniform distribution. This allows for excellent clustering into 'highly resistant' to 'highly sensitive' strains, as we can assume that having an early $T_{\mu_{max}}$ means fastest growth was achieved very quickly, as expected of a highly resistant strain. Conversely, late highest growth would indicate a highly furfural-sensitive strain as it took a protracted period of time for the strain to begin growing fully. Two other features, MaxOD (figure 3.2.3c) and T_{IP} (figure 3.2.3d), gave distributions that were next closest to uniform, particularly within a reduced range of values. These two features were therefore also deemed suitable for further analysis.

We also see in figure 3.2.3b that the C-value of a graph looks to be normally distributed and to be a poor discriminator of resistance. Most strains possess an intermediate C-value, which does not enable us to precisely differentiate between strains. This feature was therefore disqualified from the final analysis. However, in future a transformed C-value could be considered and tested. The two ratio features (figures 3.2.3e and f) showed highly right-skewed distributions and were again disqualified from further analysis.

It is interesting to note that the maximal y-values of the graphs illustrate the suitability of a feature distribution. In graphs A, C and D, we see no frequency bar

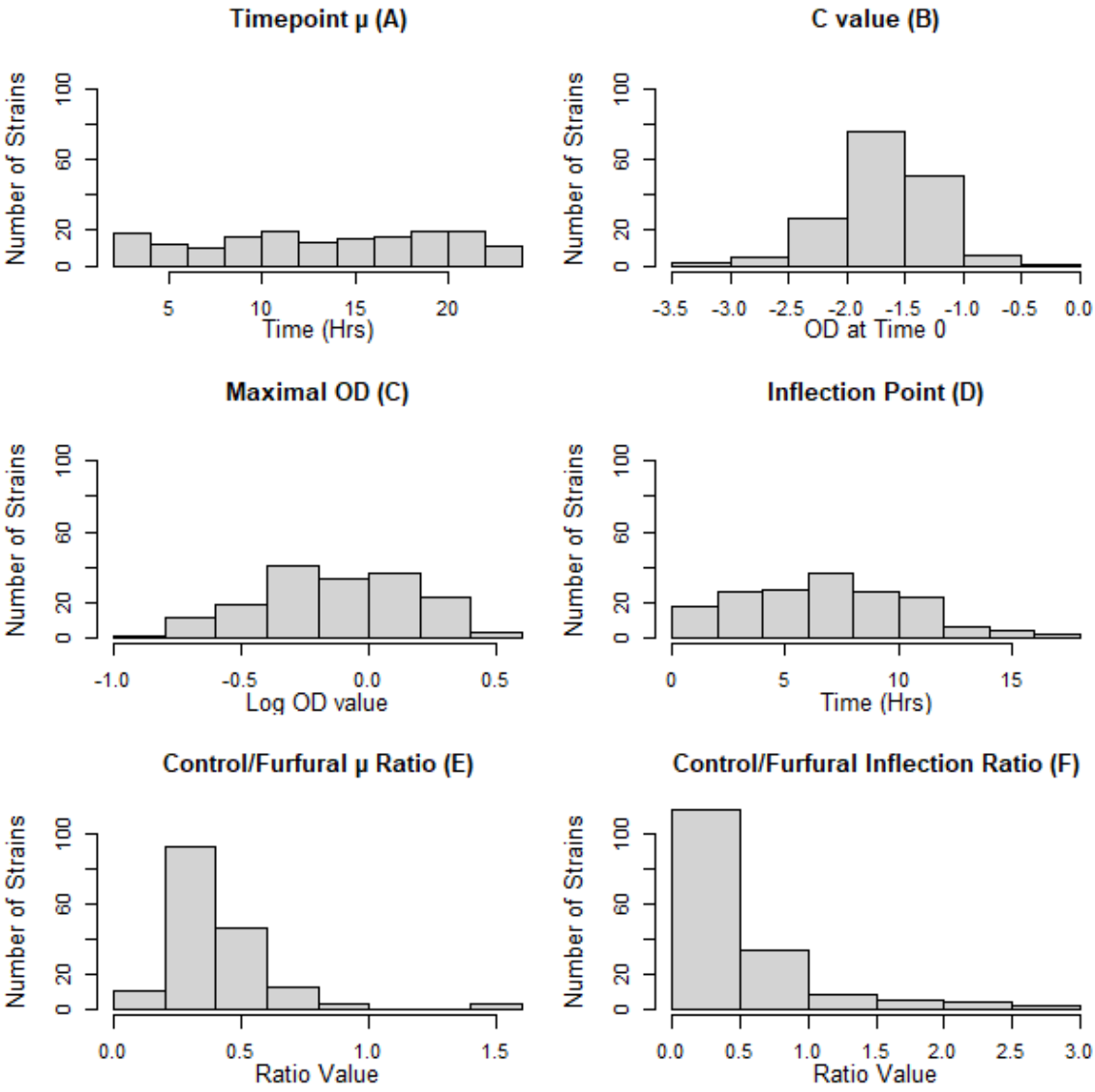


Figure 3.2.3: Frequency Distribution histograms of six curve features in an exemplar experiment measuring resistance to furfural: a) timepoint at which the maximim growth rate was observed ($T_{\mu_{max}}$), b) the intercept of the maximum-slope regression line with the y-axis (C-value), c) the maximum OD value (MaxOD), d) Inflection Point (T_{IP}), e) ratio of growth rate, μ, between strains in control media (no furfuraldehyde) and furfural media (furfuraldehyde), f) ratio of inflection ratio between strains in control media (no furfuraldehyde) and furfural media (furfuraldehyde).

reaching above 50 (i.e. the value is seen in fewer than 50 strains), while a single very high bar is present in graphs B, E and F. This example highlights how the distributions of three of the characteristics tested are highly non-uniform, with over 30% of strains in the study showing values grouped tightly together.

In conclusion, three graph features were chosen for inclusion in a resistance phenotype score for this particular experiment; the Maximal OD value (MaxOD) of the growth curve, $T_{\mu_{max}}$ (the time-point in the graph at which the growth rate μ is highest)), and the Inflection Point (T_{IP}), or end of lag phase growth. However, the

chosen features are not necessarily fixed but could be varied to suit the experiment in question.

3.2.5 Selection of k

As we saw above, three growth features (MaxOD, $T_{\mu_{max}}$ and T_{IP}) were selected for use within a resistance phenotype score based on our ability to bin their frequency distributions evenly. Subsequently, the optimal number of bins (k) for each of the three features was identified using the gap statistic [109]. The highest of the three k values was then used to bin each of the three features. This highest value was chosen so that there would be an equal number of bins for each feature, which would mean that their contributions to the resistance score would be weighted equally. Furthermore, when selecting k, it is preferable to over-fit a little, rather than under-fit and fail to explain data variation. A resistance score on a scale of 1 to k was then attached to each strain for each feature, depending on the values extracted from its growth curve. Finally, the three resistance scores (for the three features) were summed - on a scale of 3 to $3\times k$ - to provide a holistic phenotype measurement.

To calculate the value of k for a given dataset, the gap statistic is used. The method proceeds as follows, according to a custom R script (Dr Jo Dicks, personal communication). For each feature and for each possible value of k (from 1 to 20), k-means clustering (see Chapter 2) was used to bin strains into k groups for the given feature. Then, for each of the k groups, the distance between each pair of points in the group was calculated and summed, to give a value D_r (for group G_r), as shown in equation 3.2a. The distance used for this calculation was the Modified Rogers' distance, a popular choice for genetic datasets. The k values of D_r were then combined, as shown in equation 3.2b, to find the within-group dispersal measure W_k, where m_r is the number of data points within G_r. For spatial datasets, where Euclidean distances between data points can be used in equation 3.2a, W_k is equal to the pooled within-group sum of squares measure about group means. The dispersal statistic describes how well a clustering fits a set of data points. The smaller the value, the closer points within each group are to one another. Therefore, a low W_k score indicates a tight grouping. Next, for each value of k a specified number n of random datasets with similar features to the initial, real dataset were generated. The

$$D_r = \sum_{i,i' \in G_r} D_{ii'} \tag{3.2a}$$

$$W_k = \sum_{r=1}^{k} \frac{1}{2m_r} D_r \tag{3.2b}$$

$$Gap_n(k) = E_n^*\{log(W_k)\} - log(W_k), \tag{3.2c}$$

$$W_k^{'} = \sum_{r=1}^{k} \frac{2}{m_r(m_r - 1)} D_r \tag{3.2d}$$

3.2: Gap statistic equations, which can be used to identify the number of groups k within a dataset.

random dataset generation process, inspired by Jonathan Marchini's nps function within the R popgen package, creates datasets with allele frequencies identical or near-identical to those in the real dataset. The final step of the gap statistic estimation is to find the distance (the 'gap') between logarithms of the within-group dispersal measures for the expected value of the randomly generated dataset and the real dataset, as shown in equation 3.2c. The chosen value for the number of groups is either the value of k which maximises the gap function between real and random datasets, or one which shows a high gap value within a suitable range of values for k.

The W_k and weighted W_k ($W_k^{'}$, defined in equation 3.2d) values were also plotted and visualised for each value of k (k = 1 to 20) and each feature. In particular, the 'elbow' or 'inflection' of each graph was noted, the point after which increasing the number of groups for that feature only reduces the within-groups dispersal measure incrementally. This point is often used in data clustering as an alternative measure for the number of groups.

3.3 A Novel Method for Estimating the Genetic Contribution of Founder Populations to a Microbial SNP Dataset

As described in Chapter 2, understanding the contributions of founder populations to the genomes of microbial strains used within a Genome Wide Association Study is important. This information, in the form of a Q-Matrix, can be used to remove

spurious correlations between genotypes and phenotypes that arise when closely related organisms have similar phenotype values but where less closely related organisms have dissimilar phenotype values, both as a result of their ancestry. Various software, such as PSIKO, have been developed to estimate such a Q-Matrix. Within this project, the SANE (Simulating Ancestry through Nucleotide Equations) software for Q-Matrix estimation was developed using the principles of genetic distance between DNA segments (see figure 3.3.1). SANE is a Python 3 program, the source code for which can be found on github (link in abstract).

SANE begins by first finding the k strains (for a chosen number of groups, k) that are most dissimilar and distant genetically (figure 3.3.1 step 1). To find distances between pair of SNP genomes (see Chapter 2 for a definition) it uses the TamD [110] genetic distance measure. These strains are then counted as the 'seeds' for predicting k ancestral founder population genomes (at the defined SNP sites).

Next, each strain is added to its nearest genetic neighbour (using the same measure of distance) within 'seed trees', one for each of the initial 'seed' genomes, until a final set of trees has been built with all the strains in the study (figure 3.3.1 step 2). This means that the number of trees k depends on the initial number of 'seed' strains - and there could be a tree composed of only the 'seed' strain, if no other strain was similar enough to it. Then a founder 'average SNP genome' is constructed from each seed tree. This average genome is built simply by finding the most frequently observed SNP per SNP site across all strains in the tree. If, at any site, there is an equal number of two variant SNPs within strains in a seed tree (a 'tie'), a SNP is picked at random from the two possibilities for the average SNP genome (figure 3.3.1 step 3).

Subsequently, each SNP genome in the study (including the ancestral average genomes) is fractured into a number of segments (haplotypes), n (chosen manually). For a given strain, each segment is compared (again, using genetic distance calculations) to each analogous segment within the k average SNP genomes to find which is most similar to it. The number of segments within a strain's SNP genome that match to average SNP genome i is m_i (for $i = 1$ to k). The strain is therefore found to contain ($(m_i/n) * 100$)% of average SNP genome i (figure 3.3.1 step 4).

Making a list of the percentage similarity of each strain's SNP genome to the k average SNP genomes, we can build a Q-Matrix (figure 3.3.1 step 5). This Q-Matrix can then be employed in the same manner as the PSIKO Q-Matrix, using a linear

mixed model to find associations between the SNPs within the SNP genome and phenotypic values.

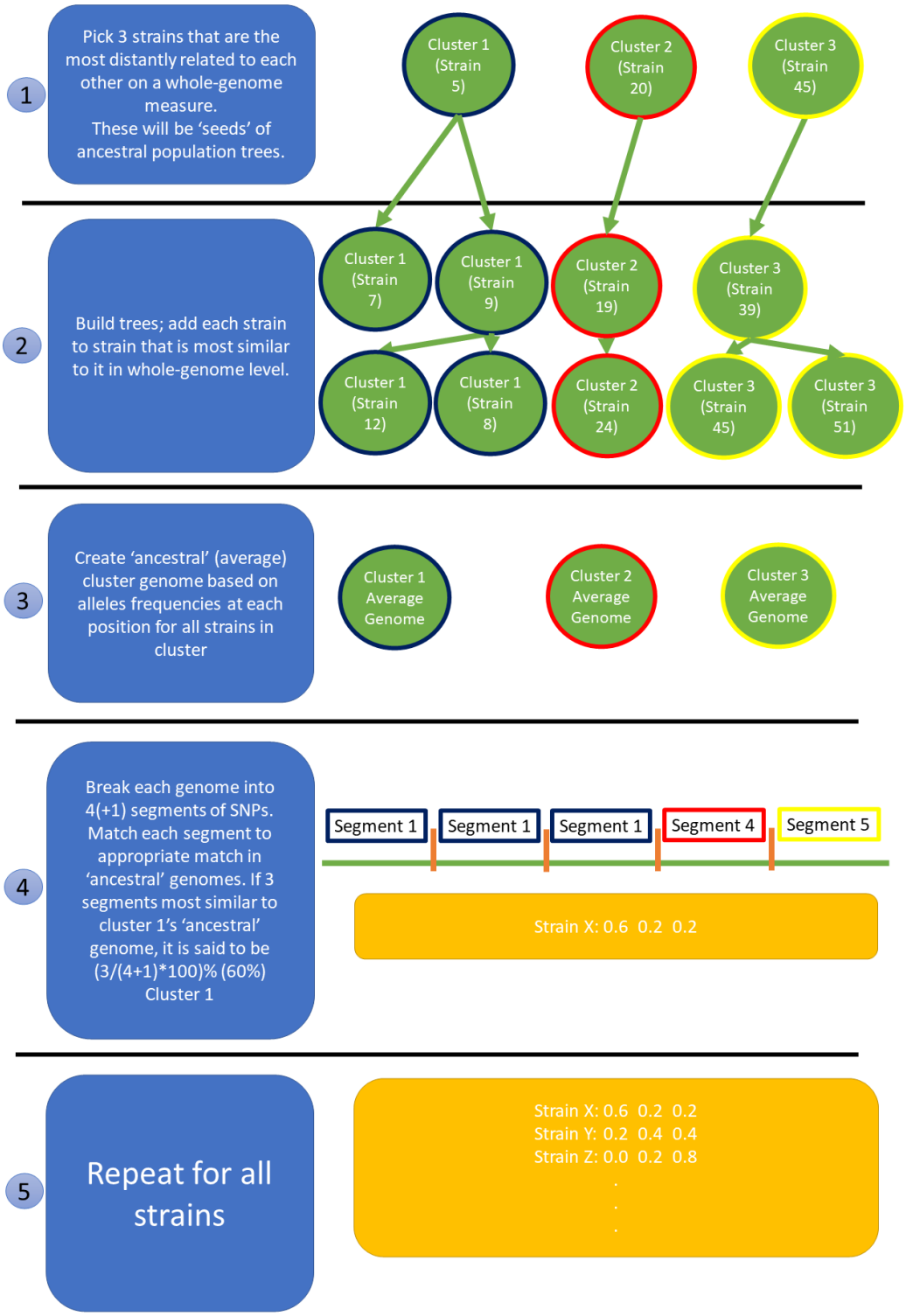


Figure 3.3.1: ***SANE Q-Matrix estimation***
Workflow of the 5 SANE steps where the number of groups k is chosen to be 3. In step 1, the 3 most distantly related strains from a given dataset are identified. In step 2, all other strains in the dataset are grouped with those most similar to it. In step 3, strains in a group are 'averaged' to find SNP genomes representative of a group. In step 4, the SNP genome of a chosen strain is broken into segments and each segment is matched to the averaged SNP genomes in step 3. The contribution of each averaged SNP genome to the SNP genome of the strain is identified and a row of the Q-matrix is calculated. In step 5, step 4 is repeated for all strains, resulting in a full Q-matrix.

3.4 A Genome Wide Association Study for resistance to furfuraldehyde in *Saccharomyces cerevisiae*

3.4.1 The Strain Set

The yeast strains selected for the study originally included the approximately four thousand strains within the National Collection of Yeast Cultures (NCYC [1]). However, the dataset was quickly reduced to those whose genomes had been whole genome sequenced (965 distinct strains). Subsequently, we selected only the *Saccharomyces cerevisae* (406 strains) within our dataset as these were most likely to have adaptations to commercial fermentation conditions and reduce FDR effects due to distant kinship relationships between species. However, some of these strains' WGS reads had been subject to inter-strain contamination, adapter contamination or were low depth (here less than 30). After removing these datasets from the study, there remained 168 *Saccharomyces cerevisae* strains with read dataset of sufficient depth and quality to enable high-quality downstream analysis.

For a full list of the NCYC sequencing plates and the *Saccharomyces cerevisae* strains used within this study, see table 2 of Chapter 6.1. The *Saccharomyces cerevisae* and other strains were taken from these plates.

3.4.2 OD analysis for resistance phenotype elucidation

To identify the resistance profiles of the chosen yeast dataset, each strain was grown for 24 hours and the OD was measured at each 30-minute time interval, using the method described in Section 3.2.1. OD readings were plotted using RStudio [69] scripts, producing 96-well graphs such as that seen in figure 3.2.1 (the negative controls without any yeast cells were removed from the figure for visualising the data).

For the furfural resistance study, it was essential that resistance could be quantified as accurately as possible. For this, visual checks of OD curves across time were insufficient; comparing hundreds of curves to identify resistance on a multi-point scale is impossible to conduct with adequate precision. Identifying resistant from sensitive strains is possible, yet the wide-spectrum resistance to furfuralaldehyde was likely to be highly variable. To resolve this, the computational method described in Section 3.2 was used to analyse the hundreds of curves in a

reproducible, reliable and verifiable manner. When a model is proven insufficient to the task, it is easy to update with the new understanding to encompass previous failures. In fact, the very model itself can tell us important things about resistance; for example, if the lag phase is the most variable factor, it is likely the most impacted by furfural and, consequently, the best for measuring resistance to furfural.

As described in Section 3.2.3, several features of each OD curve were initially identified, but the curve data for the 168 *Saccharomyces cerevisiae* strains (see figure 3.2.3) indicated that only three of the six measured characteristics were suitable for calculating a resistance score. These features were the Maximal OD reading (MaxOD) of a curve, the time at which the steepest slope (indicating the fastest growth rate, μ_{max}) was observed in a sliding 4.5-hour window and the end of lag phase, calculated by working from the time where μ is highest and working back to the baseline OD value. In this way, we can explain a curve in the language of characteristics; when does growth start (end of lag phase, T_{IP}), where is it highest ($T_{\mu_{max}}$), and what is the maximal OD reached, indicating highest call mass (MaxOD).

3.4.3 Calculating a resistance score

Once the three growth curve features had been chosen, the best choice for the number of groups k needed to be identified. This was done in two ways, as described in Section 3.2.5. Firstly, the Gap statistic was calculated for the three features using 100 randomly generated datasets. For time of maximum slope, the highest value of the Gap statistic in the range $k = 1$ to 10 was $k = 4$, as shown in figure 3.4.1A, and with a secondary peak at k=7. For MaxOD, the Gap statistic indicated a non-trivial (i.e. after $k = 2$) maximum value at $k = 6$ (figure 3.4.1B), with a secondary peak at $k = 4$. For time of Inflection Point, the Gap statistic gave a non-trivial local maximum of $k = 3$, with a subsequent secondary peak at $k = 6$ in the range $k =$ 1 to 10. In general, high secondary (or tertiary peaks etc.), even if they are not maximal values can represent group structure in the dataset.

Secondly, we plotted the within-group dispersal measure (W_k, a measure of how 'tight' a cluster is) and its weighted version for $k = 1$ to 20 to see where 'elbow points' could be seen in the graphs (figure 3.4.2). These turning points could be seen at $k = 4$, $k = 6$ and $k = 3$ for the time of maximum slope ($T_{\mu_{max}}$), maximum OD reading (MaxOD) and time of inflection point (T_{IP}), respectively, mirroring the

results seen for the Gap statistic. A sudden change in the dispersal measure would indicate the newly-selected k number significantly changes how the data clusters and is thus an important k number.

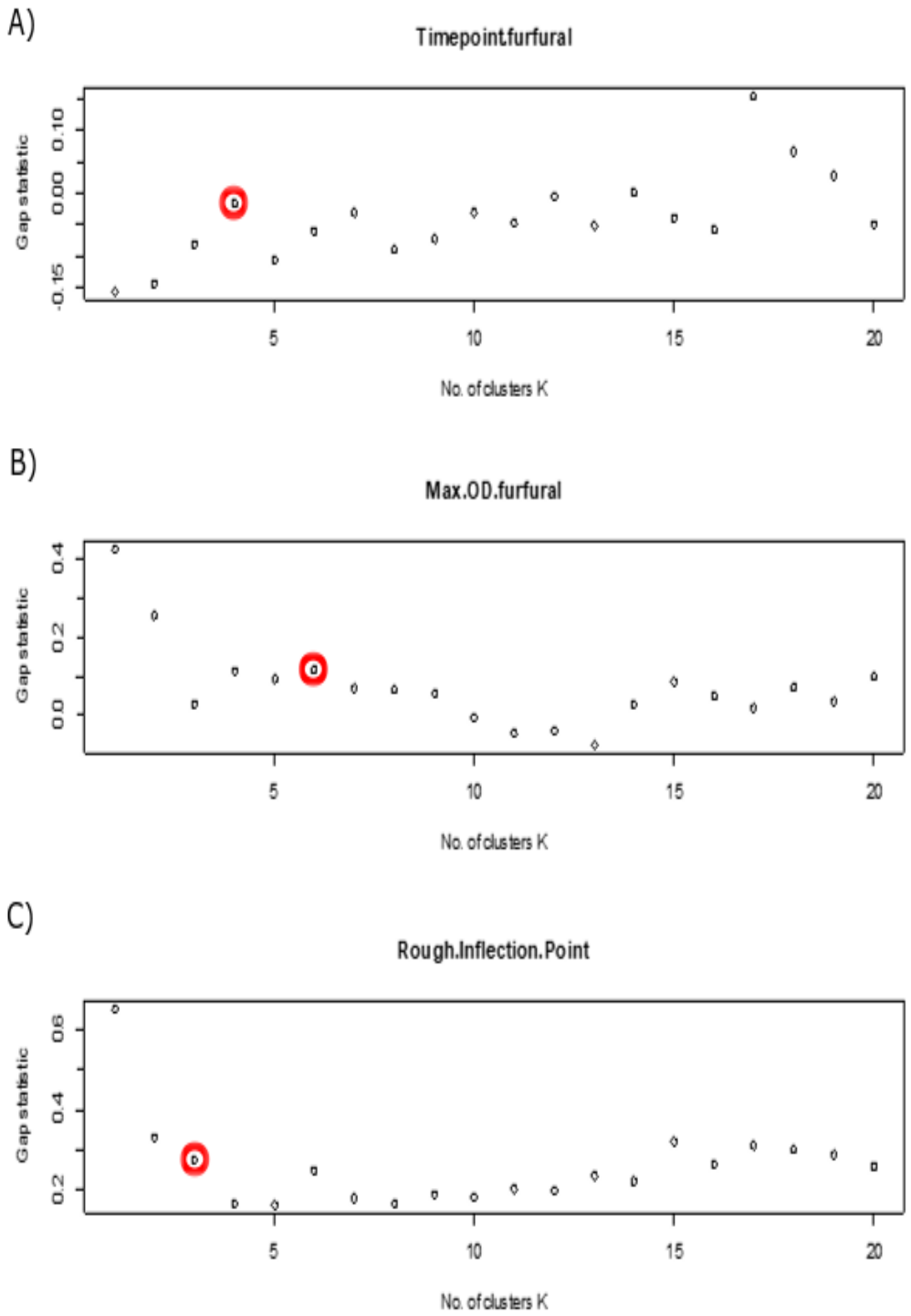


Figure 3.4.1: Gap statistics for growth curve features.
Gap and statistics for k=1 to 20 with 100 randomly generated datasets for A) Time of highest slope $T_{\mu_{max}}$, B) Maximum OD reading (MaxOD) and C) Time of Inflection Point (end of lag phase; T_{IP}). Values of $k = 4$, $k = 6$ and $k = 3$ (denoted with red circles) were chosen for these three features, respectively.

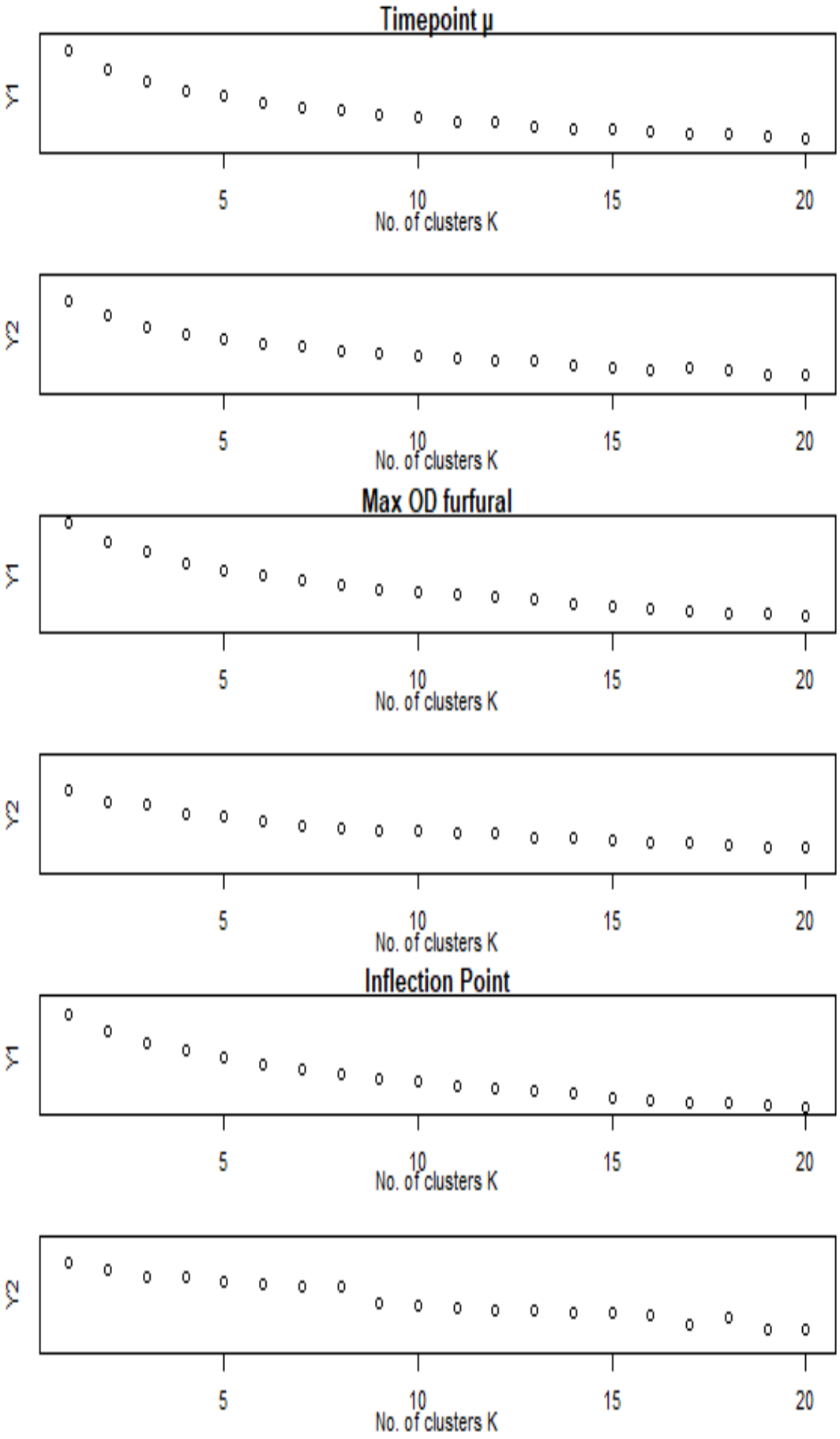


Figure 3.4.2: Within-group dispersal values for the three growth curve features.
Y1 = K-means sum of squares, Y2 = Weighted K-means sum of squares for the time of maximum slope, MaxOD and time of inflection point features, respectively.
A sudden drop in the sum of squares (SS) value, indicating the 'elbow' of a figure, would indicate the new k number clusters the data much tighter and is therefore a k value of interest. Here, we try to find differences between the unweighted (Y1) and weighted (Y2) SS calculations.

Collectively, these results indicated that $k = 6$ was the optimal number of groups for this dataset. To confirm this, distributions of the three chosen growth curve features were binned using k-means clustering for $k = 5, 6, 7$. Examining figure 3.4.3, we can see that each feature selected is roughly balanced across the strains, except for the first group (lowest values) of the Inflection Point feature, which is under-represented and moderately skews other clusters. Further, we can see the differences in strain distribution depend on the value of k. When $k = 5$ (black bars), the strains group tightly with the mode in cluster 4 for all three features (figure 3.4.3). This is undesired, as a bias towards a single cluster would skew the resulting phenotype scores.

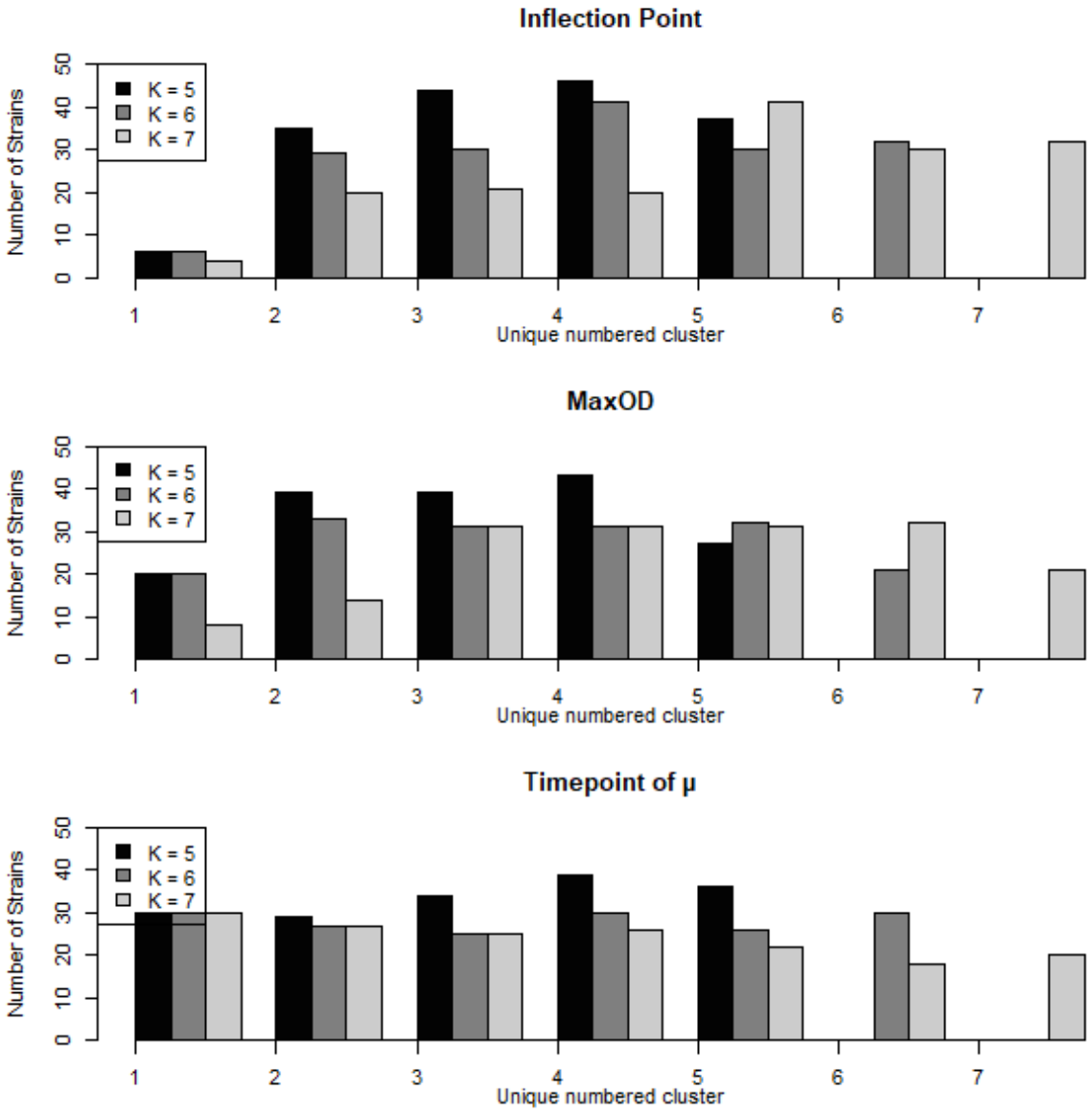


Figure 3.4.3: **Frequency distributions of grouped features**
Numbers of strains within each growth curve feature group (for k = 5 to 7) across the 168 Saccharomyces cerevisae strains within the study. Clustered features were; Inflection Point for start of growth (end of lag phase), Maximal OD, Time-point of highest growth μ_{max}.

However, when $k = 7$ the variance of cluster strain frequencies is high. For example, the MaxOD characteristic displays this point, with clusters 1, 2 and 7 under-represented. The within-cluster frequencies for the Inflection Point feature also fluctuate between clusters when using $k = 7$. Setting $k = 6$ produces the most balanced frequency distributions; there is little bias for a specific cluster while none (barring cluster 1 of Inflection Point for all values of k) of the clusters are wildly under-represented. In this way, the data is parsed more effectively and differences in strains are recognised more accurately without an overly-elevated k number causing fragmentation of phenotype.

With a final value for k now chosen, resistance scores could be determined. For each of the three features binned into $k = 6$ groups, each strain was ranked from 1-6 for each feature and these values summed to produce a final resistance score of 3-18. For example, a strain with a very high MaxOD, low $T_{\mu_{max}}$ and low (early) inflection point will score a '6' for its MaxOD, '5' for its low $T_{\mu_{max}}$, and '5' for its low (early) inflection point T_{IP}. Summing the three values, we can see the strain has a high resistance score of 16 - which would be expected as the strain's growth curve characteristics were each expected in a strain of high furfuraldehyde resistance.

The concern of a unified resistance score would be the disguising of interesting results in a single characteristic. However, unless we wished to triple the outputs, it presents a reasonable solution to concatenating disparate growth curve characteristics and, potentially, genes variants affecting all the characteristics. There are outliers for all characteristics (figure 3.4.3), however by joining them it is hoped that individual outliers in a single characteristic doesn't play an undue role in skewing the data.

Categorising the entire dataset in this way gives a holistic resistance score to each strain, with values ranging from a minimum of 3 to a maximum of 18. A Principal Component Analysis was carried out on the three initial variables of MaxOD, Inflection Point and $T_{\mu_{max}}$ for all 168 *Saccharomyces cerevisiae* strains. The results of the PCA analysis were then plotted according to the first three Principal Components (see figure 3.4.4). A K-means clustering of the overall resistance scores into four resistance categories (Highly sensitive, Sensitive, Resistant and Highly resistant) was then conducted, with all points in the PCA plot coloured according to these groups. The resulting figure, which shows similarly coloured strains close in 3D space, gives a clear view of how the raw data can be partitioned into the various

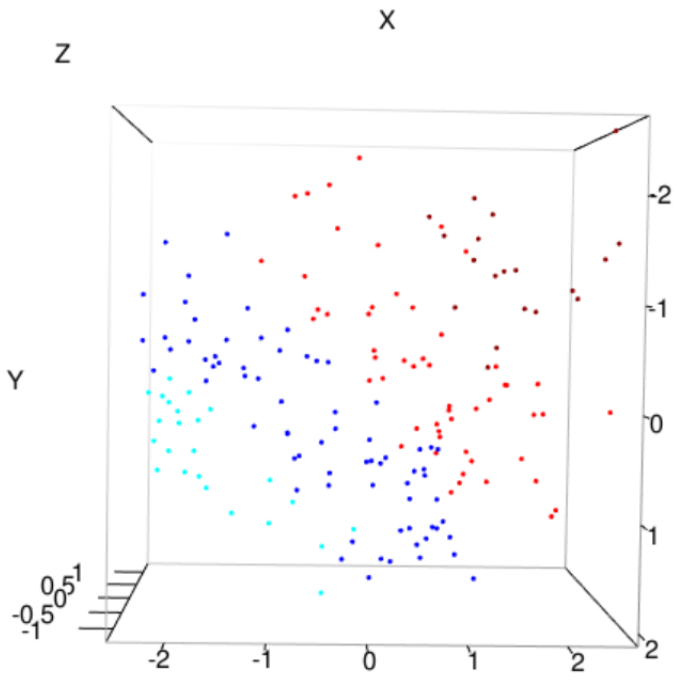


Figure 3.4.4: PCA plot of the 168 Saccharomyces cerevisiae strain dataset, with initial variables the three chosen growth curve characteristics ($T_{\mu_{max}}$, MaxOD, Inflection Point). Colouring based on holistic K-means resistance scoring. Dark red = Highly sensitive (3-6), Red = Sensitive (7-10), Blue = Resistant (11-14), Cyan = Highly resistant (15-18)

resistant vs non-resistant strains.

Furthermore, we can plot the number of strains for each Resistance Score to help parse the data and see how the strains end up clustering on the 3D plot (figure 3.4.5). On this figure, we can see how there are few strains at the tail ends of resistance, with most being intermediate through the sum of their growth curve characteristics' scores. This highlights the importance of determining the correct value of k to properly parse the highly clustered strains of intermediate resistance, since even using the measures described above, we fail to get a uniform distribution for overall resistance.

Table 3.2) shows basic information on the nine strains with resistance scores of 16 or 17 (the 'top strains'). It is immediately evident from the Habitat field that the strains in the study come from disparate backgrounds. Some strain origins are medical, while many are brewing strains, with origins of many of the older strains unknown. Similarly, table 3.1 shows the same information for the six strains with resistance scores of 3 or 4 (the 'bottom strains'). From both table 3.2 and table 3.1, we can see some expected results in terms of presence or absence of appearance of the strain in the fermentation literature. This is examined in more detail below.

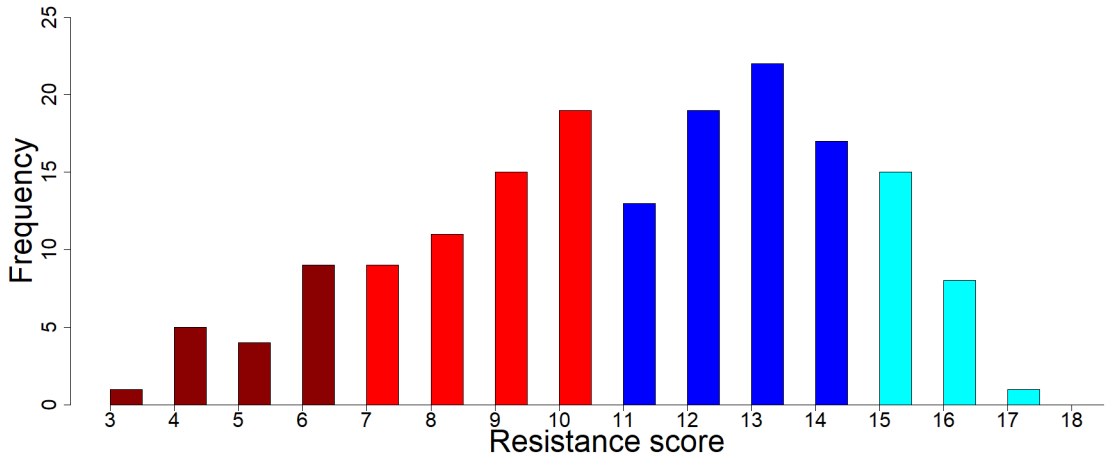


Figure 3.4.5: Histogram of strain resistance scores
Histogram showing strain distribution per Resistance Score, coloured to match the 3D plot in figure 3.4.4
Colour scheme is based on holistic K-means resistance scoring. Dark red = Highly sensitive (3-6), Red = Sensitive (7-10), Blue = Resistant (11-14), Cyan = Highly resistant (15-18)

Strain Number	Alternative Names	Deposit Name	Resistance Score	Habitat	Fermentation Literature
620	CBS 3012	Saccharomyces cerevisiae	4	Jerez Sherry production yeast from Feduchy	0
776	ATCC 12341, H.P. Klein strain LK2G12	Saccharomyces cerevisiae	4	Unknown	0
1444	Saccharomyces cerevisiae	Saccharomyces cerevisiae	4	Ale production strain	0
3313	OS 92/A, Single spore isolate of DBVPG 1853	Saccharomyces cerevisiae	4	White tecc, Ethiopia	0
3612	Mat alpha derivative of YIIc17_E5	Saccharomyces cerevisiae	4	Unknown	0
3265	0S17/A, Single spore isolate of SK1	Saccharomyces cerevisiae	3	Lab strain, USA	0

Table 3.1: Bottom Scoring Strains
Six Saccharomyces cerevisiae strains with lowest resistance scores of 3 or 4 identified through K-means clustering of three growth curve features

Strain Number	Alternative Names	Deposit Name	Resistance Score	Habitat	Fermentation Literature
200	NCTC 608	Sternberg '675'	17	Unknown	0
70	NCTC 3864	Saccharomyces anamensis	16	Unkown	0
74	ATCC 9080/24904, CBS 2354, NCTC 7014, Hillman hospital 4228	Saccharomyces carlsbergensis	16	Hillman Hospital, Birmingham, Alabama, USA	0
196	NCTC 3966	Yeast Race V	16	Uknown	0
221	A38/3, S. spore isolate from NCYC 213	Saccharomyces cerevisiae	16	Hybrid brewing strain	0
2798	MAS 6	Unknown	16	Mouth of AIDS patient	0
2826	CECT 1483, IFI 649	Saccharomyces cerevisiae	16	Grape Must	[42, 101, 111, 112, 113]
3467	OS281, S. spore isolate of W303	Saccharomyces cerevisiae	16	Unknown	0
3557	Mat a/alpha derivative of DBVPG6040	Saccharomyces cerevisiae	16	Uknown	0

Table 3.2: Top Scoring Strains
Nine Saccharomyces cerevisiae strains with highest resistance scores of 16 or 17 identified through K-means clustering of three growth curve features

3.4.4 Linking identified strains to past studies

A standard method of validation for any model is to verify if any outputs align with previous experimental evidence. As such, the top scoring strains were investigated for their past impact on the scientific literature. Many NCYC strains have been used for a diverse range of research. Tables 3.2 and 3.1 do not display papers unrelated to the study in question. For example, NCYC 74 has no literature listed in table 3.2, yet

has been used in a paper involved in mapping the transcripts of the mitochondrial genome [114]. This literature filtering was done in an effort to reduce 'noise' and a false sense of significance for strains with many publications of an unrelated nature. Literature was searched with NCYC numbers through Google Scholar.

Other strains display desired phenotypes in specific papers. For example, NCYC 2826 has been identified as possessing a desired function in the highest number of papers. It has been tested experimentally for its fermentative success to grow on specific carbon sources [112], as well as its ability to grow on various lignocellulosic waste biomass (rice stalks [112], wheat straw hydrolysate [101]), and has been assessed for its metabolic products [112]. While displaying an extended lag phase in response to furan compounds (particularly furfural [101]), it nonetheless showed specific resistance to furfurals [101] and ethanol [42].

In a previous study to experimentally test furfural resistance, six tested NCYC strains were shown to display furfural resistance (NCYC 3451, NCYC 3284, NCYC 3290, NCYC 3312, NCYC 3277 and NCYC 2826) [101]. The only one of these six strains present in our study, NCYC 2826, is similarly shown here to be highly resistant (see table 3.2). In a separate study, NCYC 2826 (Resistance Score 16) and NCYC 3445 (Resistance Score 13) were both found to be high ethanol producers in minimal fermentation media conditions [115].

Furthermore, while much information is lacking on the bottom scoring strains (table 3.1), it still provides useful clues. One of the strains (NCYC 620) is a sherry brewing yeast strain. Under environmental conditions of readily-available sugars in grapes, sherry strains have likely been selectively bred to produce vast amounts of ethanol. As a trade off, they might suffer reduced viability in higher temperatures, lignocellulosic breakdown by-products or acidic environments. However, one of the strains (NCYC 1444) is an ale strain - well used to indigestible sugars, even if brewed at colder temperatures (3-4°C [116]) and is thus unsuited to higher fermentation temperature pre-treatments.

The positioning of strains NCYC 620 and NCYC 1444 (of which we have habitat information) in the bottom of the resistance bands can be explained logically. Both are unsuited to the high temperature, low Ph and lignocellulosic byproduct-rich fermentative environments of industrial metabolite production systems.

3.4.5 GWAS analysis to identify *S. cerevisae* SNPs involved in furfuraldehye resistance

In the present study, the raw phenotypic data is multivariate. With some variables continuous (MaxOD) and others discrete (T_{IP}, $T_{\mu_{max}}$), it can be difficult to directly compare variables. Additionally, the variables had highly differing means and variances, with some being inversely correlated (i.e. high Maximal ODs should correlate with low inflection points to indicate a highly resistant strain). To resolve this difficulty of analysis, the variables were reduced to single cluster scores (all growth curve feature distributions as seen in figure 3.4.3 and their sum in figure 3.4.4). The variables were thus linearised to a single discrete holistic resistance score from the continuous multivariate input data. Although this causes a loss of information from single characteristics, it is hoped to reduce the noise of outliers in a single characteristic in a growth curve. E.g, A high final OD is less valuable in determining resistance if the strain always presents a high OD due to being an unique colour and there is little change in OD from start to end of growth (signifying little actual growth).

The SNP genomes of the 168 *Saccharomyces cerevisiae* strains were generated using a highly conservative computational pipeline (see Chapter 2). In particular, the FAT-CIGAR tool, which ensures that reads are mapped exactly to the reference genome at both ends (for a user-defined base pair length) prior to variant calling, is effective at reducing the number of false positive SNP calls. Yet even with this pipeline the study identified 84,046 high-quality SNPs (MAF > 5%) across the approximately 12.1 million base pairs (i.e. 1 SNP for roughly every 143 bases). To understand how these SNP genotype data were related to the phenotypic resistance scores, the correlation type was measured. If a correlation between the reference allele and the phenotype was negative, then the alternative allele (i.e. the SNP) was related positively to resistance. Conversely, a positive correlation of the reference allele with the phenotype indicated the alternative allele was related to a strain's sensitivity to furfuraldehyde. This distinction is important, as alternative alleles have presumably arisen via adaptive evolution. In essence, we would expect most of our top hits to relate to an alternative allele conferring greater resistance to furfuraldehyde. In table 3.3, we can see how many of our top 10 hits have alternative alleles related to furfural resistance.

The SNP dataset was then correlated to the resistance phenotype dataset, accounting for strain ancestry with the PSIKO and SANE Q-Matrices via the method described in Chapter 2. A conservative Bonferroni approach to taking into account the high number of SNP/phenotype correlations, where we would expect to see strong false positive correlations arising by chance, would give a 5% significance threshold of $0.05/84046 = 5.95 \times 10^{-7}$- which would not be quite accurate a measure due to Linkage Disequilibrium. However, we can also use less conservative corrections such as the False Discovery Rate (FDR) used commonly in GWAS analyses. Our analysis did not employ them to remain conservative with the huge number of potential hits involved[76].

If the SNPs had no 'real' correlation to the phenotype scores, we could expect 86.5% of the SNPs to be positively correlated to the phenotype in question (total number of positive/total number of SNPs). However, we find that when using the PSIKO Q-Matrix 74% of the top 1,000, 80% of the top 500 and 98% of the top 100 SNP hits are negatively correlated with the phenotype. That is to say, the alternative allele is predicted to be contributing to the desired phenotype that percentage of the time.

A second Q-Matrix was constructed with the SANE method (see Section 3.3), that utilised genetic distance estimates based on sequence similarity to predict founder populations. The genetic distance used in this study was Tamura Distance (TamD) measure [110], with genomes being fragmented into DNA segments of 5% of the total number of SNPs and with 3 original founder populations, as indicated by plots of within-group sums of squares for different values of K, analogous to the approach taken with resistance score grouping. The TamD distance is highly appropriate for eukaryotic genomes and differentiating between divergent yeast strains [110], as we see within the global *Saccharomyces cerevisiae* populations, many of which are represented within the 168 strains used here.

The fragment size was chosen to be sufficiently large (60,000bp+) for calculating genetic distances accurately. With the smallest chromosome (e.g. ChrI in *Saccharomyces cerevisiae*) being roughly 230,000bp [117], a fragment size of close to 100,000bp (1% of total) would seem reasonable. These parameters can be modified and updated as understanding of the dataset develops. For example, if the size of the genomes change, the fragment size may change or if the base assumption about the type of mutations present is updated, then the genetic distance method may

need to be updated.

To compare the Q-Matrices estimated by the PSIKO and SANE algorithms, we conducted a Mantel test, a statistical method for the multivariate analysis of spatial genetic matrices [118]. A strong correlation between the two matrices was found, with a p-value of 0.001 (r=0.4972247). This indicated that the SANE method has promise, when compared to the more established PSIKO approach. However, somewhat surprisingly, the GWAS results when using the SANE Q-Matrix were rather different, with only 58.9% of the top 1,000, 56.2% of the top 500 and 41% of the top 100 SNPs having negative correlations with the phenotype. This could indicate the genetic distance method removes fewer false positives correlations than PSIKO's half-PCA method if we assume the reduction in negative correlation signal to be a reduction in FDR. Therefore, while the matrices are highly similar, the small differences between them are important to the results of the LMM predictions used within the GWAS (All Matrices in github repository in Abstract).

3.4.6 Specific Hits

The GWAS analysis resulted in many tens of thousands of SNPs, each with a p-value illustrating its correlation to the phenotype. As a low p-value suggests a more significant correlation, it can be more intuitive to use the log of the p-values. In this case, high values indicate high significance. This information is then best illustrated through a Manhattan plot (figure 3.4.6) that shows each SNP on the x-axis and its corresponding negative log p-value on the y-axis (high y-axis equating to a high correlation) [119].

Above the numerous SNPs towards the bottom of this plot that show little relationship to furfural resistance, peaks of correlation spike for specific genomic regions and the few top scoring SNPs peak above these. In our plot, colours correspond to individual ORFs, as our 'genome' is an artificial construct of ORFs from 1,011 *Saccharomyces cerevisae* genomes [22]. In all, the plot is a useful overview of general p-values across the genomes and how they cluster. However, to investigate in a granular fashion, it is necessary to create tables to view specific SNPs, as in table 3.3.

The top GWAS hits in our study of resistance to furfurals are illustrated in table 3.3. While unfortunately none give an FDR-corrected p-value lower than 0.05,

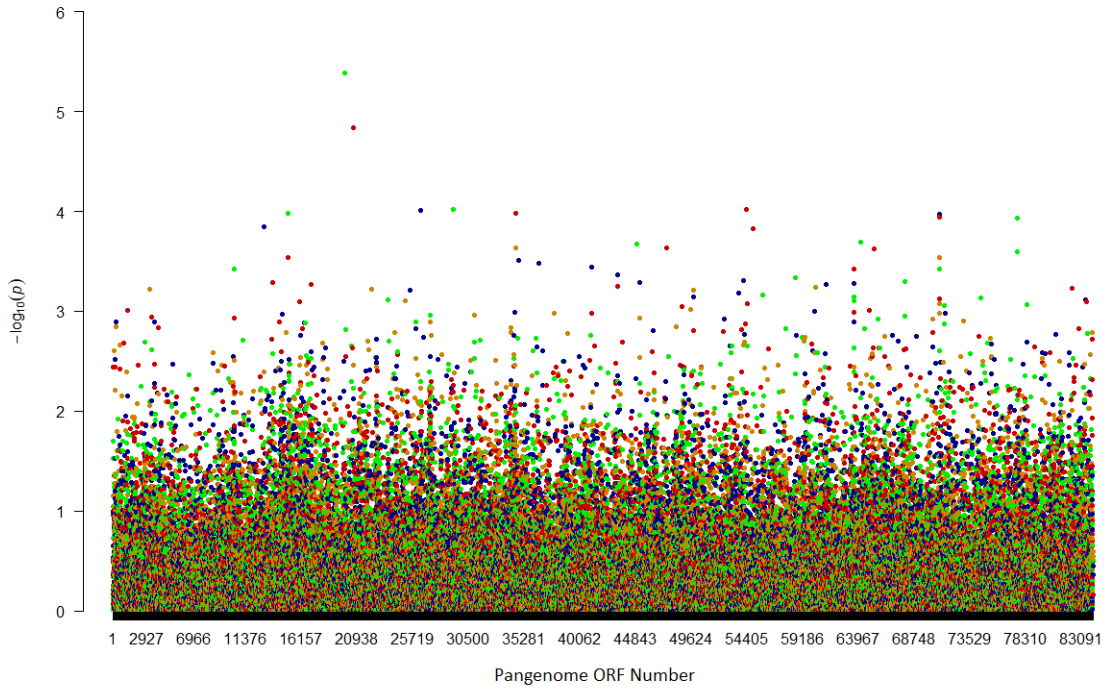


Figure 3.4.6: ***GWAS Manhattan Plot***
Manhattan Plot illustrating the log p-values (y-axis) of each SNP (x-axis), with SNPs coloured for each ORF to highlight multiple SNPs within the same ORF.

their status as lowest p-value SNPs means they and their respective ORFs require further inspection. The top scoring genes are variously important for resistance to cellular stressors (YOL105C [120]), resistance to anti-fungals (YPL056C [121]), cell replication (YLR247C [122], YGL093W [123], YOL078 [124]) and sporulation (YOL016C [125, 126]). Additionally, as an overview, most SNPs of high significance appear to confer resistance through alternative allelic functionality (positive correlation). This is seen through 64.4% of the top 500 SNPs being positively correlated to the final phenotype.

It also becomes evident that the YOL (Yeast, chromosome O/15, Left arm) genes harbour many SNPs of interest, with 12 of the top 100 SNPs being within the YOL region of the yeast genome. This includes SNPs within ORFs such as YOL136 [127], which is involved in glucose sensitivity and YOL126, a malate dehydrogenase [128]. We can therefore make a tentative assessment that this genomic region will be the site of adaptive mutations in future Directed Evolution experiments. By conducting such an experiment and sequencing the resulting strains, we could compare the evolved strains to the reference pangenome to investigate whether mutations in this region have occurred.

A second method of examining the GWAS results is to investigate only those SNPs that correlate with higher resistance in their non-reference (alternative) allelic state. In table 3.4, we can see a snapshot of this information. While the ORFS in which some of these SNPs reside appear likely to have no function on the phenotype, such as YNL054W, others show highly interesting results. For example, YOL105C appears to be involved in cell wall synthesis in stressor environments, perhaps indicating that the wrong mutation would limit its usual furfuraldehyde detoxifying potential due to the inability of the cell wall to be modified in response to furfural stress.

Previous research has identified YKL071W as involved in furfuraldehyde detoxification [129]. While a SNP (at bp 290) within this gene gave a negative correlation with the phenotype (alternative allele confers increased furfural resistance) for the GWAS with the PSIKO Q-Matrix, its uncorrected p-value was high, at 0.046. Interestingly, the same analysis performed with the SANE Q-Matrix and the Tamura Nei (TN) genetic distance [110] gave a p-value of 7.5174×10^{-4} for a SNP in this gene (bp 616 →C), perhaps indicating a similarity between PSIKO and TN distance.

YOL genes appear frequently (12 of the top 100 SNPs) in these results. Genes within this region of the yeast genome (chromosome 15, left arm) have been implicated in various functions such as Alcohol Dehydrogenases [130], cell DNA replication [131] and cytoskeleton [131]. All these functions have potential to be implicated in resistance to furfuraldehyde's broad-spectrum effects. However, some reports indicate the deletion of the region is rescued by gene duplicates elsewhere in the genome [132], suggesting that key variants may be spread more widely.

Although a SNP within the YKL071W gene was observed and found to be correlated with the phenotype in the expected direction, its p-value using the PSIKO Q-Matrix was lower than expected. Therefore, using YKL071W as a starting point, we investigated whether we could find other genes that could have 'replaced' that gene's function as the main furfuraldehyde detoxifyer. Using the YeastGenome data repository, genes with similar Gene Ontology (GO) terms to YKL071W were identified through the shared GO listed under YKL071W. Subsequently, these genes were investigated to discover whether any harboured SNPs used in the GWAS analysis (tables 3.5, 3.6). We confirmed this to be the case after checking the results of the searching process.

The YKL071W Gene has two main GO-linked functions; alcohol dehydrogenase activity and oxidoreductase activity. Searching for these two GO functions, we found interesting results within our SNP dataset. Additionally to these two GO searches, we also identified a related NADPH-dependent aldo-keto reductase (YDR368W). Indeed, many related genes were present that were likely involved in the detoxification of aldehyde-like compounds, some of which gave moderately low p-values.

There are six other SNP-bearing genes within the study that share YKL071W's Alcohol Dehydrogenase (NAD+) activity (YeastGenome GO:0004022). Table 3.5 illustrates the GWAS results for these genes. As can be seen in the table, some of the genes with a GO term for Alcohol dehydrogenase activity have SNPs with moderately low p-values. YOL086C (Alcohol Dehydrogenase 1) has the lowest p-value in the PSIKO analysis, with YGL256W the lowest when using the SANE Q-Matrix. We can also see that p-values and their ranks differ between the two analysis methods.

A secondary GO term for YKL071W is oxidoreductase activity. In table 3.6, we can see SNPs within genes that share this GO term. In both GO-related sets of genes (tables 3.6, 3.5), we can see some genes possess SNPs with moderately low p-values. This is more pronounced for the SANE Q-Matrix analysis, where p-values for these genes are mostly lower.

Finally, we see different p-values across many genes with more than one SNP. For example, the gene YKR090W has many alternate alleles positively correlated with resistance such as YKR090W at bps 436, 591, 1289 and 1824 - though all with moderately high uncorrected p-values between 0.000611 and 0.00393. It also has one negatively correlated SNP (YKR090W at bp 1576, p-value = 3.1×10^{-2}) to resistance. While these p-values are not low, this is an interesting finding that perhaps could be followed up in a larger study.

3.4.7 SANE Specific hits

As an alternative method to PSIKO (or other software such as STRUCTURE etc.) for calculating population structure within a set of strains, the Simulating Ancestry through Nucleotide distance Equations (SANE) Q-Matrix was created. Using the sequences of each SNP genome, it attempts to predict ancestral populations based

on the input sequences and attributes a fraction of each strain's genome to a specific ancestor (see section 3.3). The TamDdistance measure was selected as most suitable for the yeast dataset and a Mantel test indicated the Q-Matrix was highly similar, though not identical, to that estimated using the PSIKO software.

The GWAS analysis was repeated using SANE, with results being compared to those found with PSIKO. As PSIKO possesses a well-documented and verified methodology cited by the scientific community [85, 86], it provide a useful baseline for comparison. Furthermore, given the high correlation between the two Q-Matrices, similarities and differences between the top hits identified by the two approaches will be interesting to uncover. In table 3.7, we see the top 10 hits identified with the SANE Q-Matrix which uses a genetic distance measurement (TamD [110] here) to determine founder populations.

From the top hits in table 3.7, we see several interesting points. Firstly, all genes are different from those in table 3.3. Secondly, all correlations are negative ones. Thirdly, the FDR-corrected p-values are all lower than 0.05 and their analogous values within the PSIKO Q-Matrix analysis. Finally, and perhaps most interestingly, we see 60S ribosomes are being identified on four occasions in three genes, YHL033C, YDR012W and YHL033C. This suggests that the method could be identifying real effects, although negative correlations. This indicates that these WT alleles may be better suited to furfural resistance.

ORF/SNP	p-value	FDR	+/-	ORF function
3249-YEL016C:745-G/	4.11^{-6}	0.345	-	Nucleotide pyrophosphatase/ phosphodiesterase; activity and expression enhanced during conditions of phosphate starvation; involved in spore wall assembly
3306-YEL072W:309-G/	1.47^{-5}	0.618	-	Protein required for sporulation
3951-YGR051C:286-A/	9.53^{-5}	0.826	+	Dubious open reading frame; unlikely to encode a functional protein
5704-YLR247C:759-T/	9.64^{-5}	0.826	-	E3 ubiquitin ligase and putative helicase; involved in synthesis-dependent strand annealing-mediated homologous recombination
3738-YGL093W:1905-C/	9.78^{-5}	0.826	-	Subunit of a kinetochore-microtubule binding complex
2849-YDR159W:1747-T/	1.05^{-4}	0.826	-	mRNA export factor; required for biogenesis of the small ribosomal subunit
4325-YHR072W:1047-C/	1.05^{-4}	0.826	+	Lanosterol synthase; an essential enzyme that catalyzes the cyclization of squalene 2,3-epoxide
6903-YOL105C:714-T/	1.06^{-4}	0.826	+	involved in response to heat shock and other stressors; regulates 1,3-beta-glucan synthesis
6903-YOL105C:717-C/	1.15^{-4}	0.826	+	involved in response to heat shock and other stressors; regulates 1,3-beta-glucan synthesis
7405-YPL056C:228-G/	1.16^{-4}	0.826	+	Putative protein of unknown function; deletion mutant is fluconazole (anti-fungal) resistant and has long chronological lifespan

Table 3.3: ***Top hits from the furfural GWAS (PSIKO Q-Matrix)***
First column combines the ORF ID with the systematic yeast gene name (with any gene duplicates denoted NOG) and the location and reference allele of the SNP. The second column displays the p-value of the SNP. The third columns gives the adjusted p-value using the FDR correction. The fourth column denotes whether the alternative allele is positively (+) or negatively (-) correlated with resistance. The fifth column gives a brief description of ORF function, taken from Alliancegenome.org.

ORF/SNP	p-value	FDR	+/-	ORF function
3951-YGR051C:286-A/	9.53^{-5}	0.826	+	Dubious open reading frame; unlikely to encode a functional protein
4325-YHR072W:1047-C/	1.05^{-4}	0.826	+	Lanosterol synthase; an essential enzyme that catalyzes the cyclization of squalene 2,3-epoxide
6903-YOL105C:714-T/	1.06^{-4}	0.826	+	involved in response to heat shock and other stressors; regulates 1,3-beta-glucan synthesis
6903-YOL105C:717-C/	1.15^{-4}	0.826	+	involved in response to heat shock and other stressors; regulates 1,3-beta-glucan synthesis
7405-YPL056C:228-G/	1.16^{-4}	0.826	+	Putative protein of unknown function; deletion mutant is fluconazole (anti-fungal) resistant and has long chronological lifespan
2699-YDR009W:579-C/	1.43^{-4}	0.826	+	Transcriptional regulator; involved in activation of the GAL genes in response to galactose
6447-YNL054W:3231-T/	2.04^{-4}	0.826	+	Transposable element gene
5045-YJR138W:687-A/	2.15^{-4}	0.826	+	Iml1p/SEACIT complex is required for non-nitrogen-starvation (NNS)-induced autophagy
4325-YHR072W:126-C/	2.31^{-4}	0.826	+	Lanosterol synthase; an essential enzyme that catalyzes the cyclization of squalene
7406-YPL057C:945T/	2.56^{-4}	0.826	+	Mannosylinositol phosphorylceramide (MIPC) synthase catalytic subunit

Table 3.4: ***Top, positively correlated hits from the furfural GWAS (PSIKO Q-Matrix)***
First column combines the ORF ID with the systematic yeast gene name (with any gene duplicates denoted NOG) and the location and reference allele of the SNP. The second column displays the p-value of the SNP. The third column gives the adjusted p-value using the FDR correction. The fourth column denotes whether the alternative allele is positively (+) or negatively (-) correlated with resistance. The fifth column gives a brief description of ORF function, taken from Alliancegenome.org.

ORF	PSIKO bp & Reference Variant	PSIKO p-value	SANE bp & Reference Variant	SANE p-value
YOL086C	382-C	9.2×10^{-3}	382-C	2.4×10^{-2}
YBR145W	817-G	2.4×10^{-2}	817-G	6.4×10^{-2}
YGL256W	1044-A	7×10^{-2}	179-A	2.6×10^{-3}
YDL168W	457-G	8.1×10^{-2}	457-G	0.33
YMR303C	416-A	0.14	416-A	4.6×10^{-2}
YMR083W	620-C	0.21	620-C	0.35

Table 3.5: ***Alcohol dehydrogenase-related genes with SNPs***
Genes with GO terms suggesting an Alcohol dehydrogenase function that possess SNPs in the GWAS. Locations of SNPs and their p-values are shown, both when using the PSIKO and SANE Q-Matrices.

ORF	PSIKO bp & Reference Variant	PSIKO p-value	SANE bp & Reference Variant	SANE p-value
YMR315W	336-T	3×10^{-2}	854-A	8.2×10^{-2}
YGL157W	274-G	4.4×10^{-2}	666-G	1.0×10^{-4}
YOR246C	66-A	6.3×10^{-2}	454-A	2.0×10^{-3}
YOR120W	297-G	0.12	700-G	0.24
YDL015C	147-A	0.16	609-G	2.0×10^{-3}
YKL195W	769-A	0.16	1094-T	1.3×10^{-2}
YMR226C	703-C	0.29	564-C	0.35
YOR037W	1001-G	0.30	818-C	0.25

Table 3.6: ***Oxidoreductase-related genes with SNPs***
Genes with GO terms suggesting an Oxidoreductase function that possess SNPs in the GWAS. Locations of SNPs and their p-values are shown, both when using the PSIKO and SANE Q-Matrices.

ORF/SNP	p-value	FDR	+/-	ORF function
2362-YCR014C:975-C/	6.86^{-7}	0.019	-	DNA polymerase IV
YIL155C-NOG-2:471-A/	9.91^{-7}	0.019	-	Mitochondrial glycerol-3-phosphate dehydrogenase
YIL018W-NOG-2:883-C/	1.21^{-6}	0.019	-	Ribosomal 60S subunit protein L2B; expression is upregulated at low temperatures
YGR161C:601-C/	1.36^{-6}	0.019	-	Retrotransposon TYA Gag and TYB Pol genes
YIL122W:162-A/	1.48^{-6}	0.019	-	DNA-binding transcriptional activator; involved in cell cycle regulation
YBR132C:945-A/	2.41^{-6}	0.019	-	Plasma membrane regulator of polyamine and carnitine transport
YHL033C-NOG3:303-T/	2.85^{-6}	0.019	-	Ribosomal 60S subunit protein L8A; mutation results in decreased amounts of free 60S subunits
YDR012W-NOG-2:396-T/	3.51^{-6}	0.019	-	Ribosomal 60S subunit protein L4B
YGR292W-NOG2:249-T/	3.51^{-6}	0.019	-	Maltase (alpha-D-glucosidase); inducible protein involved in maltose catabolism
YHL033C-NOG3:165-C/	3.51^{-6}	0.019	-	Ribosomal 60S subunit protein L8A

Table 3.7: ***Top hits from the furfural GWAS (SANE Q-Matrix with TamD distance measure)*** *First column combines the ORF ID with the systematic yeast gene name (with any gene duplicates denoted NOG) and the location and reference allele of the SNP. The second column displays the p-value of the SNP. The third column gives the adjusted p-value using the FDR correction. The fourth column denotes whether the alternative allele is positively (+) or negatively (-) correlated with resistance. The fifth column gives a brief description of ORF function, taken from Alliancegenome.org.*

3.5 A Directed Evolution study to improve furfural resistance in *Saccharomyces cerevisiae*

In an attempt to develop the desired resistance to furfuraldehyde in strains of the yeast species *Saccharomyces cerevisiae*, a directed evolution experiment was designed. In particular, it was hoped that some of the SNPs with low p-values in the previous GWAS study would be similarly identified within this experiment.

3.5.1 Experimental design

Fifteen NCYC strains were chosen as the basis for a Directed Evolution experiment. Together, these strains harboured 2,398 of the 84,046 SNPs with a frequency of > 5% (i.e. MAF SNPs) within the earlier GWAS analysis of 168 strains. The strain cultures were selected to satisfy a range of criteria, including high and low resistance scores, high MaxOD values and SNPs of interest (see table 3.8). The strains were arrranged in eight strain sets, some comprising a single strain culture while others were mixes of various strains. This design allowed for the possibility that a specific strain, or group of strains (which could potentially mate), would be optimal for developing a trait of interest. By allowing for multiple strains/strain sets, we hoped to identify the optimal mix for successful adaptive evolution.

The strains or strain mixes within table 3.8 were used to inoculate a 96-well plate as shown in the rows of figure 3.5.1, with each strain set placed in all wells of the given row. Initially, along each row, the strains were grown in a range of furfural concentrations, from 0.25mg/ml (column 1) to 6mg/mL (column 12), in 0.25 increments from 0.25mg/ml to 2mg/ml and 1.00 increments to 6mg/mL, with 10g/L of glucose in 6.9g/L YNB media (table 3.9). Strains were made to grow in the furfural media for 3-4 days (with 48 hours under a plate reader) at 25°C, then rested for 3-4 days in media without furfurals before repeating the experiment. This process allowed furfural resistance to develop, while not overly taxing the cells.

To direct the evolutionary paths taken by the yeast strains towards increasing furfural resistance, we selected the most resistant strains in each cycle of the experiment. This meant that the well with highest furfural content among those displaying yeast growth was selected as the best adapted strain and was placed to 'rest' in the rest media. If there was no replicate more resistant than the others, a

Set	*S. cerevisiae* strains	Rationale for inclusion
A	NCYC 221, NCYC 2777, NCYC 2780, NCYC 2798	Mix of strains with optimal SNPs (best resistance scores)
B	NCYC 357, NCYC 3078, NCYC 3338, NCYC 3039	Mix of conventional top strains (Highest OD; 5/5/6/8 resistance points)
C	NCYC 2967, NCYC 3456, NCYC 3472, NCYC 2733	Mix of strains which together have all SNPs of interest (low resistance scores, all top-scoring SNPs).
D	NCYC 221, NCYC 2777, NCYC 2967, NCYC 3456	Mix of High and Low Resistance strains with best SNPs (From B & C)
E	NCYC 357, NCYC 3078, NCYC 3315, NCYC 2733	Mix of high and low resistance strains (OD convention)
F	NCYC 2777	Top Strain
G	NCYC 3461	Bottom Strain
H	NCYC 2777	Control Top Strain

Table 3.8: ***Strain Mixes for a Directed Evolution experiment***
Fifteen strains chosen for a DE experiment based on early results from the GWAS analysis, including those with high and low resistance scores, high MaxOD values and possession of SNPs of interest.

mix of all replicates was added to the rest plate. This was to maintain a population with all potential genetic adaptations to furfural media. After three days in the rest plate, the colonies were re-established as groups A-G for the 96-well system, placed back into the various furfural concentrations and the process repeated. In this way, adaptation was tracked and encouraged, with any evolution in furfural resistance selected for until the final, adequately resistant, strain was formed. Therefore, if Culture B (i.e. the second row of the 96-well plate) showed growth at wells B7, B8 and B9), B9 was used as the source well to inoculate the B row in the rest plate. It was hoped that those cultures growing in higher furfural conditions possessed better adapted resistance genes, which would be multiplied in rest plate growth.

Optimisation of the experimental design occurred as a result of issues arising in early experiments. For example, the rest plate was added to the procedure following early occurrences of entire cultures dying from cell stress if placed in low furfural media following high furfural conditions. As strains developed better resistance, the period of rest could slowly be reduced and, finally, removed. In addition, significant changes were made to the concentration gradient. The initial stepwise increase from 0.25 to 6mg/mL (table 3.9) was updated as the increases

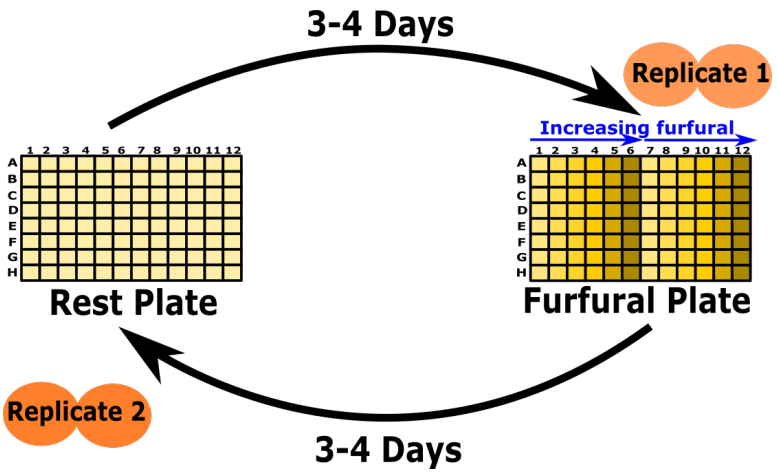


Figure 3.5.1: ***Directed Evolution Workflow***
Strains cycled through furfuraldehyde conditions (right) and resting conditions (left). As we rotate, with a rest plate or the exhausted yeast simply died, the strains gained in resistance phenotype.

at the start were too gradual, while later ones too steep, particularly as growth could not be achieved for furfural concentrations much higher than 3mg/mL. The new concentration gradient comprised just six steps, in increments of 0.5mg/mL. Firstly, it was hoped that the more incremental increase in furfural concentration would enable better adaptation and, therefore, resistance. Secondly, it permitted two biological replicates per plate, with the six concentrations repeated twice along each row of the 96-well plate. Therefore, for each row, the most resistant strain among the replicates could be carried forward to grow in the rest plate.

To allow for greater reproducability or possibly more opportunities for evolution, the whole experiment was run twice, concurrently. While one strain set rested, another (initially identical) strain set was growing in furfural conditions.

3.5.2 Analysing the results of the DE experiment

Once the two experimental runs had been completed, DNA was extracted from each final culture using the protocol in Chapter 2, with DNA quantities sought that would allow two sequencing runs per sample to be achieved. The DNA extractions were sent for the required duplicate DNA sequencing on the Illumina NextSeq platform at Quadram Institute Bioscience.

The resulting FASTQ datasets were deduplicated with BBTools and trimmed for adapter sequences, regions of low quality and low complexity using Trimmomatic v0.32. The trimmed FASTQ files were then used to identify high-quality SNPs against the reference pangenome, using the same conservative SNP-calling

Initial concentrations (mg/mL)	Updated concentrations (mg/ml)
0.25	1.5
0.50	2.0
0.75	2.5
1.00	3.0
1.25	3.5
1.50	4.0
1.75	1.5
2.00	2.0
3.00	2.5
4.00	3.0
5.00	3.5
6.00	4.0

Table 3.9: ***Furfural concentrations used in the Directed Evolution experiment***
Furfural concentrations used across 96-well plates in the initial Directed Evolution experiment (left) and modified concentrations used in the final experiment (right). The initial experiment was used to identify the conditions in which the yeast grew best; the final experiment used the updated concentrations once these conditions had been identified. The concentration gradient employed was modified to both linearise and update the concentration as most yeast were resistant to 1.5mg/ml of furfural and therefore the lower concentrations did not aid in differentiating the yeasts' variable resistance to furfural.

procedure as in the GWAS study. Variant calls that were not shared between biological replicates were computationally removed.

The OD data for every growth plate were also analysed using software scripts developed for the previous furfural resistance GWAS, with the same growth curve measurements taken as before. However, errors in experimental design (corrected with little time to continue the experiment further) resulted in the summed furfural Resistance Score increase being minimal.

3.5.3 Allelic Variation within the Directed Evolution strains

Through the final sequencing of the strains having undergone Directed Evolution, it was possible to compare their allele frequencies at specific SNP loci with those of earlier datasets. To make such comparisons, we essentially have three groups of alternate allele frequencies: all 168 NCYC *Saccharomyces cerevisiae* strains in the GWAS study (NCYC_Alt), the fifteen NCYC strains in the DE study (Selected_Alt) and fifteen sequenced strains at the end of the DE study (DE_Alt). The goal of this analysis was to see if particular allele frequencies either rose to fixation or were lost as a consequence of furfuraldehyde resistance. In figure 3.5.2, we can see this in action.

The genes used in the analysis were those top genes predicted by PSIKO. Namely, YEL016C (745G), YEL072W (309G), YGR051C (286A), YLR247C (759T), YGL093W (1905C), YDR159W (1747T), YHR072W (1047C), YOL105C (714T), YOL105C (717C), YPL056C (228G), YDR009W (579C), YLR305C (1662A), YNL054W (3231T), YJR138W (687A) and YHR072W (126C).

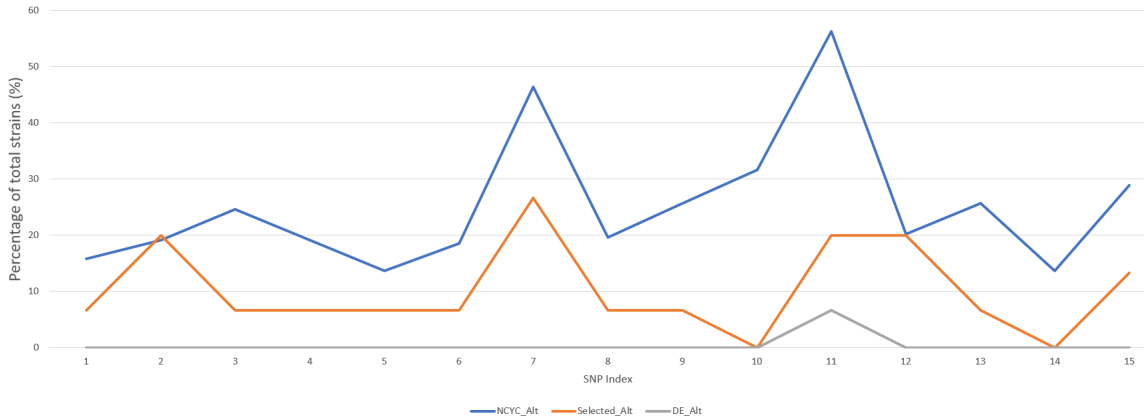


Figure 3.5.2: ***Alternative Allele Frequencies Within 3 Different Strain Groupings***
NCYC_Alt: All Saccharomyces cerevisae strains (168 strains; blue line)
Selected_Alt: Strains involved in DE (15 strains; orange line)
Evolved_Alt: Sequenced strains at the end of DE (15 strains; grey line)

From the figure, we see that the SNP frequencies of the 'Selected_Alt' dataset strains (orange line) are quite similar to those of the 'NCYC_Alt' dataset (blue line). They show that the Selected dataset represented its parent (NCYC) dataset quite well for these SNPs, even though it contains fewer than 10% of the total strains. However, the DE strains showed a quick fixation to WT/ref alleles, as seen by the 'DE_Alt' line in figure 3.5.2.

This result could be explained by the stability of the WT alleles, as they were quickly fixed across the board. However, two SNPs (bp 595, bp 607) within the YKL171W gene were raised to fixation within the DE dataset. A protein kinase with links to rapamycin, they could be SNPs linked to cell death (as rapamycin has anti-fungal effects and links to eukaryotic anti-ageing [133]).

3.6 Discussion

When measuring the broad-spectrum growth-inhibiting effects [102] of a chemical such as furfuraldehyde it is best to use holistic measures of overall resistance for a specific strain. In this way, it is possible to capture the overall effects that may be a mixture of slower growth, delayed growth and less total growth. The solution to

group and then combine different phenotypes attempts to overcome this limitation and account for a wider range of broad-spectrum effects. However, different phenotypes are not all able to be conjoined easily; even when all quantitative, they have different means, ranges and types (discreet vs continuous). Therefore, a system to compare like-for-like is necessary for true integration of disparate (but related) measures of resistance to broad-spectrum growth inhibition.

To create the generalised resistance score that carries across all strains, it was necessary to compare divergent growth curves. By selecting only for curve characteristics with a balanced distribution, it became possible to create K-means strain partitions that better grouped similarly resistant strains. Through the assembly of a general resistance score (on a scale of 3-18), it is easier to more directly compare strains in a like-for-like manner, across growth curves that vary greatly in the characteristics within those curves.

This phenotype construct was central to the approaches taken in this study. By using a holistic approach, it accounted for more variation within the initial datasets while also preventing 'true' resistance being masked by a single strain's anomalous characteristics. By adopting an equal number of groups for all growth curve features, we sought to integrate the many desired phenotypes of a strain and balance them fairly (i.e. length of lag phase, rate of maximal growth, total growth after a set time).

Moreover, the genomic component of the study utilised a broad array of software and tools. The raw read data's curation and cleaning, including subsequent filtering (e.g. removing unwanted variants such as indels) were done through well-referenced software [65, 67, 69, 87]. This ensured that the data was generated at the scientific standard. However, the final assembly of the gene matrices were carried out using wholly bespoke software and scripts developed within the study. Similarly, the software enabling the meshing of phenotype and genotype was a mix of custom scripts mixed with standard packages within the RStudio environment [69]. Lastly, the NMR preprocessing was accomplished through TopSpin3.6 using the apk0 phasing command.

Determining the success of the results relies on many factors. The GWAS analysis using the Q-Matrix estimated by the well-established PSIKO software did not indicate SNPs with corrected p-values above the 0.05 threshold. This could be a reflection of the size of the study, which at 168 strains was fairly small. Adding

strains would largely confuse the analysis; the only other high-quality sequenced strains were non-SC and any Q-Matrix would cluster all SC very tightly together and ignore most of their diversity.

However, in the analogous analysis using the SANE Q-Matrix - which was shown via a Mantel Test to correlate strongly with the PSIKO Q-Matrix - we did see a small number of SNPs with potential effects on furfural resistance, albeit connected to WT alleles. In particular, we saw a several SNPs connected to ribosomal proteins within the top hits of this analysis. Potential associations of a range of phenotypes with the ribosomal complex have been noted in previous GWAS analyses of yeast strains, so this could be related to successful growth under general stress conditions. Future work on validating the SANE approach would therefore be highly useful.

Despite the moderate p-values observed, viewing the top hits of our study (tables 3.3 and 3.7) could still be of value. The genes highlighted in table 3.3 include two (YGR051C, YPL056C) of unknown function. The second of these, YPL056C, is known to be involved in resistance to fluconazole, an anti-fungal agent. A mutation within the gene (bp 228 in this case) could plausibly confer a broader resistance to anti-fungals such as furfuraldehyde [121].

We also note that the YKL071W gene appears among the top 3% of hits within the SANE analysis with an uncorrected p-value of 0.00075174 at BP 616. YKL071W has been widely implicated in specific furfuraldehyde resistance studies [129] and was thus an anticipated result in this study. Its presence (bp 616) falls within the predicted short-chain dehydrogenase region from PANTHER (Accession: PTHR43544), lending credibility to a potential effect in this case. However, while YKL071W has been experimentally validated as a detoxifier of furfuraldehyde, an analysis of SNPs within genes that share its Gene Ontology (GO) terms show only moderate p-values within this study (tables 3.5 and 3.6). Nonetheless, that the SANE Q-Matrix identified the SNP is a good validation of strategy.

The Directed Evolution study successfully developed a series of strains that showed high levels of resistance to furfurals. Early analyses showed that important effects are related to WT alleles, when compared to the initial strain set or the wider strain set from which the starting strains were selected. However, the resulting sequence datasets have yet to be explored fully. In particular, we will investigate the prevalence of variants highlighted by the GWAS analysis within the highly resistant strains resulting from the DE experiment.

The next step of the overall study would be to build predictive models that utilise the top hits discovered, for example by engineering chosen variants in a model yeast strain using CRISPR. In this way, it would be possible to test and refine the predictions of a computational analysis with results in a living system.

The approaches followed and developed within this study have led to results that indicate a successful initial GWAS was undertaken. The verification of any results, either through expansion of the study dataset or via experimental validation of variants potentially underpinning the desired phenotype, would be a longer and more complicated process. However, perhaps more importantly, we have developed a dual computational and biological framework for the analysis of growth inhibitors in a key yeast species that can be used in future by ourselves or by others.

Even further, this study uncovers many novel results not otherwise discernable from the NCYC database data. With sparse habitat information, submitter information and even alternative names (many samples were submitted decades ago as unknown samples), the study provides new phenotype information to the NCYC. Another useful addition would be to contact old sample submitters to try to acquire more detailed strain habitat information. In providing new data, the methodology utilised to obtain the phenotypic data can be applied to the whole NCYC collection in a medium-throughput approach. Taking the study even beyond the scope of *Saccharomyces cerevisae* would help investigate the extensiveness of furfural resistance across diverse species. This novel information can also be added to the NCYC database, to increase interest from commercial buyers searching for strains with evidence of furfural resistance.

In any future DE experiment for the same phenotype, some corrections would be necessary. The experiment would have to be carried out for more generations, using the final stepped-concentration values (table 3.9). Additionally, a fine-tuning of the 'rest' period and single-strain isolation would have to be performed. Unfortunately, having carried out the DE at the start of the Covid pandemic, some errors in experimental design were not picked up upon in the confusion. With reduced lab time (and booking difficulties), correcting and re-doing experiments took a longer time than expected until we began to run short of time.

Chapter 4

Yeast Metabolite production under varied feedstock conditions

4.1 Introduction

Yeast are used throughout academic and industrial settings to produce platform chemicals at commercial scale. For example, bioethanol is produced by yeast as a renewable fuel across a vast array of industrial settings [101, 111, 134, 135].

In addition, yeast are used to produce a vast array of renewable chemicals. Especially when replacing petrochemical production systems, they present a renewable bio-based solution to finite, climate-damaging fossil fuels classically used for their production [95, 136, 137].

In addition to their broad usage in chemical production, yeast such as *Saccharomyces cerevisiae* have an extensive safety profile gained from global, ubiquitous usage for millennia [27, 138] coupled with highly diverse and adaptable genomes [46, 47, 51]. For these reasons, among others, *S. cerevisiae* strains are excellent production platforms for metabolite production. *S. cerevisiae* also produce almost all the molecules of larger, multi-cellular eukaryotes but come with the simplicity of being a uni-cellular species. Adaptable, resilient and relatively fast-breeding, they are excellent production platforms [6, 22, 28, 139].

The *S. cerevisiae* strains in this thesis were reduced to the strains with the highest read quality and depth of sequencing from the National Collection of Yeast Cultures (NCYC) to test for various metabolite expression patterns. Those strains that failed to sequence with adequate depth for SNP calling (>30) and high read quality were

excluded. This helped ensure a high quality in our downstream analysis.

The metabolomic requirements of each strain required by industry and academia is different. For thousands of years, even the relatively simple demands of brewers resulted in many different strains, each adapted to the specific needs of the individual brewers [140]. Further, funders and industrial producers have vastly different demands to the relatively basic demands of brewers.In academic and pharmaceutical environments, scientists have used yeast to produce antibodies for a range of diseases [25, 141]. This antibody production has vastly different demands on the metabolism of yeast than when *S. cerevisiae* are used for brewing alcoholic beverages [25].

A major focus of funders, for example, is the production of antibodies for medical applications. Most antibodies are produced in mammalian systems, often as immune responses to specific stimuli but also using human cells in human-animal chimeras [142]. However, this is not always possible without manipulating the host's immune system as well as issues with subsequent purification from the rest of the blood [142]. These mammalian systems also have the benefit of correct post-translational modifications such as with glycation and chaperone-assisted protein folding. However, financial and technical difficulties mean mammal production systems (cell lines included) make it difficult for smaller companies to compete [143]. This is where yeast can come in, with more cost-effective antibody production.

As a whole, yeast platform production systems are inexpensive and interest has grown in their antibody-producing capabilities that do not suffer from bacterial systems' incorrect post-translational modifications of proteins. Issues remain, but upscaling would be a relatively easy matter [25]. Therefore, *S. cerevisiae* present a potential solution to many industrial issues, from pharmaceutical and biomedical production to secondary metabolite products. This broad adaptability illustrates the usefulness of *S. cerevisiae* as a production platform. However, our focus remains on the small-molecule metabolites produced by the global *S. cerevisiae* populations.

In contrast to antibody research in *S. cerevisiae*, many commercial ventures require high expression rates with efficient feedstock conversion rates to produce high quantities of relatively low value products. Especially with products generated from low-cost non-renewable oil, the pricing of the produced chemical can be both low and fluctuating. This means any bio-based yeast alternative must ultimately

be both more reliable and cheaper than any oil derivative [95, 134, 137, 30]. This presents yeast production platforms with divergent needs; flavourful brewing, high-value biomolecules and low-cost platform chemicals. Even in these vague fields of interest, there is plenty of diversity. For example, brewers are no longer caught up in a race for high-efficiency ethanol-producing living machines. The new goal is for flavourful drinks with low alcohol content to fulfil the more health-conscious drinking habits of a new generation [144].

This chapter in the thesis, therefore, attempts to measure and predict metabolite production (especially of flavour and TCA cycle metabolites) in a GWAS setting. With a desire for low ethanol prevalent, the focus was to identify low-ethanol genetic markers for *Saccharomyces cerevisiae* strains that co-coincide with SNPs for flavour metabolites. As a side interest, it was hoped the entire metabolic composition of optimal strains could be identified, and their genetic components elucidated. In conjunction, strains could be created (or predicted) using genetic data to create low-ethanol producers that maintain a specific flavour profile.

To achieve this goal, it is necessary to measure the metabolic output of each strain. The supernatant (of strains fully-grown within malt media) of each strain can be assayed to investigate its components, measuring the metabolic output of a specific organism. However, traditional assays entail time-consuming, expensive experiments that struggle with accurate concentration measurement. A faster, more quantifiable and cheaper method was necessary to analyse the medium-throughput metabolic data of many *S. cerevisiae* strains growing in malt media common in the brewing industry.

To analyse the metabolites from an organism is expensive. Doing it quantitatively is difficult. In the past, the usual solutions involved assays to detect the presence of a metabolite or Mass Spectroscopy (MS) to determine the presence of metabolites in a sample (when coupled with various chromatography techniques). Although highly sensitive, MS is expensive, can suffer from reproducibility issues and lacks the quantitative ability of Nuclear Magnetic Resonance (NMR) while struggling with the similar molecular weight of small molecules [89].

In addition to data concerns, MS requires extra preparation steps such as separation. All of this takes time and money. In contrast, NMR is much more successful at quantitative analysis of metabolites with limited lab setup besides media preparation and centrifugation [89].

It is also much more amenable to medium-throughput analysis as machine setup is largely limited to cleaning NMR tubes and setting a program. For this reason, we selected NMR as the method to analyse the metabolomic state of our yeast media post-fermentation.

However, not all metabolites are to be analysed with the same purpose of high-conversion efficiency in mind. Due to increased public health awareness of the physical damage of alcohol consumption, not least of which relates to obesity and cancer, consumption of alcoholic drinks has decreased over time [145]. As people seek to reduce their alcohol consumption, they sometimes wish to maintain the social aspect of drinking, so beers with minimal alcohol content become desirable. In the past, many yeast were bred to produce ethanol as a necessary energy-management step of anaerobic growth. In newer studies, GMO techniques have been employed to create low-ethanol producing strains that may still be used for brewing previously alcoholic beverages [146].

When industries produce beer, they encounter the issue of elevated ethanol content in a period where low-alcohol beers are desired by the consumer market [145]. Therefore, a search for low-alcohol beers that maintain the same flavour profile of their more alcoholic cousins is highly desired by industrial brewers. We attempted to fill this societal niche by identifying low-alcohol genomic traits and strains in our yeast library.

To best match industrial environments, we chose malt extract media in which to grow our yeast cells, which were then analysed through quantitative NMR to measure the quantities of a range of flavour metabolites, including ethanol. We then made a comparison of growth under Malt conditions to those of laboratory YNB media with glucose as the carbon source. By comparing the two resulting datasets, we hoped to identify key differences in metabolites produced.

By the end of the project, it was hoped it would be possible to elucidate various genes associated with the production of specific metabolites. Matching these genes back to biomolecular pathways, desirable genetic abilities could then be identified. Therefore, simply sequencing a genome could be used to predict a given strain's general metabolite profile, and allow for rapid identification of strains to study experimentally.

Additionally, with the low-cost methods utilised, it could become possible to identify the genetic underpinnings of any metabolite. Therefore, with amply

available genetic tools, it could become possible to mutate any strain to possess the desired traits. Alternatively, it could be used to elucidate biomolecular pathways involved in the metabolic flux of the cells in question. Besides general research, this could aid in understanding the total metabolic flux of an entire cell culture.

Therefore, the aim of this research would ultimately be to find or create yeast strains capable of producing high yields of flavour compounds and other metabolites while maintaining low ethanol content. Perhaps the low ethanol status of these strains would also push more carbon towards production of the other desired metabolites.

4.1.1 Strains in study

To discuss what was done with the data, it is important to mention the strains included within the study. The full data is available in the GitHub repository, with full metabolite data. In the lists below, the 'NCYC' designation is excluded from each strain number but is common to every strain used.

Full strain lists used for each study (YNB/Malt) is included in the Appendix (table 3). Species is included for clarity; there were 168 *S. Cerevisiae* strains in the Malt study, with 50 strains shared with the YNB strain dataset which contains 362 total strains.

4.2 Metabolite Laboratory and Computational Methods

4.2.1 Mapping and matching SNPs with phenotypes

The mapping of genome specific SNPs to the metabolic profiles of each strain was done by borrowing the GWAS methodology used in Chapter 3. The full genomic process is explained in the Methods chapter, in Section 2.2.

Rather than a weakness of analysis, this usage illustrates the versatility of the methodology and highlights its scalability. Similarly, this allows like-for-like comparison of SNPs highlighted in the furfuraldehyde resistance study. Using the same method, and base DNA data, we ensure that any differences or similarities identified are not due to changes in the underlying data.

4.2.2 Yeast protocol

The protocols for growing the NCYC yeast strains in both YNB and Malt media were identical. This was to reduce the amount of variability between strains' expression patterns. The only variable changed within any experiment carried out in this section was the media in which each strain was grown- with only *Saccharomyces cerevisiae* grown in the malt media and not all *Saccharomyces cerevisiae* grown in the YNB. This partitioning was due to time constraints, whereby we attempted to prioritise a broad spectrum of strains for the YNB growth, while ensuring all *Saccharomyces cerevisiae* strains were grown in malt media. This allowed for the focussing on important results required by researchers and industry, respectively.

The protocol was fairly straight-forward, outside of creating buffers and media. The 96-well plate was grown with 31 yeasts in triplicate for 5 days anaerobically at 25°C, with media as either YNB or malt extract as appropriate. This time period allowed for the yeast strains to achieve their final growth stage. The media was then ready for extraction.

The plates were subsequently centrifuged at 3000rpm for 15 mins (cells sink to bottom as pellets) and 400µL of supernatant from the wells was added to individual micro centrifuge tubes. 400µL NMR buffer was added to each sample (section 2.5.2) micro centrifuge tube to a total volume of 800µL. Tubes were then centrifuged at 3000rpm for 15 mins again (insurance). The top 600ul of the supernatant was moved to a clean NMR tube and placed in the 500 MHz NMR machine. The spectra were then cleaned on TopSpin (Bruker NMR Software) and then quantified with the CHENOMX software (CHENOMX Inc, Canada).

The protocol demanded 5 days of growth to reach the predicted end of growth stage, and to use most of the energy source. This functioned well for the YNB media with low glucose. However the maltose was not completely metabolised due to its very high concentration. The culture was nonetheless pelleted, mixed with buffer solution in a 1:1 mix and then centrifuged again before being placed in the NMR for quantification. The buffer was essential as it provided a reference compound (TSP) of a known final concentration (2.5mM) with D2O to calibrate the NMR spectra (section 2.5.2).

The pelleting used within the protocol might have compressed and extracted some content from within cells, but likely most of the metabolites quantified were

those secreted into the supernatant. This is important, as any chemicals essential for cell life are unlikely to be fully secreted and would skew any quantification analysis. This probability was tested with ATP/ADP; if the energy carrier, ATP, was found in the supernatant this would only have been the product of cell lysis. Consequently ATP was used as a biomarker for cell lysis.

4.2.3 OD analysis for confirmation of growth

For the malt metabolite experiment, a simple binary test of growth/lack thereof was necessary. Therefore, visual checks were all that were necessary to verify each strain could grow in the media. Once a strain was confirmed to grow in the malt media, it could be (in separate replicates) quantitatively measured for its metabolic production. It was discovered that a slight alteration in temperature (27°C compared to 25°C) resulting from a faulty incubator significantly affected the strains' glucose metabolism, so the corresponding datasets were discarded. Subsequently, once a strain was demonstrated to grow adequately within the malt media, the experiment to quantitatively measure the metabolic products of the strains in malt media could be commenced.

4.2.4 Gene variant identification

Comparing the metabolite expression levels (tables 4.2 and 4.3) of each strain mapped to their SNP genomes, it is possible to identify SNPs correlated with specific metabolite production. Using this technique, a list of SNPs could be generated for each metabolite. Subsequently, a strain's metabolic profile could be predicted based on the presence or absence of said SNPs. Potentially, a strain could also be engineered to possess desirable SNPs in a quick, single-generation mutation.

Therefore, with a full SNP map for yeast strains, it could be possible to identify strains of metabolic interest based upon a SNP profile. This first computational step would massively reduce the number of strains to analyse experimentally. With a library of sequenced strains, a predicted metabolite profile could be constructed for each strain. By then refining through each experimental validation, the model would build in accuracy and predictive value. For metabolites where there was a limited cross-over between genes involved in each metabolite, the SNP profile could potentially more easily predict the metabolic profile.

Any company with sufficiently large genome databases might then assess yeast strain suitability for a specific metabolic profile based on genomic data alone. For example, a brewer might choose to create a beer with low ethanol, high acetoin and low acidity compounds by selecting for or breeding/engineering a strain with the desired allelic combination.

However, for predictions of genomic variations to be of the highest accuracy and quality, other genomic variation should be included within the predictive model if possible. For example, the Copy Number Variations of specific genes or the aneuploidy of entire chromosomes, both of which have broad phenotypic effects in various organisms. These factors are known to have large effects on yeast phenotypes and should be carefully considered and integrated into any model for the full predictive effects of genomic variations to be utilised [11, 22, 46, 47, 51].

4.2.5 Metabolite GWAS

To perform a GWAS on the metabolite data to pinpoint SNPs potentially involved in each metabolite's level, it was necessary to couple the genetic data of each strain to their metabolite quantity. The SNP data used in this study was simply a re-usage of pipelines constructed for the elucidation of the genetic basis for furfural resistance (section 2.2). Therefore, it had few issues in implementation, validation and analysis.

With over 80,000 SNPs used as inputs to the GWAS, a broad swathe of *Saccharomyces cerevisiae*'s biological pathways were tested for correlation to the phenotypes. Reference alleles were again being quantified as '0', while non-reference (alternative) alleles were '1'. Therefore, a positive correlation to increasing metabolite quantity indicates the alternative allele at a specific locus was potentially contributing to increased metabolite quantities.

An interesting note would be the potential cross-over between metabolites within strains. For example, two metabolites with the same genes (and perhaps SNPs) contributing to their quantity would complicate future metabolome engineering. A mutation for one desired phenotype (e.g, high succinate) might cause an undesired phenotype (e.g, high ethanol). This would make all future implementation more difficult.

However, as the metabolome is one whole system, some overlap is unavoidable.

With finite carbon sources, an increase in one metabolite necessitates a decrease in others. Avoiding a single pathway that directly affects two metabolites might also be undesirable; a SNP in a single gene might, in some cases, confer more than one desired phenotype. For example, it could reduce an enzyme's specificity for one metabolite's precursor while increasing the specificity for another. This would shuttle carbon from a more even distribution towards producing mostly one outcome. Therefore, overlap of genes or SNPs is not necessarily good or bad; it is the effect of an SNP on all metabolites that is of interest.

4.3 Media Usage

This study attempted to evaluate the metabolic productions of diverse yeast strains in two different media. In practice, this involved pelleting the cell biomass, and extracting supernatant to mix with a sterilising NMR buffer (section 2.5.2). The samples were then analysed with a 500MHz NMR machine to obtain spectra peaks which were cleaned with topspin3.6 using the 'apk0.noe' command. The cleaned spectra were in turn analysed with CHENOMX.

CHENOMX is a software package for analysing NMR spectra from a range of frequencies with standard libraries provided with hundreds of chemicals. When given a reference compound, and its concentration (e.g, TSP), spectra can be measured quantitatively. CHENOMX is a versatile tool that is capable of automatically assigning concentration values to spectra peaks. However, it can struggle with automatic peak assignments for compounds of smaller concentrations or unexpected media properties (e.g, incorrect pH input) where the NMR spectra may deviate from the norm. It is thus necessary to manually curate areas of low-fidelity.

This limitation is not an issue for high-concentration compounds within a media. For example, ethanol is usually highly expressed in yeast strains (especially within the *Saccharomyces cerevisiae* species) and thus easily identified through automatic annotation through CHENOMX. However, 'flavour' compounds and other secondary metabolites are usually present in lower concentrations and are thus occasionally miss-annotated in the automatic spectra-fitting.

4.3.1 Metabolomic Variation in YNB Media

Each strain was grown in 6.9g/L Yeast Nitrogen Base (YNB) media with 10g/L glucose. In table 2.2, we can see the components of the media when made up to the required concentration. The metabolome was subsequently analysed quantitatively to measure the variation of each metabolite per strain. Using CHENOMX (CHENOMX Inc) ensured accurate, unbiased metabolite quantification. Performed through medium-throughput NMR analysis of the supernatant, it permitted a holistic view of each strain's metabolome.

With the YNB media, and diverse yeast strains employed, we hoped to gain a broad view of the metabolome in academic/industrial research environments. With everything needed to grow, but falling short of thriving, the YNB media should give us a good 'baseline' production. Every strain should comfortably grow in the media and provide good metabolic overviews.

The extensive list of maxima, minima and the standard deviation of expression between all strains' metabolite expression is displayed in table 4.2.

4.3.2 Metabolomic Variation in Malt Media

With a view to analyse the quantitative metabolome of *Saccharomyces cerevisiae*, 168 *S. cerevisiae* strains were grown in malt media. These strains were then analysed for metabolites to then attempt to identify SNPs correlated with the phenotypes of interest. The malt media was chosen to represent conditions seen in the brewing industry, who desire a low-ethanol malt-metabolising strain. The *Saccharomyces cerevisiae* strains had little issue metabolising the abundant maltose (glucose disaccharide). However, the remnant amount of maltose, and glucose, varied per strain.

The extensive list of maximums, minimums and the standard deviation of expression between all strains' metabolite expression is displayed in table 4.3. This makes it easier to visualise the breadth of variation present even within this *Saccharomyces cerevisiae* dataset.

4.4 Metabolomic profiles of the Yeast Strains

When considering an 'optimised' strain, there are multiple factors to take into consideration. We will focus on three main factors; carbon source metabolism, major metabolite production (ethanol/glycerol/methanol/...) and all other, often nearly trace, metabolite concentrations. The data from table 4.2 had anomalously high metabolite-concentration strains removed as a precaution against NMR errors. While all biological triplicates were comparable, there was not time to re-run all the samples.

From tables 4.2 and 4.3, we see a broad view highlighted within the specific examples above (all SNPs and metabolite tables provided on GitHub). The Malt media yeast strains produced higher amounts of flavour compounds (citrate, acetoin) and had the stronger statistical correlations to illustrate it. This is expected, as brewers would select conditions that increase the production of these rarer flavour compounds, from selectively growing specialised *S. cerevisiae* strains to creating media better suited to producing these rarer metabolites (malt media). The diverse, largely non-*S. cerevisiae* strains grown in YNB media could not compete with the generations of selective breeding in optimised media for the production of rarer flavour compounds.

4.4.1 Carbon source efficiency

An important metric when evaluating a strain as a production platform is its ability to efficiently metabolise various sugars. A strain's ability to produce a maximal amount of a desired metabolite might depend heavily on its ability to metabolise its carbon source. A sugar poorly matched to a strain would result in low output yields, giving an unfavourable view of the strain as a production platform. Maximising carbon metabolism is essential to pursing high yields [2, 42, 113, 135, 147].

This is doubly important, as often an industrial process will capitalise on a cheap feedstock to increase its margins of profitability. Feedstocks can vary widely, with various characteristics (e.g, lignocellulosic, whole-crop wet mill and green grass feedstock) and methods to pre-treat them (heat, acid, long water baths, alkali treatments). All of these different feedstocks and pre-treatments result in greatly varied potential feedstock mixes (and hurdles for growth) for any organic system [148].

When using a living production platform it is important to ensure that the organism is able to readily digest the input energy source. Growth of an organism on one media is not indicative of growth on another media. For example, it is not necessarily true that *Actinosynnema pretiosum* optimised to produce ansamitocin with fructose will be able to do so with glucose [6]. This is important because as feedstock prices change, it becomes more attractive to investigate the adaptation of strains to alternate carbon sources, and the interest in organisms capable of digesting a wide array of feedstocks becomes highly desired [147].

Lately, volatility in renewable feedstock prices is un unavoidable fact. Weather such as rainfall, snow depth, soil temperature, and soil moisture content, and many other factors influence the final cost of feedstocks [149]. Prices of feedstocks therefore vary (costing up to 38% more when weather is ignored [149]), and being able to not only manage these changes but benefit from them would be a great support to the economic viability of any organic chemical production system.

For example, woody residues dominate the biomass market but, as prices rise above 50 USD per dry ton, the economic feasibility of value extraction is greatly reduced and industries pivot to other carbon sources that are easier to digest- such as corn stover and wood itself. Therefore, it is essential to be able to either use many feedstocks as standard (to reduce price volatility from any one feedstock price fluctuation) or possess the ability to quickly pivot to other, cheaper sources [150].

For these reasons, we include in our analysis the final concentration of the carbon source (glucose/maltose), as it is an important variable; how effectively can the strain utilise our carbon source? A strain with little carbon digestion and low desired metabolite production is not necessarily ruled out as being of interest.

4.4.2 Major metabolite production

Each strain's metabolome is highly complex and variable in expression levels [147]. However, there are some metabolites that are produced by most yeast strains that represent most of their carbon output. The two main products are closely related to glycolysis and energy conversion during fermentation conditions; ethanol and glycerol. These two metabolites generally act as huge carbon sinks for the organism, using much of the carbon metabolised. Generally, a drop in the expression of these two metabolites is matched by a drop in carbon source metabolism.

It is also necessary on a practical level to separate these higher-expression metabolites (including medium-level metabolites such as malate) from those less-expressed. Without this separation, heatmaps of expression quantities for the main metabolites would be difficult, or impossible, to differentiate from other metabolites.

Lastly, a strain that acts as a poor expression platform for the major metabolites might be ideal for the production of other, rarer, chemicals. The biochemical pathways would shuttle less carbon to the usual carbon sinks, and the organism would instead more readily convert its carbon to the desired metabolite.

4.4.3 High value and lower expression metabolites

When considering microbial production platforms, not all outputs are created equal. It is obvious that the economic feasibility of a production process relies on two factors (assuming production and fixed costs remain constant); feedstock prices and value of the produced chemical. Each feedstock is different, and each presents difficulties and opportunities [150]. Economies of scale, using a single feedstock, also play a large role in economic viability [151]- perhaps due to the investment needed to extract the full value from a carbon source. Trying to swap feedstock needs a re-invention of the entire biorefinery [150], from organism used, to pre-treatments and specialised machinery. Due to the difficulties involved in extracting energy from the carbon source, trying to express the highest-value end product is natural.

While bioethanol production has been the focus of much research, it is generally low-value and has to contend with the fluctuating pricing of oil-derived fuels. Other biofuels suffer from similar issues, with the added difficulties involved in lipid-based products [29, 152]. However, as fossil fuel reserves deplete and climate change worsens, it is imperative to be prepared for future economic environments [29, 152, 153].

Therefore, research often focusses on attempting to produce higher quantities of rarer metabolites [137] and getting them up to theoretical maximum yields. Here, we have carried out two different feedstock studies to analyse some of these desired phenotypes. However, their genetic components contributing to these higher expression levels are not to be taken in isolation.

Just as important to the production of high amounts of a rare metabolite is the high metabolism of the base carbon source such as glucose, glycerol, or other carbon sources (section 4.4.1) and the expression of high-quantity metabolites such as ethanol (section 4.4.2). That is to say, it is nearly impossible to achieve high fractions of theoretical maximal yields if most of the carbon source goes either undigested or shuttled into lower-value metabolites. Additionally, the purer the final solution, the simpler the eventual purification and extraction of the desired metabolite is to perform.

In conclusion, it is not possible to look at a metabolite in isolation. Each metabolite exists in a complex network with every other metabolite in an organism's biochemical pathways- in aggregate, its metabolome. As such, it is necessary to take a holistic view of the metabolome when considering the optimisation or genetic modification of an organic production platform to express elevated levels of a desired metabolite.

4.5 Results

4.5.1 Overview of Metabolite Correlations

Two tables were constructed for the metabolite analysis. Table 4.2 detailed the highest and lowest metabolite concentrations for single strains within the study with YNB media. This study included many diverse yeast strains of various species. Table 4.3 attempts a similar study, but for *Saccharomyces cerevisiae* strains grown in malt media.

There are many differences between the two studies in terms of metabolite concentrations. Generally, however, the malt extract study (table 4.3) has higher concentrations than those in the YNB study (table 4.2). This is likely simply due to the abundant carbon source (maltose) in the maltose extract.

The YNB study had 10g/L of glucose, while the malt study contained approximately 100g/L of maltose. This is a hugely significant increase in carbon availability. However, even with this huge difference, expression only increased by a factor of approximately two to three. This could mean that an energy source was no longer the rate-limiting step of cell growth in the malt media, as it could have been in the YNB media. This factor is shown clearly by the remaining glucose/maltose

levels. The glucose concentrations in the YNB was near-zero for many strains(table 4.2), while the malt growth strains never dropped maltose concentrations below 9mM (table 4.3).

At a glance, the results of these two studies can be a challenge to interpret. However, heatmaps enable easy identification of relative metabolite expression levels. In heatmaps, all values are compared together and each value is assigned a colour for its expression level relative to other metabolites. That is to say, a heatmap interprets the relative expression of each metabolite per strain.

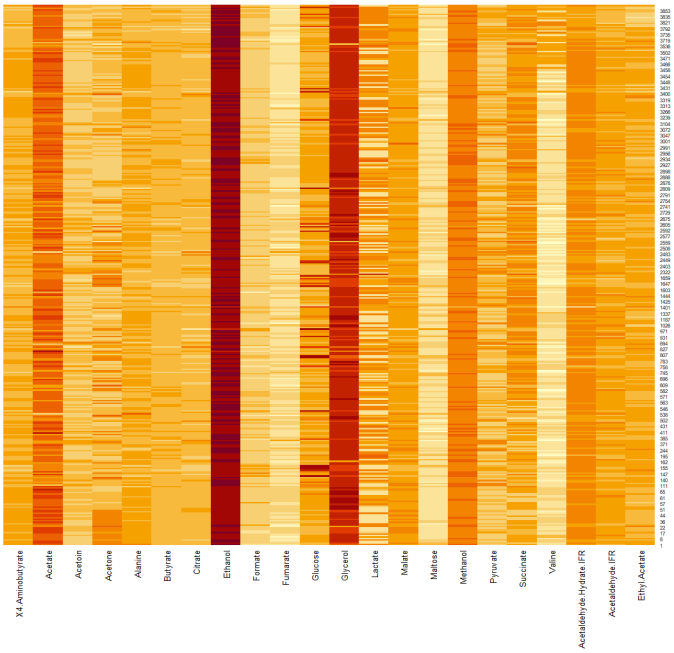


Figure 4.5.1: ***The log10 expression of various metabolites of strains grown in YNB media*** *In this figure, we can see how ethanol is highly expressed, with some glucose being mistaken for maltose.*

Figure 4.5.1 shows the log10 metabolite expression of the YNB study. In it, we can see clear patterns of expression. When ethanol production is reduced and glucose metabolism is high, glycerol expression is elevated. The Pearson correlation between glucose and glycerol is -0.168, glucose and ethanol is -0.368 while glucose correlated to glycerol and ethanol combined is -0.440. This indicates that both metabolites are a significant carbon sink from the glucose carbon stock. There is also a single strain (NCYC 820) with extremely high acetate expression which replaces ethanol.

Figure 4.5.2 shows the log10 expression of the Malt media metabolites from the second, Malt study. Compared to figure 4.5.1, maltose is clearly present in the media. The YNB media did not contain any maltose. Additionally, we can see

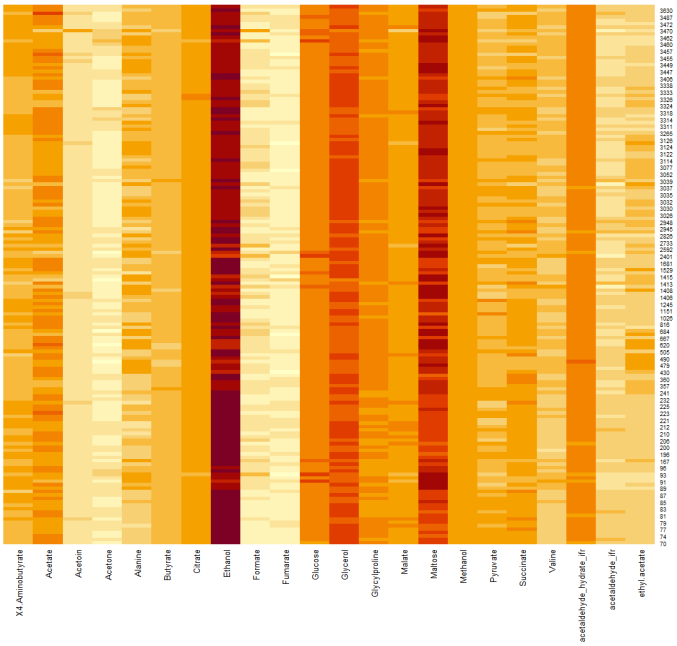


Figure 4.5.2: ***The log10 expression of various metabolites of strains grown in Malt media***
In this figure, we can see that ethanol expression appears to depend on consumption of glucose and its disaccharide maltose

changes in expression for other metabolites such as malate.

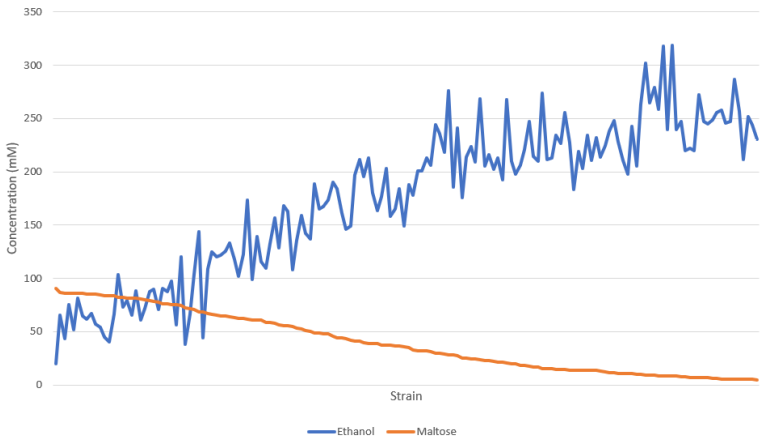


Figure 4.5.3: ***Maltose left in cell vs Ethanol produced***
In this figure, each Saccharomyces cerevisiae strain (168 total) is a data point; graph ordered by decreasing malt concentration. In the final quantification of metabolites after all growth, we see that the strains that have consumed the most maltose have produced the most ethanol (r = -0.936709864)

From figure 4.5.2, we can generally see that strains that struggled to metabolise all the maltose had a correspondingly lower expression of ethanol (figure 4.5.3). This is expected. Interestingly, our amino acids tested (Alanine r = 0.90, Valine r = 0.87) decreased with maltose levels (table 4.1).

Encouragingly, we can see clear increases in the flavour compounds succinate (r = -0.667), malate (r = -0.543) and butyrate (r = -0.705) (table 4.1) as maltose

decreases. This is encouraging for researchers who wish to uncover the genetic underpinnings of these metabolites as clear links exist between carbon consumption and metabolite production which can be exploited- instead of being a metabolite that always maintains a constant expression level.

Importantly, these high correlations indicate a high variability in metabolite concentration between strains. High phenotypic variability is essential to determining the genetic underpinnings of phenotypic traits (i.e, metabolite concentrations). This could explain why malt media, which has more significant correlations between carbon source consumption and metabolite production, has more low p-value SNPs that correlate with these metabolites. Some specific examples are detailed within section 4.5.3.

4.5.2 Metabolite Quantities

The next stage in the analysis was comparing the metabolite data from the malt study to the SNP genomes constituting the malt strain dataset. This would result in an indication of the relative importance of each SNP to a specific phenotype (in this case, metabolite expression levels).

In designing a strain with the desired metabolic profiles, it is necessary to consider three main variables; carbon source digestion efficiency, genomic variants affecting the concentration of major metabolites and, finally, the genomic variants affecting the concentration of the desired phenotype (section 4.4).

As such, we will detail any SNPs potentially involved in expression levels of the metabolites desired, while selecting strains with high carbon efficiencies. A final modification of the strain could involve the addition of previously discussed SNPs for furfural resistance to aid in the digestion of sugars present in pretreated lignocellulosic waste (table 3.3).

We identified SNPs strongly correlated to both ethanol and succinate expression levels (section 4.5.3). With these two metabolites, combined with information on carbon source utilisation [2, 42, 113, 135, 147] and furfural resistance SNPs (table 3.3) we hope to build information that could lead to the design of various yeast strains used in a range of production pipelines.

When deciding which metabolites to include within the final analysis, any metabolites from Tricarboxylic Acid cycle (TCA) cycle were thought to be useful.

Metabolite	Malt Correlation	YNB Correlation
4-Aminobutyrate	0.192	-0.221
Acetate	0.005	-0.048
Acetoin	0.064	0.054
Acetone	0.057	0.096
Alanine	0.904	-0.122
Butyrate	-0.705	-0.127
Citrate	0.275	-0.061
Ethanol	-0.937	-0.367
Formate	0.373	0.436
Fumarate	-0.032	0.036
Glucose	0.262	1.000
Glycerol	0.000	-0.168
Glycylproline	0.604	-0.031
Malate	-0.543	-0.124
Maltose	1.000	0.421
Methanol	0.245	-0.17
Pyruvate	-0.447	0.024
Succinate	-0.667	-0.110
Valine	0.874	-0.049
Acetaldehyde Hydrate (QIB)	0.236	-0.096
Acetaldehyde (QIB)	-0.546	-0.213
ethyl acetate	0.361	0.100

Table 4.1: ***Metabolite correlations per media to main carbon source***
The first column is the metabolite being analysed. The second column is the correlation of these metabolites in the malt media to the levels of maltose in the same media. The final column is the correlation of these metabolites in the YNB media to the levels of glucose in the same media.
This table gives a quick glance at metabolite conversion from carbon source. If the correlation is negative, the metabolite is produced as the carbon source is consumed. If the correlation is positive, the reduction of the carbon source correlates with a drop in the metabolite, perhaps due to consuming the metabolite as an energy source as the level of carbon source remaining decreases.

Containing many high-value metabolites and, as a core biochemical pathway, found in all yeast, the TCA is of immense interest. As such, figure 4.5.4 highlights all the metabolites included within the analysis. While not the only metabolites comprising our analysis, they represent a resource of huge value and interest.

To achieve a broader view of all the TCA cycle metabolites analysed, a figure was crafted highlighting the entire TCA cycle. In figure 4.5.4, we can see the different compounds analysed in our study. Green represents potential metabolites that we did not pursue for further analysis. Meanwhile, blue represents the metabolites selected for analysis and red the metabolites without spectra peaks present within the CHENOMX software library.

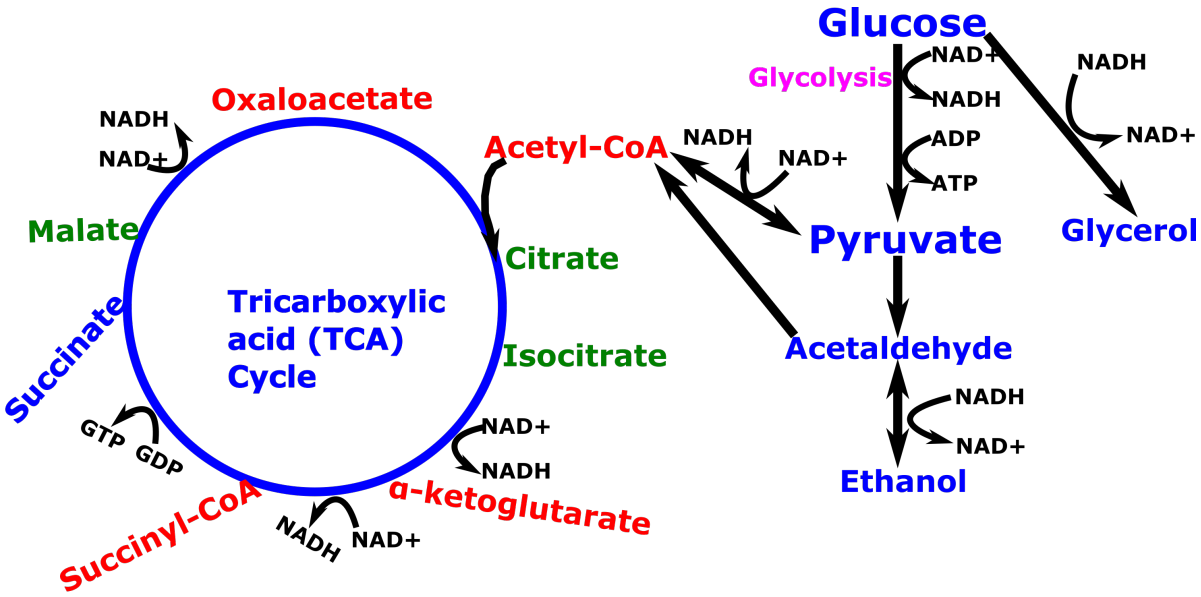


Figure 4.5.4: ***The TCA cycle with relevant biomolecules***
Biomolecules present in CHENOMX software for metabolite analysis (green), fully analysed (blue) and molecules with identification difficulties (red). A quick overview reveals that many of the TCA molecules themselves are either not in the CHENOMX compound library or too messy to be used (red). By contrast, glycolysis metabolites are highly represented in the compound library (blue).

While a heatmap is useful for a broad overview of the data, it is less helpful for the concrete evaluation of specific metabolites. Therefore, a table was created with all the metabolites within the YNB media. The variation of metabolite quantities between strains, which were from diverse species, was extensive. The variation of across species grown in YNB media metabolites was summarised within table 4.2. This analysis contained a few *Saccharomyces cerevisiae* strains, but mostly consisted of strains from other species.

Metabolite	Minimum	Maximum	Mean-+StD	Fold Diversity
4-Aminobutyrate	0.00260	0.0726	0.0244-+0.014	27.900
Acetate	0.00270	22.700	0.7120-+1.31	8420.0
Acetoin	0.00113	0.1210	0.0118-+0.0163	107.00
Acetone	0.00140	1.3700	0.0538-+0.126	980.00
Alanine	0.00153	0.3550	0.0297-+0.0258	232.00
Butyrate	0.00295	0.0382	0.0152-+0.00544	12.900
Citrate	0.00140	0.3060	0.0166-+0.0282	219.00
Ethanol	6.96000	46.000	32.400-+7.07	6.6100
Formate	0.00000	0.0964	0.0070-+0.0092	–
Fumarate	0.00000	0.0527	0.0029-+0.00525	–
Glucose	0.00210	32.800	0.9080-+3.97	15600
Glycerol	0.01800	38.600	5.5900-+4.96	2150.0
Glycylproline	0.00000	9.5800	0.1030-+0.505	–
Malate	0.00437	0.3920	0.0560-+0.0554	89.700
Maltose	0.00000	0.0399	0.0020-+0.00394	–
Methanol	0.09140	0.2880	0.2420-+0.0291	3.1500
Pyruvate	0.00177	0.2050	0.0264-+0.0382	116.00
Succinate	0.00073	0.7610	0.0804-+0.0799	1040.0
Valine	0.00000	0.0994	0.0067-+0.0174	–
Acetaldehyde Hydrate (QIB)	0.0066	0.2890	0.1090-+0.0519	44.000
Acetaldehyde (QIB)	0.00250	0.1290	0.0445-+0.0252	51.600
Ethyl acetate	0.00125	0.1330	0.0317-+0.0226	107.00

Table 4.2: ***Metabolite (and carbon source) Concentrations (mM) analysed through quantitative NMR after growth in YNB (+glucose) media.***
Highest and lowest performing strain values shown alongside standard deviation in expression across all strains in the study. The relatively high variability between low and high concentrations (fold diversity) is often explained by a very low minimum in a strain. Acetaldehyde metabolites' spectra added to compound library from the Quadram Institute Bioscience (QIB). Numbers rounded to a s. f. number to align values.

As not exclusively *S. cerevisiae* strains, ethanol production varied from a high of 46 mM with strain NCYC 17. Known more widely as *Hanseniaspora valbyensis*, it is a grape must yeast well known to have high ethanol producing capabilities [154]. The species with the lowest level of ethanol production, at just 6.96 mM, was *Rhodotorula graminis* (NCYC 1401), a species known for very high lipid expression levels [153].

These results are expected for these strains, as *Rhodotorula graminis* is known as a strain with very low nitrogen but very high lipid content of up to 54% w/w with glycerol as a feedstock. This extreme lipid expression is being investigated as a production platform for biodiesel. Its low ethanol production is expected, as

most of the carbon would have been sequestered to other biochemical pathways (i.e, lipid production) [153]. Meanwhile *Hanseniaspora valbyensis* is an expected high ethanol producer. Found in grape must, and the first step in grape fermentation, high ethanol production for the future wine is expected [154]. Other metabolites of interest include Succinate, which varied from a near-zero of 0.00073 mM to a peak of 0.7610 mM.

This variation hints at the taxonomic and metabolic diversity encompassed within the NCYC, with the evolutionary constraints of each strain's home environment resulting in specific metabolomic profiles. Even in the same environment, the genome of each strain causes vastly different metabolomic outputs.

As with the YNB media, there was a high degree of variation between the metabolomes of the strains within the Malt dataset. Unlike the YNB media, which contained many diverse yeast species, the malt media growth was achieved with solely *S. cerevisiae* strains. The strains grown in the malt media presented varied metabolomic profiles, illustrating their genetic diversity even within the *Saccharomyces cerevisiae* species. As an example of this metabolite diversity, ethanol varied from a low of 20.1 mM in strain NCYC 2041, with a high of 318.6 mM in strain NCYC 1413. This huge variation presents the potential for SNP identification that affects expression levels. Both strains are isolated from wine making processes, which implies one strain is better adapted at growing on the malt media.

The huge diversity in species used in the earlier YNB growth study was readily grasped in table 4.2, where we saw huge fold-change differences between the lowest and highest producing strains of a metabolite. That strain dataset included both *Saccharomyces cerevisiae* and many diverse yeast species. In contrast, table 4.3 contains only *Saccharomyces cerevisiae* strains. This difference in diversity is clearly represented in the metabolite quantities. For example, all *Saccharomyces cerevisiae* strains are generally assumed to produce high quantities of ethanol. In contrasting the two tables, we see the *Saccharomyces cerevisiae*-only table 4.3 has an 11.3-fold change between the lowest and highest glycerol producers. Table 4.2, by contrast, has a 2150 fold difference (due to a very low base).

It is interesting to consider these results, as different media appear to affect fold diversity in metabolite production between strains. As media change, the efficiency of a yeast strain on the carbon source becomes an important variable.

Metabolite	Minimum	Maximum	Mean-+StD	Fold Diversity
4-Aminobutyrate	0.07230	0.364	0.25600-+0.0762	5.03
Acetate	0.07190	22.70	1.02000-+1.82	316
Acetoin	0.01200	0.335	0.02520-+0.0261	27.8
Acetone	0.00307	0.327	0.01600-+0.0287	107
Alanine	0.01690	0.531	0.19200-+0.15	31.3
Butyrate	0.08280	0.349	0.18300-+0.0454	4.21
Citrate	0.24200	0.888	0.51000-+0.1	3.66
Ethanol	20.1000	319.0	175.000-+70.2	15.9
Formate	0.00597	0.382	0.02420-+0.0386	64
Fumarate	0.00050	0.074	0.00993-+0.00748	148
Glucose	0.72400	19.40	1.72000-+2.25	26.8
Glycerol	1.32000	14.90	7.47000-+2.64	11.3
Glycylproline	0.43600	1.310	0.94700-+0.183	3.01
Malate	0.12400	0.9220	0.59900-+0.147	7.45
Maltose	4.71000	90.80	39.7000-+27.6	19.3
Methanol	0.45300	0.731	0.64600-+0.0519	1.61
Pyruvate	0.03710	0.884	0.23200-+0.139	23.8
Succinate	0.05980	1.380	0.43200-+0.214	23.1
Valine	0.03440	0.284	0.12600-+0.07	8.26
Acetaldehyde Hydrate (QIB)	0.34900	6.080	1.24000-+0.494	17.4
Acetaldehyde (QIB)	0.00543	0.229	0.04250-+0.0278	42.1
Ethyl acetate	0.01280	0.849	0.11700-+0.136	66.2

Table 4.3: ***Metabolite (and carbon source) Concentrations (mM) analysed through quantitative NMR after growth in Malt media.***
Highest and lowest performing strain values shown alongside standard deviation in expression across all strains in the study. The relatively high variability between low and high concentrations (fold diversity) is often explained by a very low minimum in a strain. Acetaldehyde metabolites' spectra added to compound library from Quadram Institute Bioscience (QIB). Numbers rounded to a s. f. number to align values.

From table 4.3, it is instantly visible (from Fold Diversity/Change), that there exists a huge dynamism within the *S. cerevisiae* strains. While all of the same species, they are nonetheless specialised to specific roles. Some strains are used in ale production, while others make wine and still others are from clinical environments. As such, their broad diversity in metabolomic outputs is not unusual. While generally more flavour compounds were produced (citrate, acetoin) in the malt media (table 4.3) than in YNB media (table 4.2), this discrepancy could probably be partially explained by more of the carbon source.

Nonetheless, our analysis reveal the great metabolic diversity of yeasts, which are a great resource to any potential buyer. Once coupled with the WGS of strains

within the NCYC dataset, it might become possible to both predict phenotypic traits based on genotypic variations as well as the evolutionary relationships between strains (Q-Matrix, section 2.2.4).

4.5.3 Specific Metabolites Within The Malt Yeast Strains

The following subsections will attempt to illustrate the great variability in genomic predictions based on the varied strains and media utilised using select metabolites as examples. Malt media possesses more stored carbon energy (maltose, a glucose disaccharide) than the YNB media, so is expected to have more metabolites produced within it. As only *S. cerevisiae* strains were used in the malt media, which are generally good for the brewing industry, many more 'flavour' compounds are expected to emerge within it. More over, the founder effect predictions made would affect the final correlations- between the Kernel PCA-based PSIKO and the genetic distance SANE, there exist some differences in p-value correlations for specific SNPs.

Far from a weakness, this shows how there are many tools available to a bioinformatician. When one analysis fails to explain the data, a second method might reveal a novel insight. Only when many tools and methods are tested can the data be fully analysed and exploited for its full research potential.

In this analyses, we had many variables. There existed both YNB and Malt media analyses, each with 23 metabolites analysed quantitatively. We also employed 2 separate Q-Matrices; the statistical PSIKO and the TamD genetic distance SANE. To avoid displaying 92 SNP tables, these were added to the GitHub directory, while some selected metabolite and SNP examples will be explained below. They are summarised in table 4.4- this is not an exhaustive list.

As the strains in the YNB media were very diverse, they did not map well onto the reference *Saccharomyces cerevisiae* pan-genome assembled [22]. For this reason, the uncorrected p-values were very high and were not included in this section. However, their SNP data is still included alongside their metabolomic outputs.

The data in table 4.4 is obtained from a GWAS that correlates data from the SNP genomes of each strain to the concentrations of specific metabolites. An example of the data is in table 4.5; full data in GitHub repository ("Strain_Metabolite_Comparisons.xlsx"). Excepting some strains expressing more glycerol in YNB media than in Malt media, all metabolites were upregulated in

Metabolite	Gene	Base pair	p-value	Q-Matrix
Glucose	YGR287C	687	1.25×10^{-5}	SANE
Ethanol	YGR292W	913	6.84×10^{-8}	PSIKO
Ethanol	YCR048W	336	9.59×10^{-9}	SANE
Glycerol	YGR192C	865	4.6×10^{-22}	SANE
Glycerol	YJL052W	561	9.27×10^{-22}	PSIKO
Succinate	YAL054C	1332	1.3×10^{-10}	SANE
Citrate	YNR073C	327	1.30×10^{-20}	SANE
Acetoin	YAL060W	312	0.023	SANE

Table 4.4: ***Specific hits for metabolite concentrations***
The table displays some SNPs highlighted in this Malt media section. They are not an exhaustive list, as many other SNPs exist with similar uncorrected p-values; these are a few selected from the 92 SNP tables with 51,744 (Core Genome) SNPs per analysis each: 23 metabolites, 2 Q-Matrices (SANE/PSIKO), 2 media types (YNB/Malt).

Malt media in comparison to YNB media. This can be due to Malt media having more glucose available (as a disaccharide) than the YNB media.

Strain Malt	Acetoin	Citrate	Ethanol	Strain YNB	Acetoin	Citrate	Ethanol
232	0.0267	0.4408	220.4	232	0.0045	0.0075	37.875
235	0.0269	0.4630	211.5	235	0.0044	0.0075	38.167
360	0.0218	0.4145	197.9	360	0.005	0.0127	31.907
361	0.0273	0.5166	245.8	361	0.0076	0.0126	31.931
505	0.0216	0.4442	206.5	505	0.0107	0.0078	38.533
667	0.0254	0.4704	224.7	667	0.0039	0.0043	38.030
695	0.0245	0.4515	213.6	695	0.0094	0.0160	30.461

Table 4.5: ***Metabolic outputs of specific Saccharomyces cerevisiae yeast in YNB and Malt media***
Table illustrates the change in metabolite production for specific Saccharomyces cerevisiae yeast strains present in both media. Strain number for Malt media under 'Strain Malt', strain number for YNB media under 'Strain YNB'. Acetoin, citrate and ethanol shown as example metabolites. Malt media presents much higher concentrations (mmol) of all metabolites chosen.

Ethanol

Ethanol, known colloquially simply as 'alcohol', has been known to humans for thousands of years as a constituent of recreational, pathogen-free alcoholic beverages [138]. An electron donor to NADH in anaerobic and Crabtree [37] (i.e, high sugar) conditions, it allows cells to respire without oxygen. Yet, it is not essential and performs a similar function to glycerol (figure 1.3.2).

In contrast to the past, recently the usage of ethanol has shifted drastically. While brewers and health experts have begun to focus on low-ethanol beers [155], industrialists have attempted to increase ethanol production efficiency for use

as a biofuel [134]. Much research then focusses on producing bioethanol from lignocellulosic waste biomass [111, 134, 135, 156, 157]. Both of these avenues demand the same thing; elucidation of the genetic elements responsible for the biological pathways that affect ethanol production. While industrialists simply desire high ethanol production and viable yeast, brewers would also want similar flavour profiles to their current brewing strains.

As much of the carbon evidently goes towards ethanol (figure 4.5.2), we can conclude that anything affecting glucose metabolism would affect ethanol levels. Therefore, we can note that glucose concentrations appears heavily reliant on a alpha-glucosidases (such as gene YGR287C at bp 687 (p-value 1.25E-05, SANE Q-Matrix) [158]. This is an expected result that further validates the GWAS model. It is also an indication of a glucose-metabolising gene that increases ethanol (and all other) metabolite values.

Similarly, a mutation in metabolising glucose's disaccharide (maltose) can be predicted to affect ethanol levels. Indeed, this is what was found with YGR292W at bp 913 (SANE: bp 913 6.84 x 10^{-6}, PSIKO: 2.87 x 10^{-8}) [159]. When considering ethanol, and other metabolites, it becomes important to consider carbon source catabolism.

For ethanol-specific genes in *S. cerevisiae* strains growing in Malt media, we found many genes. Each SNP is a potential insight into the network effects of various genes within the regulatory pathways of ethanol. For example, YCR048W appears with an SNP at bp 336 (SANE: p-value 9.59 x 10^{-9}, PSIKO: 3.37 x 10^{-9}) and is the Acyl-coenzyme A gene.The interactions between ethanol and Acyl CoA have long been known so it is not surprising that a SNP within the gene might correlate with various ethanol concentrations[160].

Glycerol

This metabolite is synthesised largely within industrial processes or derived from fossil fuels, such as a by-product of biodiesel production [152]. Used in everything from shampoo to food products to vibration dampeners, it is widely desired by the world economy. While disposal of excess can be an issue [152], a renewable source would be essential to decarbonise and find alternatives to fossil fuels before they disappear.

As glycerol interacts, and competes with, molecules in the TCA cycle it is

expected to heavily vary depending on specific genetic variation that emphasises different metabolic pathways. It also acts as a serious carbon sink (figures 4.5.2, 4.5.1).

Among the many SNP correlations for glycerol concentration, one was found at bp 865 (SANE: p-value 4.6 x 10^{-22}, PSIKO 4.60 x 10^{-22}) for gene YGR192C. YGR192C is a Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) that is differentially expressed in stress conditions [161]. Prior to full validation, it is not inconceivable to consider this a 'true' hit. As dehydrogenation of the alcohol version (glycerol) gives glyceraldehyde, it is possible to see that the dehydrogenation of glyceraldehyde would affect glycerol levels.

A further validation of GAPDHs in the role of yeast glycerol concentrations is illyustrated by gene YJL052W at bp 561 (PSIKO: 9.27 x 10^{-22}, SANE: 9.27 x 10^{-22}) [128]. The identification of two disparate GAPDH genes with very low uncorrected p-values indicates they might be ideal targets for the alteration of glycerol concentrations.

Succinate

Succinate is a TCA metabolite produced in the mitochondria of cells, but is found throughout cells and even extra-cellularly. Due to it being used in the TCA cycle, it is found at some level in almost all eukaryotic organisms. This makes it an excellent target for GWAS, where we can then potentially identify genetic variants that affect metabolite levels [26, 162, 163]. In vivo cells, it can act as everything from DNA transcription modulator to enzyme inhibitor though interaction with histones and α-ketoglutarate-dependent enzymes and propyl hydroxylases [163]. In industry, succinate fulfils many other roles. It is used as a food additive, feedstocks for pharmaceuticals and agricultural animals as well as the general bulk chemical market which includes using succinate to produce nylon, various plastics and select chelators [137]. In short, it is needed in high concentrations for ease of purification and high volume for best pricing [137].

Succinate is a chemical of high industrial and, thus, research interest. While sourcing of succinate was derived through oil, renewable bio-based methods are being investigated as replacements in a greener system [137, 136]. We have therefore selected succinate as a biomolecule of interest due to ubiquity of both industrial interest and yeast metabolic production.

Fortunately, we see in the Malt media there are many SNPs with low uncorrected p-values correlated with succinate concentration. For example, the SNP with the lowest p-value is YAL054C on bp 1332 (p-value 1.3 x 10^{-10}, SANE Q-Matrix), which is a Acetyl-coA synthetase gene- shuttling more carbon into the TCA cycle (figure 4.5.4) [164]. Another gene is YAL061W at bp 549 (p-value 3.4 x 10^{-9}, SANE Q-Matrix), which is a putative medium-chain alcohol dehydrogenase with similarity to BDH1 and very tentative relations to succinate respiration [165].

Citrate

Another flavour compound important to the brewing industry is citrate. The compound causing the 'citrus fruit' flavour, it is produced in the biochemical pathway that bears it's name; The Citric Acid (TCA) cycle. An important flavour compound, its presence was expected in Malt media, and plenty was found (figure 4.5.2).

The top correlated SNP among many, the gene YNR073C has a mutation at bp 327 (p-value 1.30 x 10^{-20}, SANE Q-Matrix). The gene is MANnitol dehydrogenase, an oxidoreductase that affects cellular levels of NAD+/NADH. This is important as NAD+/NADH are pivotal to citrate levels. In the TCA cycle, both of citrate's substrates (oxaloacetate and acetyl-CoA) require NAD+ to be produced themselves. However, it is important to note that the gene YNR073C increases NADH levels; when NAD+ is needed for both oxaloacetate and acetyl-CoA. Therefore, the allele in question might reduce its ability to catalyse the conversion of NAD+ to NADH [166].

This gives an interesting target for the increase of citrate; both YNR073C, but also oxidoreductases and NAD+/NADH genes in general. Affecting the NAD+/NADH ratio and concentrations can be predicted to have large effects on citrate levels. This highlights the complex network effects of genes and their interactions; a gene may affect a completely separate gene due to modifying core metabolite levels.

Acetoin

As a metabolite with a pleasant 'buttery' smell, acetoin is often added to butter-alternatives (such as plant-based butters) to add the distinctive 'buttery' taste. Due to this gentle flavour, Acetoin is a useful metabolite in beer. As such,

elucidating the biochemical pathways leading to its production is of industrial interest.

Acetoin is produced by many organisms from *Lactococcus lactis* [167]to various bacteria and *Saccharomyces cerevisiae*, and found in many foods from yogurt to blackberries. Already in use as an additive in everything from butter [167] to electronic cigarette 'vape' products [168], it is widely accepted as safe for use.

The *S. cerevisiae* strains had many correlations to their acetoin levels; there were 3354 SNPs with p-values below 0.01- with gene YAL060W that catabolises acetoin [169] found with an uncorrected p-value of 0.023 at bp 312 (SANE Q-Matrix). With a P-value too high be valid, it is nonetheless interesting. The huge number of SNPs with very low p-values (228 with less than 0.0001) highlight malt extract as a media well-suited for the production of flavour compounds from *S. cerevisiae* strains.

4.6 Discussion

To analyse the results of the experiment and to assess the potential consequences of these results, it is necessary to take a deep dive into the literature. Firstly, it is necessary to select a SNP, with associated gene information, that is correlated to the desired phenotype. Secondly, the gene in which the SNP is located in is investigated and its general Gene Ontology (GO) and other functions verified.

Such a strategy might already reveal a possible mode of action, if the functional data is evident enough and the correlation is strong. Subsequently, if the gene is appropriately annotated, the specific locus of the SNP can be identified. This allows for a much more accurate prediction of the potential method of a resultant phenotype. For example, a mutation in the active site of an enzyme that creates the metabolite can likely be linked to the enzyme being mutated to increase its efficiency and/or how active it is.

The experiments carried out here, while providing possible solutions for the genetic basis of phenotypic traits, remain predictive in scope. Further studies, for example by targetted genetic mutation, would test the hypothesis of the SNPs' role in the phenotypic outcome. However, there is also another possibility; a directed evolution experiment to increase a metabolite, followed by a WGS to verify if the expected mutations gained prominence could be contemplated, for example in the high-value metabolite succinate.

To increase succinate production in yeast strains, we would need to produce an environment where an increase in expression would result in an increased fitness. In this way, we could select for genetic traits that cause higher succinate. One possible avenue is to grow the strains in a competitive inhibitor for succinate; malanoate. This would mean that for the cells to utilise succinate in the levels needed in the TCA cycle, mutations might have to evolve to allow for better succinate specificity, elevated expression or, unfortunately, a method for cells to remove melanoate from the cells.

In this study, we found various genomic variants potentially underlying metabolite expression changes. For example, we found an enzyme known to directly interact with acetoin (YAL060W) as a potential gene affecting acetoin levels. Other interesting results were uncharacterised regions of chromosome 15 potentially being pivotal to ethanol production (in YNB media data). In a GWAS system, there could be many reasons for such indirect genes causing increased expression of specific metabolites, such as a reduction in competition for a scarce carbon source. In finding known, fully characterised genes where expected (acetoin YAL060W) we verify the accuracy of the GWAS performed. In finding tangential or unexpected hits, we attempt to provide novel insights into the network effects affecting a metabolite's expression level. Overall, the GWAS was roughly successful, even with the relatively small number of strains included in each study.

With more data, and better reference genomes for diverse yeast strains, better GWAS may be carried out. As it stands, we were able to obtain good metabolomic profiles for each strain and evaluate some of the predictive ability of the study with some expected SNP hits. In future, an expansion of the assayed strains would expand the potential of the GWAS.

4.6.1 CHENOMX profiling issues

The following examples will present some of the difficulties encountered in assigning concentrations to a metabolite's spectra. In general, difficulties in assigning peaks to a metabolite are due to peak-shifts perhaps caused by an unexpected pH, to trace amounts of metabolite that are difficult to identify or, to peaks in a 'messy' area of the spectra. All these factors can make it difficult to determine the quantity of a specific metabolite. For these reasons, malate, citrate

and isocitrate were rejected for analysis as their concentrations could not be reliably identified.

For the metabolites carried forward for further analysis, some additional issues remained. The first is the issue of two metabolites with single peaks very close to one another. These two peaks may be confused for each other when quantified via automatic annotation. A prime example of this phenomenon is acetoin, a 'buttery' flavour compound common in beer and likely an important factor in a beer's flavour profile. Acetoin's NMR spectra peak is adjacent to that of acetone. Acetone, while produced in low amounts as a secondary metabolite, is usually a signal for contamination from the acetone washing of NMR tubes. It can be difficult for CHENOMX (CHENOMX Inc, Canada) to automatically fit both peaks correctly and they must thus be corrected manually.

A separate issue is one of consistent peak shift. Due to deviations in media pH or other media components, a metabolite peak is consistently found at an unusual second location in the spectra. This is the case with Succinate, another metabolite of interest. An acidity regulator and an input for plastic production, it is an important metabolite for industry [162]. Through close analysis, and personal consultation with NMR lab manager Dr. Colin MacDonald, it was easy to identify a recurring mis-annotation of the CHENOMX software and adjust for it.

A final issue is when a compound library does not contain the required metabolite at the desired NMR frequency. For example, to identify acetaldehyde, it was necessary to use the 600MHz reference values on a 500MHz spectra. Due to mismatches, only manual annotation was possible with the spectra peaks. Additionally, due to limited intersection with glucose/maltose, it was a useful cross-check for glucose/maltose concentrations.

However, none of these issues were found to affect high-concentration compounds or those whose peaks were distinct from their neighbours. Ethanol is a perfect example of a high-concentration compound present in all strains. It is thus correctly and confidently annotated automatically by the CHENOMX software throughout the study.

Chapter 5

Operon Prediction of Cytochromes through genomic analysis

5.1 Electrogenic bacteria

Electrogenic bacteria are bacteria that can produce electrical potentials to shuttle electrons across their membranes. *Shewanella oneidensis* and *Geobacter sulfurreducens* are both isolated from the anaerobic sediments of lake water and are the best understood electrogenic model organisms [56, 57, 170]. The anoxic subsurface is not the only place that electrogenic bacteria are found, for instance the human microbiome has also been shown to electrogenic bacteria such as *S. aureus* (ATC 6538), *E. faecalis* (ATCC 19433), *S. agalactiae* (A909), *L. rhamnosus* (GG), and *L.reuteri* (ATCC 23272)[171]. Some showed even greater electrical productivity than *S. oneidensis* [171]. This illustrates that a diverse array of hitherto-uncharacterised bacteria have the potential to be electrogenic- including some in our own gut microbiome.

Electrogenic bacteria have a wide array of potential biotechnological uses. The three most common and of interest to industry are bio-remediation, microbial fuel cells and microbial electro-synthesis [172].

In bio-remediation, electrogenic bacteria can be used to remove uranium from contaminated groundwater in defunct ore-processing facilities or waste water treatment [172, 173]. This is vastly cheaper than gathering and processing all the soil, and water, separately.

Microbial fuel cells are another focus of research, for either wastewater

remediation, electricity production from waste or soil or a mix of both [57]. The broad range of substrates that electrogenic bacteria are able to thrive on has even been exploited to act as a sensor for specific substrates, where a substrate's presence is measured by the electrical current generated by a mutated bacterial population [174].

Lastly, electrogenic bacteria have been investigated for their ability to undergo electro-synthesis. This is when a lower-value substrate is turned, with electrical energy being supplied as an energy source, into a higher-value product. For example, the final goal of directly turning carbon dioxide into higher value multicarbon extracellular organic compounds with electricity [43]. When coupled with 'green' electrical supplies, it presents a potential alternative to oil-derived chemicals.

5.1.1 Biochemistry of Electrogenic Bacteria

Cyrtochromes are proteins that use redox-active iron atoms in haem co-factors to act as electron transporters or catalyse redox reactions. When multiple haems are arranged in a linear conformation within a protein and across a membrane, they can be used to shuttle electrons between a cell and its environment. This allows for the reduction of insoluble mineral oxide grains of Fe (III) and Mn (III/IV) in a microorganism's environment. This ability can sometimes evolve into essential respiratory functions for metal-respiring (electrogenic) microorganisms[175, 176].

An electrogenic bacteria that can use this ability to 'breathe' metal (in place of oxygen) is *Shewanella amazonensis*. This bacteria is capable, in anoxic conditions, to utilise the metal oxides and solids in its environment as terminal electron acceptors [176].

The complex that allows electron transfer across the outer membrane of S. oneidensis is composed of three subunits, which in *Shewanella* are named MtrA, MtrB and MtrC. MtrC acts as the extracellular protein interacting with the extracellular metals in the environment, MtrB acts as an insulating transmembrane porin β-barrel with MtrA nestled within MtrB. Subunit MtrC interacts with MtrA to pass electrons back and forth from the cytoplasm, through MtrB, to the extracellular environment [56].

The cytochome structures predicted are organised into 4 general structural

groups. The MtrCAB structure is the largest and consists of a porin with two haem-containing cytochromes (figure 5.1.1), while the MtrAB (figure 5.1.2)is similar but lacking MtrC which means it cannot interact with its environment as easily. The Cyc2 (figure 5.1.3)'fusion' type is similar to MtrAB but is composed of a single protein that is both a porin and an iron-containing electron-transport haem chain (MtrA and MtrB joined into one protein). Finally, we group any cytochromes that do not fit into these 3 groups into 'other'. These may turn out to be 'weird and wonderful' structures which improve our understanding of cytochromes as a whole.

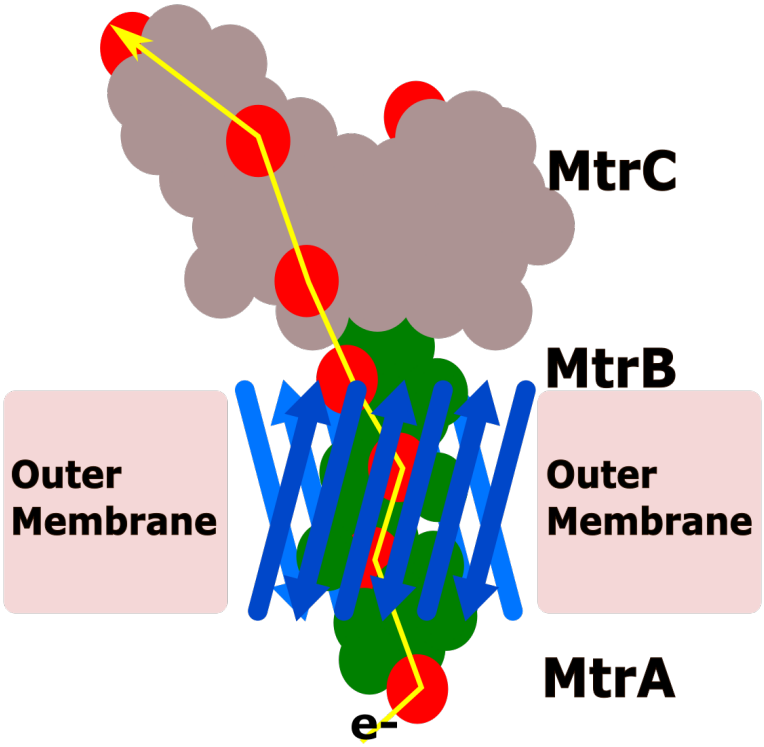


Figure 5.1.1: ***MtrCAB Operon Illustration***
MtrCAB Operon is visualised in an artistic representation of the general protein structures. MtrA (Green) is sheathed within the MtrB (Blue) transmembrane porin and connects to the MtrC (Grey) extracellular cytochrome. Electrons are carried from within the cell towards MtrC (then to the environment) along the electron path (Yellow line) which follows the chain of C-type haems (Red circles, number solely for illustrative purposes).

Knowing these general structures, and each subunit's makeup, it becomes possible to predict if a genetic sequence could be one of the three subunits. The genetic sequence gives us plenty of clues, from a predicted number of haems in a protein the to number of β-strands and its cellular localisation signals. This can be coupled with knowledge of analogus systems, that might comprise of a one- or two- component system as opposed to the usual three-component one. All together, it becomes possible to screen genetic sequences for likelihood of being a

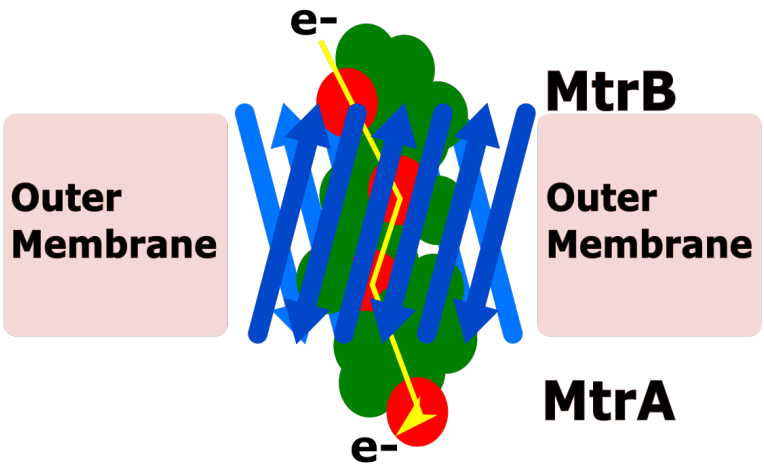


Figure 5.1.2: ***MtrCAB Operon Illustration***
MtrAB Operon is visualised in an artistic representation of the general protein structures. MtrA (Green) is sheathed within the MtrB (Blue) transmembrane porin. Electrons (Yellow line) are passed along MtrA's c-type haems (Red circles, number solely for illustrative purposes) and through MtrB.

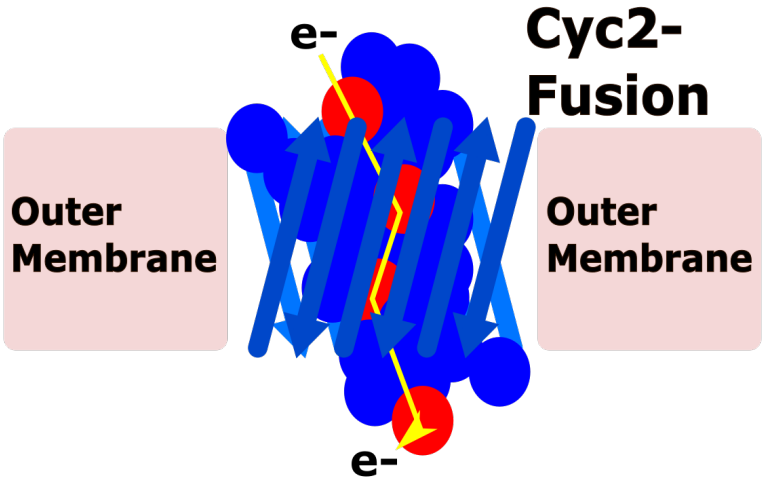


Figure 5.1.3: ***MtrCAB Operon Illustration***
Cyc2 Operon is visualised in an artistic representation of the general protein structures. The blue Cyc2 fusion protein acts as both porin (directional 'barrel' arrows) and as electron transport chain with c-type haems (Red, number solely for illustrative purposes). The electron path (Yellow) leads through the electron transport chain's path. The electrons are drawn into the cell.

cytochrome-porin operon.

In our computational approach, we attempt to predict these cytochrome-porin operons across all sequenced bacteria in the RefSeq database. This would allow for an analysis of all sequenced bacteria, to allow us to narrow down the database to suspected cytochrome-containing bacteria. Once all potential cytochrome operons were identified in the bacterial genomes, they could be refined to only those of specific properties (such as a specific number of haems). Further, each cytochrome

was classified as one of the various distinct cytochrome types (figures 5.1.3, 5.1.1, 5.1.2) that allow us to better predict their phenotypic capabilities.

This novel approach allows us to classify all cytochromes present in all genomes within the RefSeq database.

5.2 ETMiner

Electron Transfer data Miner (ETMiner) is a Python software tool developed within this project to predict putative cytochrome operons from bacterial CDS genome files. Able to predict operons from CDS genome files, it allows a user to interrogate a database with desired operon parameters.

Able to automatically generate various figures such as heatmaps and scatterplots, ETMiner enables quick analysis of the data to allow the researcher an intuitive understanding of the data structure.

5.2.1 Metagenomic approach

ETMiner was used to predict all cytochrome operons present in the entire sequenced CDS bacterial dataset within the Refseq database [64]. As such, it captured most bacterial species found to date. (2020, when RefSeq CDS genomes were downloaded). The single-database approach has the added benefit of ensuring all genomes analysed adhere to the same quality standards instead of attempting to reach a parity between multiple databanks with diverse data submission standards. In this way, we can ensure all data were submitted under similar principles and hold the same base biases.

This breadth of data utilised makes the discovery of novel operons highly likely. Once operons are discovered, they can be backwards placed into their genomic context.

5.2.2 Bioinformatics pipeline

To identify all cytochrome operons within the entire bacterial genome dataset, it was necessary to create a bespoke pipeline and software application. Electron Transport data Miner (ETMiner) allows for the identification of various cytochrome operon types that match specific characteristics such as c-type haem count and molecular

weight ratios.

To construct this application, it was necessary to create a BLAST nucleotide database that was searchable by a BLASTn (i.e. nucleotide to nucleotide) call. The database created was also, through headers within the data files, matched to genomic location data for future operon construction. This division of information, including splitting the database into four segments, was due to computational limits and the size and breadth of the data used.

Cytochrome-like (containing c-type haems) proteins with known periplasmic signal peptides were BLASTed to this sub-divided database. This would hopefully identify all cytochromes that are located on cell walls- even in completely uncharacterised genomes with purely predicted CDS regions. Once a protein match was found, with an e-value below 0.0001 and query cover over 85% (percentage match between the query and hit), it was then placed within its predicted operon.

A gene is assumed to be within a single operon with a neighbouring gene if the two genes are less than 100bp apart. The chain of genes all within this 100bp of their next neighbour are assumed to be an operon. This attempted to extract the putative cytochrome entirely- with any associated proteins. One of these associated proteins could be the porin required for cross-membrane electron shuttling.

To verify if associated proteins (or the hit protein itself) were porins, the transmembrane beta-strands of each protein had to be counted. The TM region count per protein was predicted using PRED-TMBB [16] while the haem count was predicted through the Scan Prosite [17] tool. Once this had been done, each predicted operon was assessed for its cytochrome potential using the ETMiner application. This new software speeds identification of potential cytochromes of interest in the broad range of bacteria via the selection of matching criteria. The identified operons can then be narrowed down further manually for experimental validation. This is useful, as it gives new insight into regions of interest and where to begin searching. As it carries out a broad search of all bacterial species, ETMiner is also able to identify regions in bacteria that have very limited gene annotation but nonetheless match cytochrome profiles.

Unlike in standard GWAS systems, we translated coding DNA sequences (CDS) into proteins after our initial BLASTn searches. Next, we focussed on matching protein sequences to known motifs and, from there, constructing predicted operons. In this way, we obtained a roughly genome-wide prediction of phenotypic

characteristics based on genetic sequences.

The entirety of the pipeline meant that it was possible to search an entire CDS genomic database for potential c-type haem cytochrome porin-containing metal-respiring operons without experimentally validating genes. The BLAST commands would identify likely candidates, while the Haem and TM motif searches would then verify cytochrome likelihood. Once this was completed, candidates could be reviewed with manual experimentation. All this data was later joined in ETMiner (figure 5.7.1).

5.3 Bacterial genome selection and curation

As a project designed to capture the breadth of cytochrome operons in bacterial genomes that have potentially remained as-yet undiscovered, it was decided to use every genome in the Refseq database. The Refseq database is a high-quality database of sequenced genomes that would give us reliable results of real operons [64].

To refine any missing gene/protein data, the non-redundant (NR) protein database at the NCBI was used to fill in the gaps. This helped ensure that the manual data curation remained of high quality. However, there were significant issues with assigning taxonomic identification (TaxID). Many CDS and proteins in the study had no TaxID information, or it was outdated, or had changed as genomes were re-classified.

These issues were largely alleviated through manual curation. For example, using NCBI's ftp system allowed many conversion files to be downloaded for quick conversion to the accessions needed (https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/). Using these files, TaxIDs can accurately be attributed to specific WGS genomes by refseq. Secondly, any strains with missing species info can be assembled from NCBIWWW's BioBlast python3 packages that convert TaxIDs to species information.

With these steps, and converting old TaxIDs into new ones, it was possible to eventually find as many TaxIDs as possible from all WGS and CDS information. Some rare discontinued TaxIDs and anonymous protein IDs were completely inscrutable.

5.4 Bioinformatics and explanation for choices and development of methods

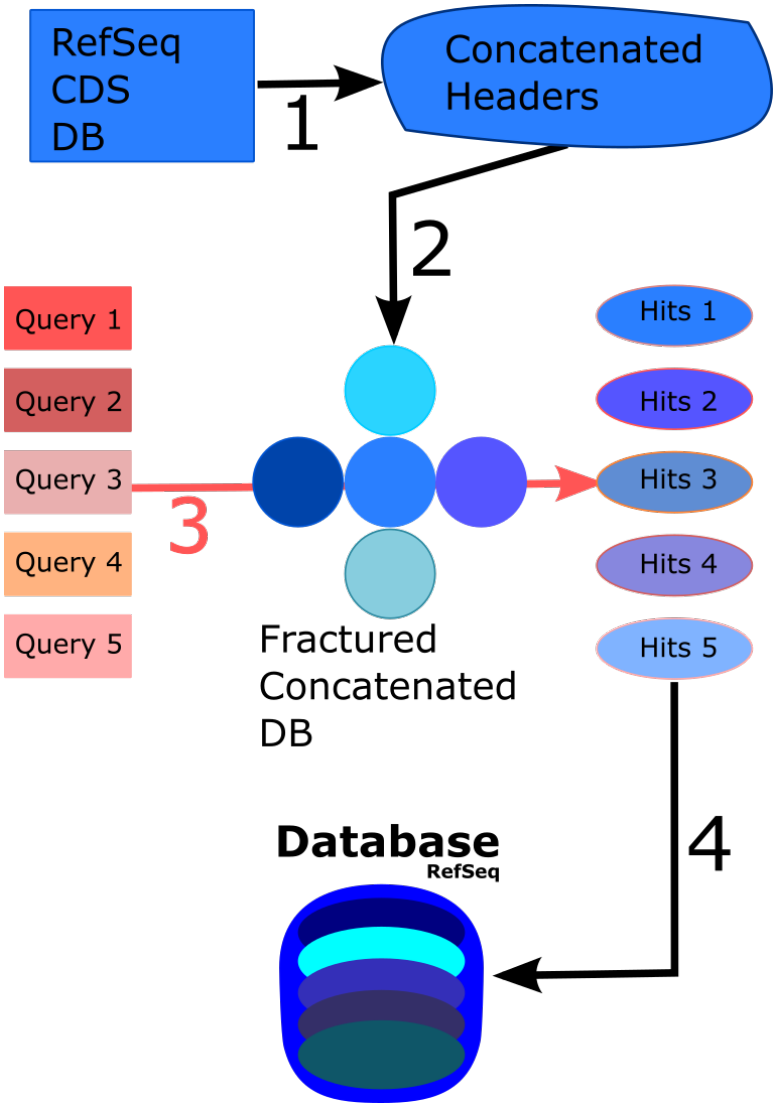


Figure 5.4.1: ***Creation of DB for ETMiner app***
The RefSeq CDS genomes were downloaded and placed into a single file (RefSeq CDS DB). The data was too large to host on the server, so it was reduced (700+Gb to 300Gb) by concatonating headers (Concatenated Headers) in ***Step 1****. In* ***Step 2****, this was fractured into multiple smaller databases to be small enough to BLAST on available HPC cores. In* ***Step 3****, BLASTn was used for queries against the fractured databases, and hits stored. In* ***Step 4****, the hits were reinserted into their putative genomic operons, converted to protein sequences and then joined together into a single database (Database RefSeq).*

To explain the later stages of the analysis, it is first necessary to explain how the databases was created. Two databases were made; one with sequence information, and one without.

The RefSeq CDS Genomes were downloaded from the NCBI FTP website directly. As there were too many genomes at the time (2020), the database was

over 700Gb and reached the limits of our storage permissions. As such, in **step 1** a database was created and reduced by exploiting redundancy. Every redundant DNA sequence received a header of all the matching genes in all the genomes concatenated. This database was, in turn, too large for BLASTn to query with the CPUs at our disposal. This had the benefit of allowing us to simultaneously BLASTn many of our queries at once with many cores- important due to our large volume of query sequences (250,000).

Therefore, in **step 2**, the concatenated-headers database was separated randomly ('fractured') into multiple smaller databases. These were then queried in place of the larger database for **step 3**. The e-values of true hits were nonetheless still distinct when compared to non-hits by many orders of magnitude (decimal points). The hits were plugged back into their original CDS genome files and neighbouring genes extracted. These hits, with their neighbouring operon genes, were then joined into a single database in **step 4** (Database RefSeq). They were also converted to their relevant protein sequences for subsequent Hidden Markov Model (HMM) analysis of specific protein motifs. One version was maintained without any sequence information in the eventuality sequence data was redundant to a researcher.

The database creation was initiated with the relevant blast command (equation 5.1) below.

$$makeblastdb -in\ NonRedundantDB.fna\ -title\ DB\_for\_BLASTing\ -dbtype\ nucl \tag{5.1}$$

Once the databases were created, each query sequence was BLASTn-ed against the smaller databases. As a nucleotide-nucleotide match, we used the blastn command (equation 5.2).

$$blastn -outfmt\ 7\ -query\ query.fa\ -db\ databaseN\ -evalue\ 0.0001\ -qcov\_hsp\_perc\ 85 \tag{5.2}$$

Output format 7 has been chosen as the preferred visual output for data parsing, with $database_N$ (N being databases1-4). The evalue and query cover (percentage of query matching target in database) were selected to be broad enough to capture enough novel variation, while remaining significant.

Hits were individual gene matches that were mapped to their parent genome

(s). This was possible by both preserving all the 'headers' of the hits (step 1, figure 5.4.1) and output 7 (-outfmt 7) which displayed everything necessary. The genes, mapped to their parent genomes, were then inserted into predicted operons which were then converted to protein sequences.

To accomplish the operon creation, a 'window' of 300 genes on either end of the re-inserted gene was maintained to either end of each gene hit. The single-gene operon was then expanded sequentially to include any genes within 100bp at either end of the operon's ends (to the 300 limit). This sliding 'window' was utilised to counter computing resource limitations (step 4, figure 5.4.1). The reduction in computing difficulty also resulted in faster data processing.

Once the operons were thus constructed and assembled through BLAST and custom scripts, all the proteins (converted from the DNA sequence) within the operons were listed. This list was analysed for haem and transmembrane motifs. The end result was a dictionary of Haems and transmembrane motifs per protein distinct from the operons the proteins reside in. The cytochrome operons were then predicted using the haem and transmembrane information acquired earlier.

The ETMiner app joined c-type haem, transmembrane β-strand and operon data with thresholds (haems per protein,..) and overall structure (Cyc2 cytochrome-porin fusion structure, etc) to output predicted operons of each type (Cyc2, MtrCAB, MtrAB, Other). This data included the molecular weight of each protein within the operon. This extra molecular weight data was used to filter, sort and rank operons to highlight those of most interest to a user.

5.5 ETMiner and Cytochrome novel operon prediction

The ETMiner app takes as its input 3 files.

The file containing data on the operons was constructed via the HPC scripts. It is, by far, the largest as it contains gene names, sequences, genomic locations and genome information. It was processed and produced via bespoke software, and refined through manual data curation. For example, some gene names were missing and had to be replaced via BLASTing back to the non-redundant (NR) protein database. This allowed for the extraction of missing NR protein accessions for genes.

The second file was the haem data file. The prediction of haems and, therefore,

cytochrome relied on the CXXCH/CXXXCH motif. This is because the motif allows the incorporation of a c-type heme into the protein when it is transported to the periplasm. These predictions were obtained via running through haem prediction software Prosite [17] which predicted the number of CXXCH/CXXXCH motifs in a protein sequence. This allowed us to detail how many CXXCH/CXXXCH were in each gene and operon in our operon file for the final analysis [175, 176].

It is generally accepted that MtrCAB cytochrome porins have at least 20 haems, MtrAB have more at least 10 and fusion (Cyc2) need just one to function. Therefore, haem number becomes an important factor when considering if a protein is a true cytochrome porin [170].

The final file contained information on putative porin structure (via TM strand count) obtained through another researcher (Konstantinos D. Tsirigos [16]) who made specialised software for the purpose (TransMembrane Beta Barrel predictor, TMBB2 [16]). The TMBB2 tool predicts the number of β-strands within a single protein- which then is used to predict the porin's structure. This data, verified through randomly checking some predictions, appeared accurate.

Once all three files were unified, it was possible to make predictions within certain constraints. For example, it became possible to locate operons for which we predict 10 haems in a single protein, flanked by two CXXCH/CXXXCH proteins. By filtering with different constraints, we could check the individual veracity of predicted operons. While some were the expected, some presented novel insights- including cytochrome-porin electron transporters found in bacterial species previously believed to have none.

In this way, we can bring manually curated data together to form intuitive outputs that are easily translatable into operon data that can be investigated experimentally. While most operons should be readily identifiable, many novel ones could be investigated to learn something of both cytochromes and sequence predictions.

5.6 ETMiner Usage

The app designed for the elucidation of cytochromes was by necessity option-dense. In the case where the user might require rare functionality, many often unused options had to be included. This could present itself as a bewildering profusion

of choice to the new user (figure 5.8.1).

Therefore, full usage and functionality is spelled out below.

From figure 5.8.1, the options present at start-up are in evidence. The first option is simply naming of the output operon file. The format can be seen in figure 5.8.1 with explanation of barcode in 5.8.5. The default would include TM and Haem numbers mixed with the date the run is processed.

The second option is the selection of the operon 'type' to look for. I.e, fusion (Cyc2), 2-component (MtrAB) or 3-component (MtrCAB). There is also the options of both 'other' and 'custom'. Other allows the search for any non-characterised operons, which could be entirely novel in structure, and any added to a custom search (as per 'Custom Operon Search' button in the bottom left).

The TM threshold is simply the number of predicted beta-strands to find in a protein to count it as a TMBB protein. The CXXXCH is the number of haems motifs to identify within a protein to count it as a cytochrome.

The 'top hits to print as image' it allows the user to print a set number (default 10) of the top hits (lowest kDa/haem ratio) into operon CDS images. This could be output as a variety of formats (and size, default 25)- notably SVG which could then be further manipulated by the user in Inkscape or other similar software.

The ratio of kDa to haems has been identified as of potential scientific interest, with specific ratio parameters being vital. As such, it was added as an option.

Custom operons structures can also be searched for (figure 5.8.3). This allows for much more versatility and future-proofs for when other general cytochrome structures are elucidated and need to be searched for in the current database.

To create images from each operon at will, it is necessary to input the row data from the output (a series of accessions from figure 5.8.5 separated by a comma) into the app's image generator (figure 5.8.2). This will output the relevant images (figure 5.8.4) for consideration by the user.

To interpret the output data itself, it is necessary to read the headers for the output CSV and then each predicted protein within the operon which is decoded via the example in figure 5.8.5.

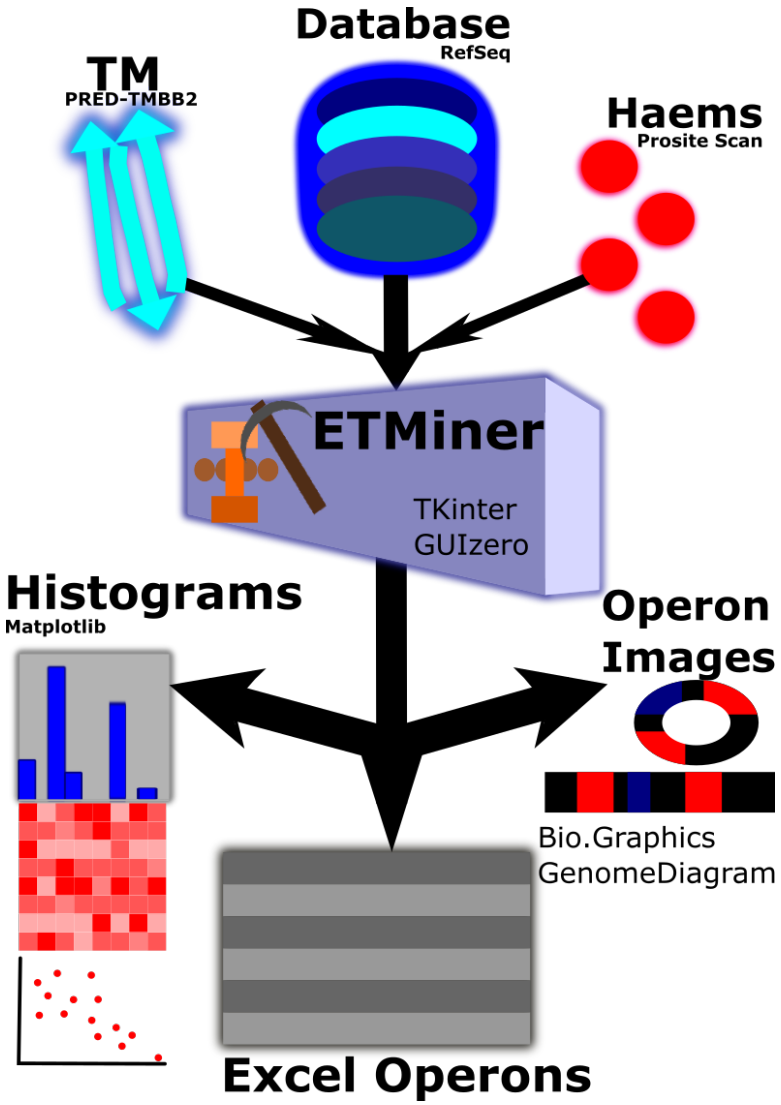


Figure 5.7.1: ***Basic data workflow of ETMiner app.***
ETMiner requires Haem and TM strand datafiles to predict transmembrane porin cytochrome complexes in bacterial operons (here RefSeq DB from figure 5.4.1). ETMiner uses the following Python Packages: Bio, reportlab, guizero, EasyTKinter, datetime, numpy, math, pandas, glob2, pillow, openpyxl, matplotlibmath

5.7 ETMiner Backend data handling

These predicted raw (Database RefSeq, figure 5.7.1) operons are assessed for adhering to operon 'types' based on TM and Haem numbers within each protein of the operon. The data for each is saved as in individual files.

As seen in figure 5.7.1, the basic data workflow is of conjoining data together from disparate sources and predicting results with visually pleasing outputs.

The operon database is the most important element in the setup. Without this basic database, no local analysis can be performed. The operon database, 'Database Refseq' (figure 5.7.1), is created in a linux server as per figure 5.4.1 and contains information on our predicted, raw, operons.

The Haem files (CXXCH, CXXXCH) contain information on the predicted number of haems for every protein within the database, while the TM file contains the predicted number of beta strands for every protein within the database. These were obtained through methods detailed in section 5.5.

The haems and TMs counts were only counted as one of the known 'types' (Cyc2, MtrAB, MtrCAB) if they existed in the known conformations. For example, to be a Cyc2 protein, the Haem count and TM count must both be above the desired threshold and on the same protein within the Operon. If the haems and TMs are not on the same protein, the operon cannot be a Cyc2 fusion cytochrome and the Haem/TM count is thus not counted on any Cyc2 data plots.

Joining the Haem/TM counts with the raw operon database itself, it can predict operons into a CSV format with images of select operons as well as create histograms and heatmaps for a visualisation of the data.

The ETMiner app used a variety of packages, with the main ones being TKinter and GUIzero for GUI construction, matplotlib for all figures and charts, and finally GenomeDiagram from Bio.Graphics for all operon visualisations. It works on a local computer as a desktop executable app.

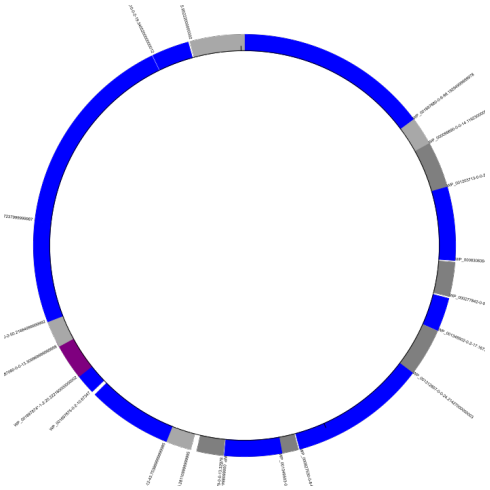
5.7.1 Operon Figures

To visualise operons directly, ETMiner permits the printing of entire operons in both linear and circular formats. For example figure 5.7.2 shows a operon classified as 'Other' in type but passing both TM strand and haem thresholds.

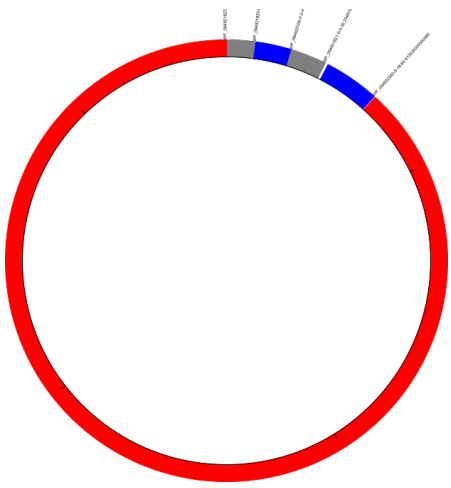
While it is possible to create some of the 'top' operons as a test of the select top operons, it is also possible to directly print an operon. This allows any user to quickly print out any operon to visualise. With a SVG output mode, it is possible to undergo subsequent manual tinkering. This permits near-infinite modification from a base template.

This simple tool allows for the instant comprehension of an operon to confirm its structure visually. As they include to-scale genetic distances of the CDS regions, it is a useful visual tool. In the figures, blue represents TM- containing proteins, red is haem-containing proteins while purple are fusion proteins, with both haem and TMs.

Not necessarily a final publication-grade figure, nor for use in plasmid design,



(a) An example of a complex operon classified as 'Other'



(b) An example operon classified as 'MtrAB'.

Figure 5.7.2: ***Operon figures classified according to the ETMiner rules***
The figures illustrate circularised operons (NOT PLASMIDS) crafted from RefSeq protein CDS location information.Blue represents TM-containing proteins, red is haem=containing proteins while purple sections are proteins with fusions of the two. Grey sections are intergenic regions or proteins without either TM strands or haems. The raw file is an SVG and nearly infinitely scalable for HQ images.

it is a useful starting point for a researcher wanting a quick visual overview of an operon. From this point, it is possible to then further design the next step- having understood the structure from the automated figure generation. The option of linear and circular simply allows the user to choose that which intellectually resonates with them best.

All the following figures were created programmatically and were not manually manipulated to present an accurate idea of the outputs of ETMiner. Any slight inaccuracies can, however, be manipulated and tweaked with the excel (XLSX) files provided with each figure.

5.7.2 Scatterplots

One type of figure created automatically by ETMiner is a suite of scatterplots, one for each operon type (Cyc2, MtrAB, MtrCAB, custom and Other). The scatterplots aim to identify proteins /operons of particular interest by plotting the $\log_{10}$ of the molecular weight (kDA) to haem number, then plotted against haem number on the x-axis.

That is to say, we find the average amount of amino acids (in kDA) per haem in the protein. This tells us if the haems are unusually sparse in the protein or if they make up a relatively large percentage of the total weight of the protein (low

kDA/haem). This is illustrated in equation 5.3

$$\begin{aligned} X-Axis &= Haem\ Number \\ Y-Axis &= log_{10}(kDa\ of\ protein/haem\ number) \end{aligned} \tag{5.3}$$

The $\log_{10}$function aims to reduce the spread of data to a more visually pleasing range. In this way, it is possible to view the entire spread of data in a single plot comfortably. In figure 5.7.3, this becomes clear as we can see everything from 1000kDa (3, y-axis) to nearly 0 kDa/haem.

The output in figure 5.7.3 is taken directly from ETMiner's output for transparency. However, the raw data making up the figure is automatically also presented in an XLSX file.

This means that interesting data points, such as the protein with 62 haems on the scatterplot, can be identified clearly. Because of the thresholds set in the ETMiner app (figure 5.8.1), we know each of the proteins found here have at least 12 TM strands. This number was selected as the conservative lower threshold of TM strands needed for the protein to plausibly possess a porin in it's 3D structure.

In this case, the species responsible is *Geopsychrobacter electrodiphilus* (TAXID:1121918). The bacteria is a known anaerobic, psychrophilic metal-respiring bacterium [177]. This makes it a good candidate for research into metal-respiring cytochrome-containing organisms that do not require high temperatures to function.

The protein with 62 haems within the species is WP_020674948. With a very low kDa weight per haem (4.4) it is a very interesting hit as it shows a fusion protein with a huge number of haems that represent a large fraction of the overall protein's weight. Due to it being largely uncharacterised, it is an excellent potential research target obtained from the overview granted by the automatically generated scatterplots.

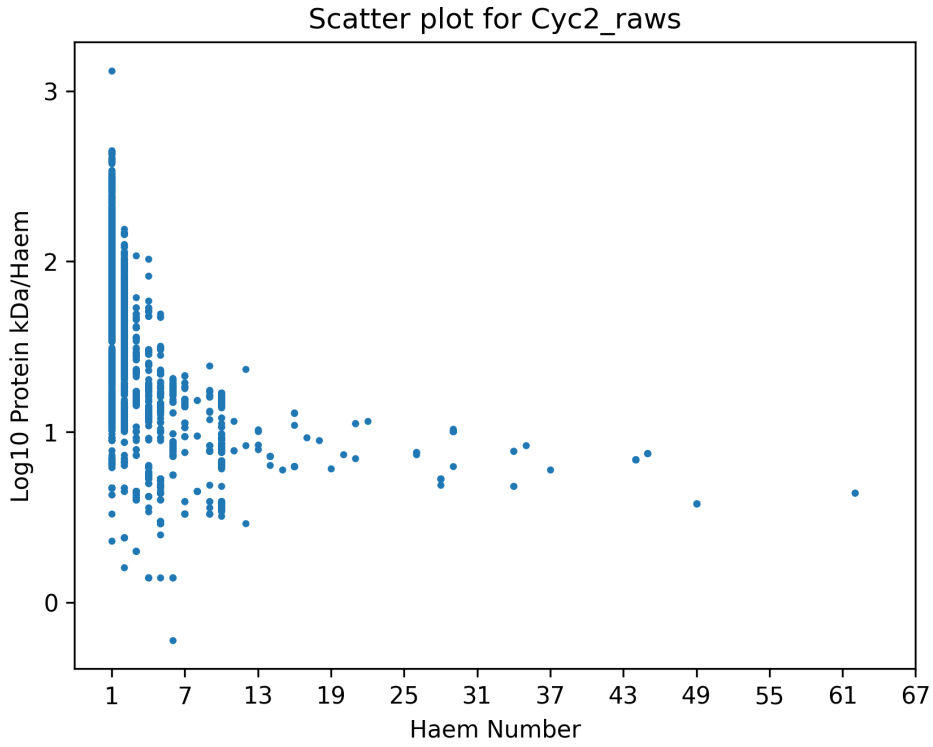


Figure 5.7.3: ***Scatterplot with weight per haem (log_{10} (kDa/haem)) plotted against total haems in protein.***
X-axis ticks automatically selected to be a broad spread to reduce cluttering. All proteins have 12 or more TM strands. Actual organisms (each dot) can be pulled from a related XLSX file sat beside the output plot.

5.7.3 Heatmaps

Occasionally, a more holistic overview of all possible operons is necessary to understand the larger picture involved. For this, we have plotted the TM strand number in the putative porin protein of the operon against the haem number of the putative cytochrome protein of the operon. In this way, we can see trends in the distribution of the data.

ETMiner's automatic heatmaps generated present a wealth of data. With intensity of colour illustrating the (log_{10}) number of operons. This shows what combinations of TMs and Haems are most common, and where they cluster in frequency.

In figure 5.7.4, we can see some interesting results. There appears to be a preference for 1, 6 or 10 haems for Cyc2 (fusion) proteins. In general, however, the broader trend is clear; it is much more common for Cyc2 operons to have lower numbers of TMs and Haems. With our automatic figure generations (with one each for Cyc2, MtrAB, MtrCAB, Other and custom operon types) we can see all of this clearly at-a-glance. Colour can be changed with an option in ETMiner- simply select

the preferred colour from a list (figure 5.8.1).

As well as the general trends, we can pick out the 'outlier' operons. Such as Cyc2 operons with 44 or 45 Haems and 22 TM strands. An interesting cluster for its sheer number of c-type haems and worthy of further scrutiny. Anyone wanting to fully investigate the proteins, their host organism and genomic locations can look in the associated CSV/XLSX files that are printed with every figure.

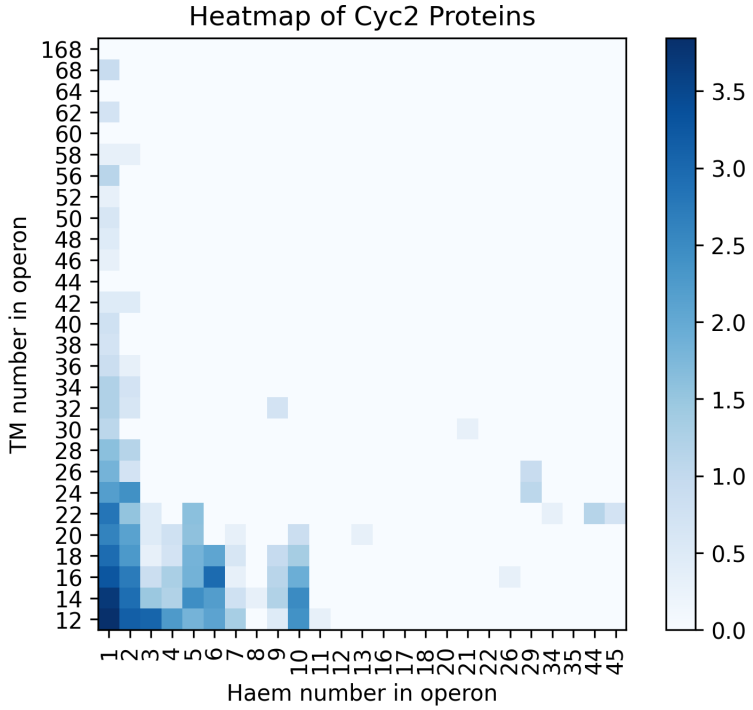


Figure 5.7.4: ***Heatmap of TM strands in an operon's putative porin vs number of hemes predicted from a cytochrome sequence***
Using this heatmap, it is possible to see what TM & haem numbers are most common in bacteria. The clusters might indicate something fundamental about structure and function. Dark blue square indicates a high number of bacterial cytochrome operons in our analysis have that specific TM-Haem count. The specified bacterial operons, and host species, can be found in an accompanying XLSX.

5.7.4 Histograms

Occasionally, researchers want to investigate a single parameter. That is, what is the distribution of Haems among operons where the porin has N number of TM strands?

This is especially useful for when a TM strand number has been identified as interesting. This can be found, for example, in figure 5.7.4. There appear to be many Cyc2 operons with 18 TM strands (as seen by colour intensity), so we choose to investigate the histogram for that specific TM number. As we can see in figure

5.7.5, it matches figure 5.7.4 and highlights count differences.

In figure 5.7.5, we can see the number of haems in each operon that has 18 TM strands for its porin. Automatically generated, an XLSX exists with all the data to be manipulated further. At a glance, we can see that a single haem is the most common for a Cyc2 operon with 18 TM strands, followed by 2 then 6 haems.

Using these two figures, we can get a clear insight into the structure of the data. Delving into their individual XLSX files, we can find the specific operons relating to each figure for future analysis.

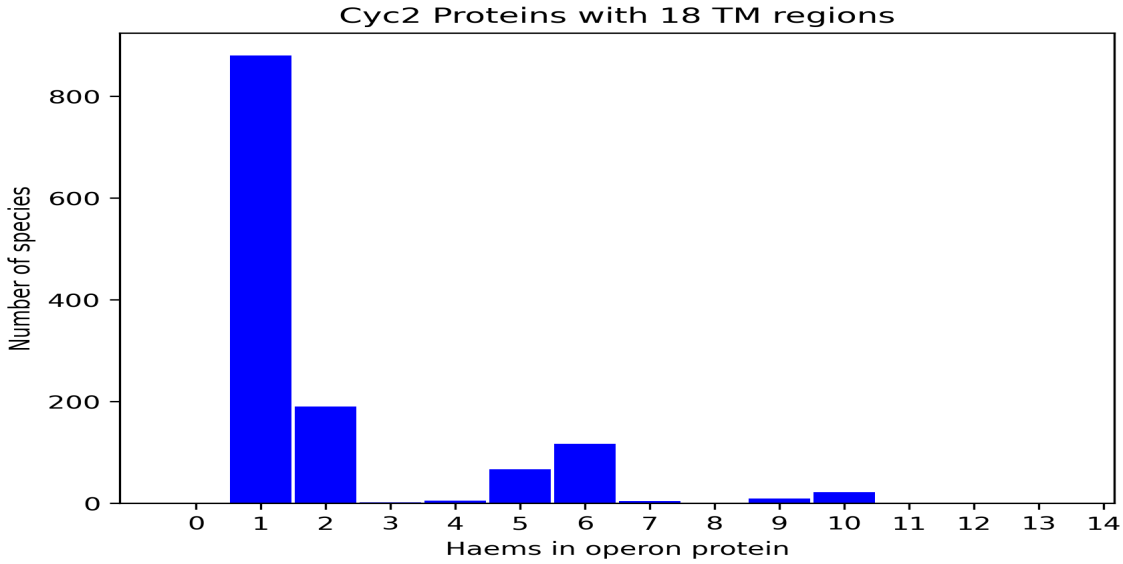


Figure 5.7.5: ***Histogram showing occurrence of Haem numbers in operons with 18 TM strands***
Automatically generated by ETMiner, the figure's graphics are not optimal and simply act as a quick at-a-glance guide of the raw data which is also available in a CSV.

5.7.5 Interactive Tree of Life files

To assist in creating Interactive Tree of Life (iTOL) files, specific files are automatically output by ETMiner. This includes 'mocked-up' iTOL files with some configurations pre-set and all phenotypic data (in this case, count of operon types per taxon ID). Additionally, a list of Taxon IDs are created for phylogenetic tree creating on NCBI and a file listing.

Due to the need to use online resources, instructions for using the data and converting it to a functional tree of life are included in a text file. This provides a step-by-step guide on creating the Phylogenic Tree from the iTOL files output by ETMiner.

The benefit of using online resources to craft the tree of life is that the evolutionary relationships between taxa is kept up-to-date by the NCBI. Mixed with some mocked files for the online resource, it is a quick, easy, and relatively up-to-date system to procure phylogenetic trees.

Following the instructions file, we create a phylogenic tree from the Taxon list. This is then combined with the phenotypic information we decided; in this case, the count of Cyc2, MtraAB, MtrCAB, Other and 'custom' operons found in each species. Using the interactive Tree of Life (iTOL) [18] tool, we are able to visualise this (figure 5.7.6). From it, we can see the broad spread of operons. This is highly relevant to researchers who want to investigate an unusual number of operons in a specific bacterial species.

Moreover, as the operon files, iTOL files and high-res image are all included in the supplementary data, it is possible for researchers to quickly identify an operon of interest to investigate. With both a taxon ID and protein ID (WP redundant code), they can find the exact DNA sequence giving rise to the cytochrome in a bacteria. This DNA sequence can then be investigated in-vitro for insights. For simplicity of reading, the bacterial classes were split into separate colours; green for Gammaproteobacteria, red for Alphaproteobacterial, blue for Betaproteobacteria, purple for Epsilonproteobacteria, orange for Deltaproteobacteria, pink for Bacteroides and, finally, black for everything else.

From figure 5.7.6, we can see that the presence of cytochromes is distributed across many bacterial species- with notable spikes of operon numbers in some species. This gives the image of most bacterial groupings using cytochromes bound to porins for various purposes but with a spread species containing a multitude- perhaps as an indication of the importance of cytochromes within the species.

In any case, this provides plenty of fodder for research with many species that could become of great import to cytochrome research.

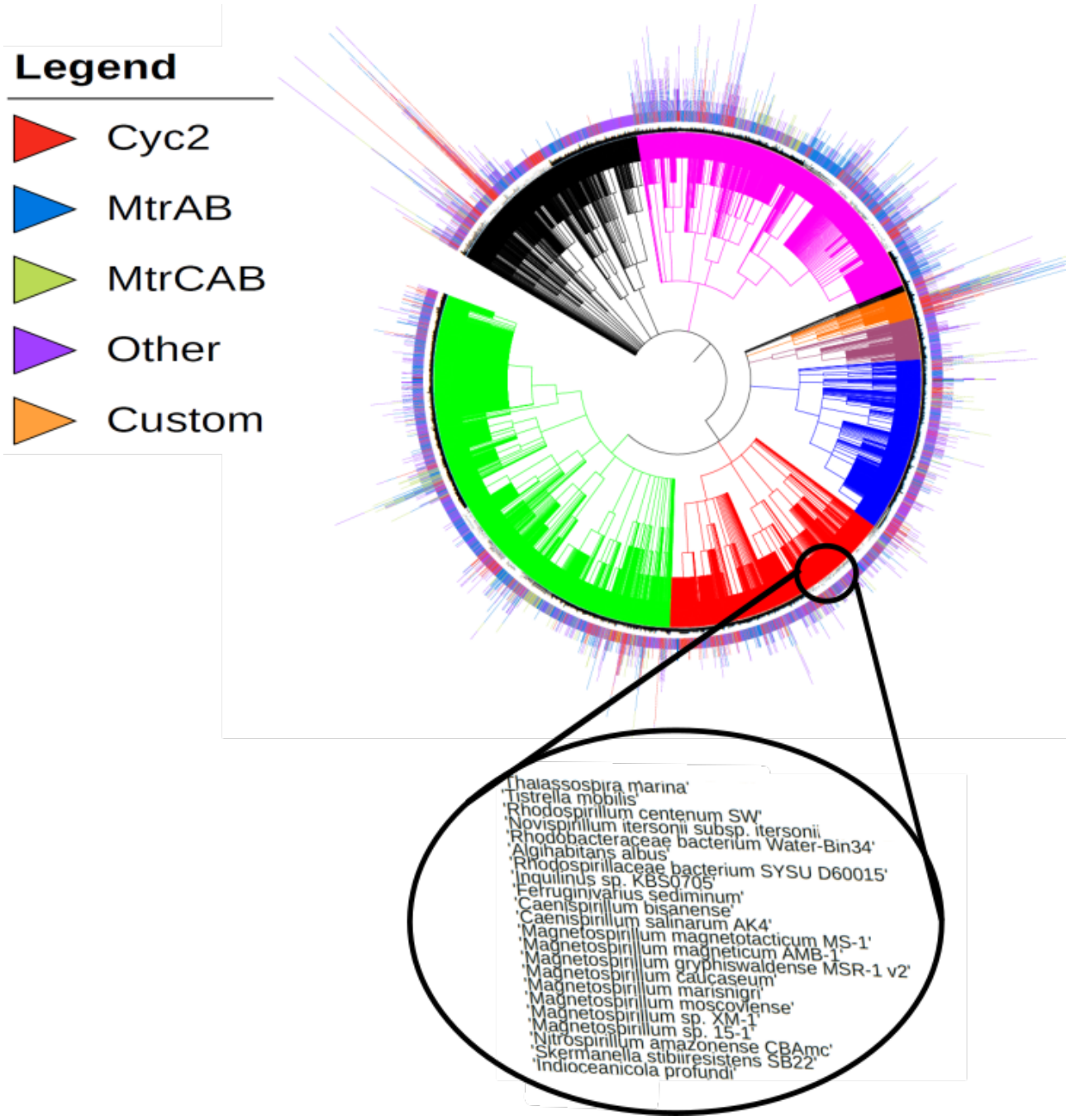


Figure 5.7.6: ***Interactive Tree of Life of all cytochrome operon types across the entire bacterial kingdom***
Not all bacterial species are included; only those we have identified (through Taxon ID) as possessing at least one cytochrome operon were included- with more cytochromes indicated by a longer radiating bar (indicating a higher operon count). Type of operon identified is illustrated by colour (see legend).
Blow-up shows a list of species from the Alphaproteobacterial class (red).
Species lineage is denoted by bacterial class and were split into separate colours; green for Gammaproteobacteria, red for Alphaproteobacterial, blue for Betaproteobacteria, purple for Epsilonproteobacteria, orange for Deltaproteobacteria, pink for Bacteroides and black for everything else.

5.8 ETMiner Images

Figures for explaining the GUI of ETMiner app. In these images, we can see the basic functionality of the ETMiner app and how inputs can be translated to outputs. By manipulating the selections appropriately, we can create the optimal output desired.

In figure 5.8.1, we see the main window of the GUI. By selecting the 'create image' option (figure 5.8.2), it is possible to create an operon image straight from the CSV data of the operon file. Alternatively, the operon file can be searched with a custom structure, for example a porin-haem-porin operon (figure 5.8.3).

By using the 'file' option to add the haem/TM count files as well as the operon CSV file, it becomes possible to quickly analyse the operon database according to any criteria set within the ETMiner app. This enables a user to quickly search for any operon that might match their profile; from molecular weight per haem to haem number and TM number. Multiple colour profiles exist for the automated figures output to match various user preferences/needs.

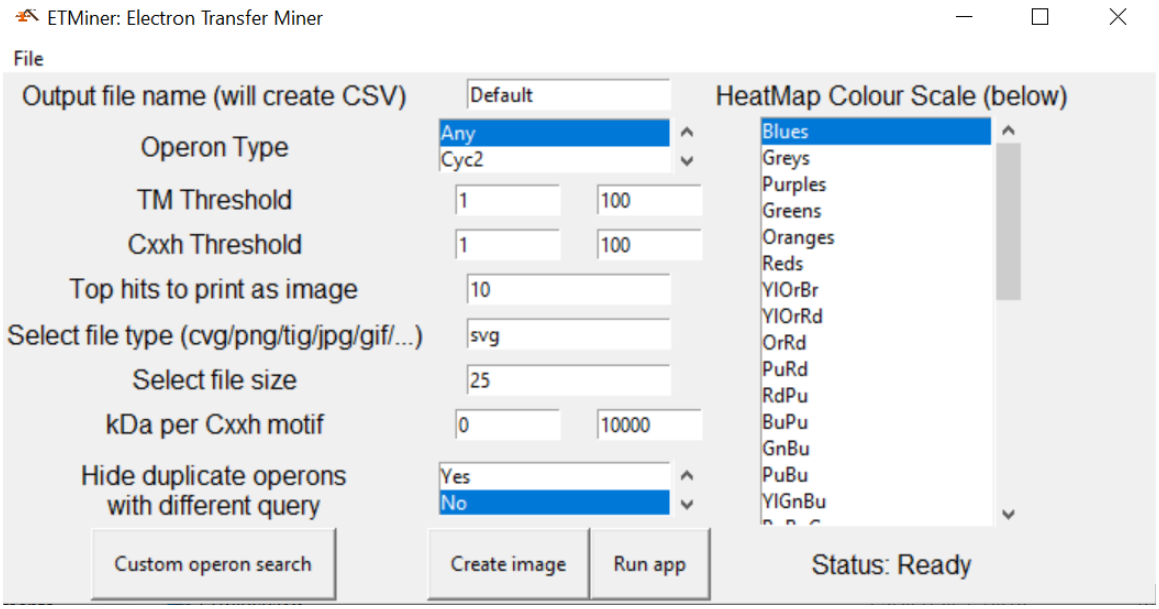


Figure 5.8.1: ***Main ETMiner GUI***
The main window as seen when ETMiner is opened. It holds options such as range of TM/Haems to count in the analysis, weight ratio per haem, colour of resultant heatmaps and more.
The File menu allows selection of the Haem/TM count files, as well as the sequence/operon file.

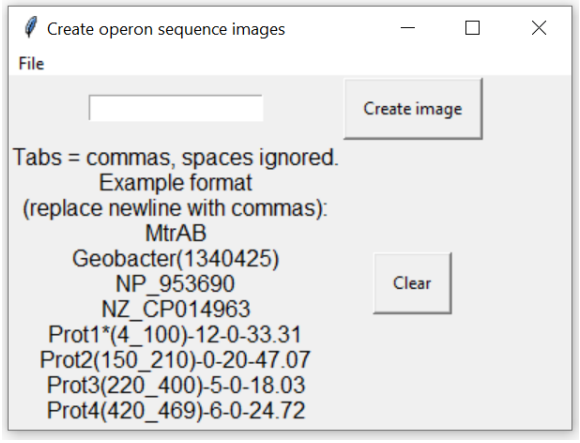


Figure 5.8.2: ***ETMiner image creation from operon***
Allows the conversion of a text-based operon descriptor (CSV format) to be turned into an operon image. Useful for printing out an operon for easier visualisation.

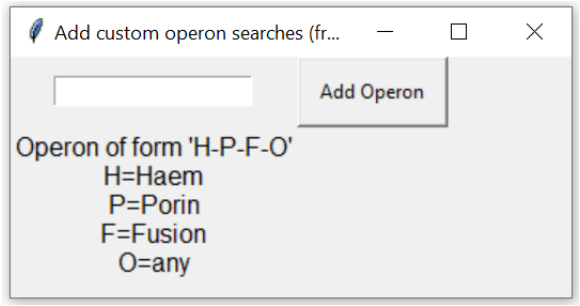


Figure 5.8.3: ***ETMiner custom operon type prediction***
Add custom operon format to search for. For example, a haem flanked by two porins (P-H-P). The haem must be within the boundaries set in figure 5.8.1, and the TM counts for the porin must also be within the range set in 5.8.1

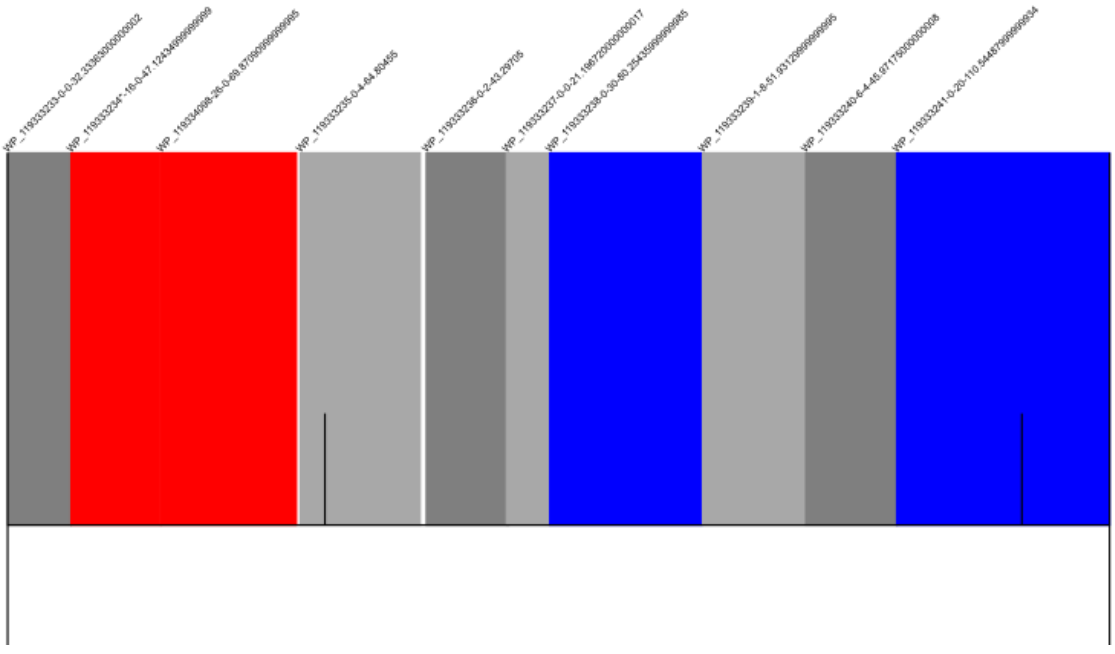


Figure 5.8.4: ***ETMiner example output (single row)***
An operon printed out automatically using figure 5.8.2's functionality. Linear format is shown, but circular is also output. Asterix (*) on operon protein accession indicates match to query protein in BLAST search.

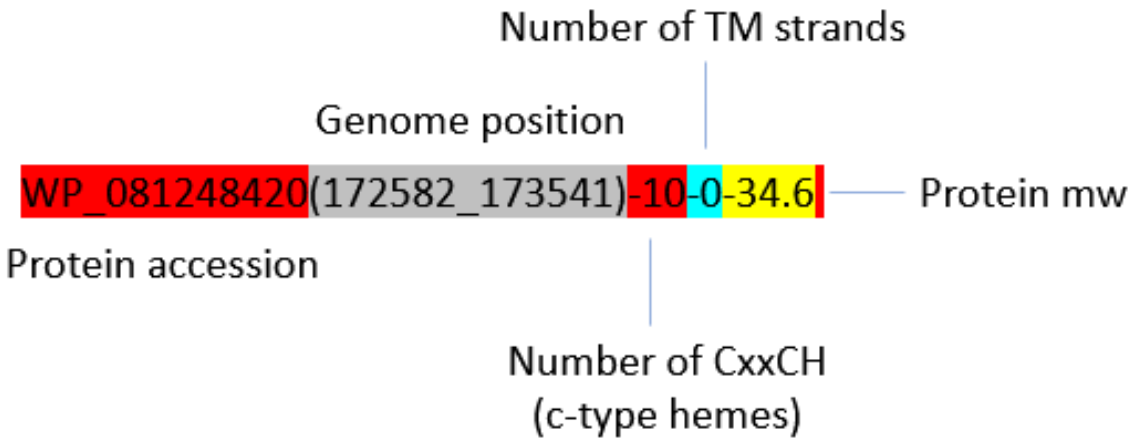


Figure 5.8.5: ***ETMiner output operon barcode explanation***
An explanation of how the barcode in the CSV operon files work. The red is the non-redundant (NR) WP protein accession, followed by it's genomic location in grey and then the number of CXXXCH motifs (haem section), number of TM strands (for porin) and then the molecular weight (MW) in yellow.

5.9 Discussion

From simply the raw data obtained from the RefSeq database, we were able to curate the whole genome sequenced (WGS) bacterial genomes into which possess cytochromes- and whether they belong to the three known types we have previously discussed (figures 5.1.3, 5.1.1 and 5.1.2).

Recent studies have attempted similar; such as FeGenie, which attempts to identify 'iron genes' and 'iron gene neighbourhoods' [178]. However, where we differ is in approach; FeGenie uses Hidden Markov Models (HMM) to predict 'iron genes' in a genome based on protein motifs and requires both a dataset of known 'iron genes' and tests putative 'iron genes' in an iterative fashion. In opposition to FeGenie, this thesis utilised nucleotide matching across the entire bacterial kingdom to identify all, including entirely unknown, cytochromes fitting a rough profile (BLAST[15], section 5.4) and joins them into a resource for all researchers. FeGenie is useful when you want to verify if a protein sequence you've identified matches the 'iron gene' profile, while our ETMiner system helps you locate the sequences in the first place. if you don't have any sequences to test, ETMiner is good at finding a subset of possible sequences.

This is a powerful system that can be applied more generally, with some streamlining and increased computational power for the initial database-searching. Using only information already present in the public domain, we are able to predict haems, TM strands and, eventually, cytochrome type. This allowed us to investigate thousands of bacteria at once (all currently sequenced) for any unusual cytochromes both in structure and species.

The study permits the identification of novel points for potential research. Not only can identified proteins be investigated for function and structure experimentally, experiments can also be undertaken on unusual species to validate their cytochromes. If applied more broadly, various other operons can be identified and studied without the need to test every protein in a database.

Additionally, the process allows for a higher-confidence analysis than a simple BLAST search for homologues. In our approach, we verify haem and TM numbers; something not necessarily identified through BLAST. It is also able to identify interesting proteins that do not share sufficient homology with currently identified cytochromes to be isolated through a BLAST search.

With the automatic creation of figures, and the data placed into XLSX files for ease-of-recreating, we gain an accurate at-a-glance assessment of the data. Any points of interest are then taken up firstly by the XLSX, then carried into the usual bioinformatics analysis of interesting hits. In the future, the proteins can even be analysed experimentally.

The project has highlighted the number, and type, of cytochrome-porin operons across the entire bacterial species. Available publicly as an interactive Tree Of Life (iTOL) phylogenetic tree (figure 5.7.6), raw xlsx data, and iTOL files, it is a valuable resource to identify species of interest or to approximate the spread of operon types across the bacterial kingdom.

This is a significant contribution that can act as a starting point for those wishing to explore various electrogenic bacteria experimentally. However, the usefulness doesn't end there. The methods we've employed can be employed to any other operon people wish to identify.

Following the methodology, any sized database can be analysed for specific operons. By first splitting the database into manageable chunks, BLAST can be launched on hundreds of HPC cores. By itself, BLAST is insufficient. The hits from it provide a useful starting point, however, and the resultant re-inserted hit genes can be analysed in the context of their operons. These operons can be sequentially checked for porin β-sheet motifs, or any other motifs, to analyse their probability of being a gene cluster warranting further study.

Where other studies use HMM to analyse genomes on a as-needed basis, we can identify entirely uncharacterised genomes from the entire repository of bacterial genomes submitted to any fasta-format database. Indeed, this is what we have accomplished in many cases (e.g, *Zobellia uliginosa*).

This broad approach allows for a more holistic view of the bacterial genomes known. Knowing the distribution and type of cytochrome-porins is an important step in the study of all bacteria. In the future, knowing how bacteria interact with, and exploit, minerals in their environment could become an integral concept to understanding the life of any bacteria.

5.10 Conclusions

Overall, the operon prediction part of the thesis can be said to be successful. While the expected operons were drawn out of the RefSeq database, many novel, and some unusual, operons were also identified to act as the basis of future research. Of particular interest may be those operons of unknown type, classified as 'Other', as they could broaden our understanding of electrogenic bacteria- or better define their limits.

The automated figures generated (figures 5.7.3, 5.7.5, 5.7.4) grant an instant overview of the data structure. For example, figure 5.7.3 highlights *Geopsychrobacter electrodiphilus* as a bacteria with a very unusual, as yet experimentally uncharacterised, predicted cytochrome protein with an attached porin.

While the bacteria and its operon might have been identified without the figure, the scatter-plot highlighted it instantly. With the various figures, it is possible to 'lock on' to specific operons and offer conjectures for the more common TM-to-Haem numbers within operons.

In our analysis of all bacterial CDS genomes, we identified the expected culprits when looking for cytochromes. This includes *Shewanella onidensis* [179] (and other *Shewanella* species), *Geobacter sulfurreducens* [180] and *Sideroxydans lithotrophicus* [181] among others. All these bacteria are known electrogenic or metal-respiring bacteria [179, 180, 181]. As expected results, they confirm that we have correctly identified a instances of various cytochromes across the bacterial kingdom.

However, in our identification of bacteria with cytochrome-porin operons, we have found species previously not categorised as electrogenic. Indeed, some have either very little or no information about them, except perhaps some sequencing in the literature (such as *Zobellia uliginosa*[182]).

For example, *Dyella amyloliquefaciens* is a forest soil microbe found in aerobic environments [183] which we predict to have 2 MtrAB cytochrome porin operons. This is surprising as we would expect these operons to be found in more anoxic environments. However, it is still a soil microbe, unlike *Zobellia galactanivorans* which is a marine bacteria found in degrading algae [184] and is predicted to possess 3 Cyc2, 4 MtrAB and 1 'Other' predicted cytochrome porin operons. This lack of a MtrCAB could indicate, as a marine bacteria, it does not need a protein interface (MtrC) to interact with extracellular deposits of metals.

These results indicate that the distribution, function and versatility of these cytochrome-porins are much wider than previously expected. This increases the number of potential commercial applications of the bacteria possessing such operons as there is a higher probability of identifying an operon with characteristics desired by industry or academia. In identifying so many putative operons, ETMiner opens up the world of electrogenic bacteria as never before.

Chapter 6

Discussion

The practice of mapping phenotype to genotype is not a new approach. Far from it; the 'personalised medicine' of the future depends on it [185]. If we are ever to learn to tailor biology to our needs as a society, we will have to master the genotype-phenotype gap. In humans this is often limited to the identification of novel pharmaceutical treatments for rare genetic disorders and diseases. New applications of GWAS technology have recently been carried out for a variety of reasons, such as uncovering the surprising number of genes involved in human eye colour [186].

Each organism has its own genomic peculiarities; bacterial genomes are packed with coding DNA in neat clusters, or 'operons', while yeasts have huge chromosomal variability. There are benefits and drawbacks to each organism in academic and industrial settings. Bacteria are easier to manipulate, but may lack the molecular machinery to create human-like biomolecules. In picking an organism to work with, it is first necessary to consider what is the goal of the project. Yet microbes have many benefits as a whole.

The benefit of microbes as model organisms is the ability to rapidly validate and test any genetic hypothesis through genetic manipulation techniques such as CRISPR- including for modifying the metabolome [187]. In humans such genetic manipulation for basic science is, for obvious reasons, strictly controlled. In addition, microbes' rapid life-cycles make it relatively easy to monitor populations over many generations.

Secondly, using allele frequency variations across many strains is a known technique for population mapping and ancestry prediction, for example in the analysis of human populations [188]. Excepting genes known to be under purifying

selective pressure (for example in protein coding genes), it can be assumed that DNA variations mutate and accumulate randomly over time with few deleterious effects. Often, these gene locations are used to measure evolutionary relationships with models that account for random variation [189, 190].

In a very simplistic (not real-world and without full mathematical proofs) example with only one non-coding locus with alleles A and a, we can attempt a prediction of ancestry for an individual. If allele A is present 90% of the time in population 1, we could assume a 90% probability of any individual with allele A as belonging to population 1 [188]. Once clusters of loci are identified in specific populations and sub-populations, probabilistic models can be tailored for more complex situations such as the origin of entire populations (e.g, human Out Of Africa model) [189]. Similar reasoning is used to build Q-Matrix (chapter 2.2.4) estimates to predict ancestry of members of a species, or sometimes a genus.

Seen in everything from RNA-Seq to GWAS analyses, genetic correlation studies are a well-known tool for discovery and narrowing down of targets of interest *in silico* before experimental studies are carried out. While inconclusive on their own, they provide substantial support to justifying future experimental approaches for validation or further analysis.

Industry and research often focuses on 'high-throughput' techniques to discover novel phenotypes in organisms. However, being able to target microorganisms with probabilistic scores based on genotypes would be a novel approach. The approaches used and discussed present new methodologies for discovering the link between genotype and phenotype.

With ever-cheaper sequencing technologies coupled with the high cost of high-throughput analytical techniques, a new system could be created. This could drastically reduce the time frame needed to identify beneficial strains, as the target group could be curated to remove unlikely strains based on genomic features.

As a whole, techniques used to link genetic variants to phenotypes have been a feature of biological research for many decades. In the studies discussed here, tens of thousands of genetic variants are assessed concurrently. Furthermore, each additional cycle of model usage and validation would improve the predictive value of the model and further improve future predictions (figure 6.0.1).

Using yeast as the model organism, (figure 6.0.1) it is inefficient to perform high-throughput analysis on all yeast strains directly. Because while experimental

analysis is expensive, computational time is cheap. Any reduction in throughput load would improve the model, ensuring gradual increases in accuracy and speed. As both increase, N2 (top hits from model) decreases from dataset N1 (all genomes).

Being able to reduce the number of strains to be analysed would dramatically reduce the time wasted on irrelevant strains. Yeast species are still some one of the many microorganisms that would benefit from this approach; in fact, as eukaryotes, they are among the most difficult to build models of due to their complexity.

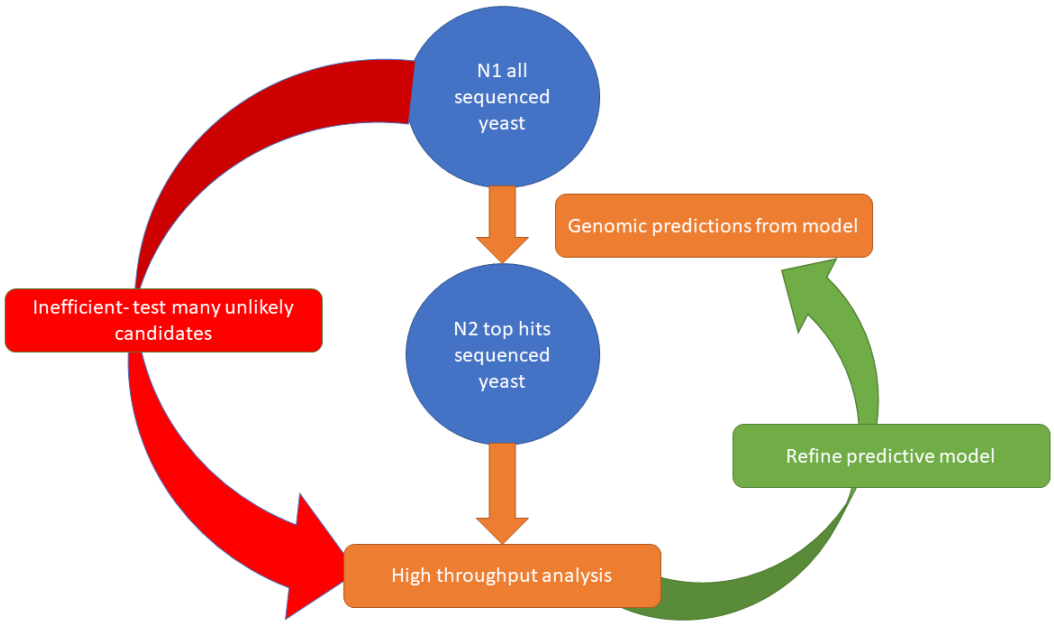


Figure 6.0.1: Recursive improvement of model for trait prediction for high-throughput preparation

Yeast species are therefore not the only ones to benefit from such computational predictions; it applies to everything including bacteria and archaea. Moreover, not all genomic predictions are based upon SNP data. Some models utilise the variability in Copy Number Variations (CNVs) [11] between strains (or individuals) of a species, and many other computational models predict the structure [13] and function [14] of proteins produced by genomic regions.

With our SNP data as the basis, we did a furfural resistance study on 168 *S. cerevisiae* yeast strains. To accomplish this, we developed a new computational method to predict a holistic resistance score from yeast growth curves (as seen in Chapter 3). This method appears to work well and can be used in other studies. We developed a new computational method (SANE) to estimate a Q-Matrix representing the ancestry and inter-relationships of our yeast strains (section 3.3). A Mantel test applied to our yeast dataset suggested the results are highly similar to

those estimated using an established method (PSIKO), though they are not identical, and they share 2283 out of the top 10000 hits. More work would be needed to fully validate and study the approach.

With our resistance phenotype, SNP genomes and custom SANE Q-Matrix, we did a GWAS for the strains. We did not find any compelling results from the analysis, likely because the study was too small. However, some potentially interesting results were identified; we found SNP hits in our study for genes important to resistance to cellular stressors (YLR247C, YGL093W, YOL078) such as the predicted broad-spectrum effects of furfuraldehyde. We then progressed to a Directed Evolution experiment. While resistance scores did not grow appreciably, we did find some interesting results such as the fixation of many of the top predicted alleles in the main study. Some shortcomings were both lack of time (started at the beginning of the pandemic and lab access was reduced) and small number of strains. If the study was scaled up, however, it would be a complete methodology for the improvement of yeast strains in regards to inhibitor resistance.

We did not limit ourselves to inhibitor resistance as our phenotype for the genotypes' GWAS. We performed a metabolomic study on two yeast datasets. Firstly, we looked at a broad set of 362 yeast strains (from 118 species), growing on YNB (with 10g/L glucose) media, to simulate basic growing conditions. Secondly, we examined the 168 Saccharomyces cerevisiae strains from the furfural study growing on Malt media, to replicate brewing conditions. NMR metabolite levels were quantified using the CHENOMX software which showed a huge variability between species and even strains. Several notable effects were apparent.

Generally, metabolite quantities from yeast growing on malt extract were considerably higher than on YNB. Examining a set of 50 S. cerevisiae strains common to both datasets showed that fold levels were as high as 218,200 times for the Malt study (Acetate for NCYC 3612). A GWAS conducted on the second, Malt dataset indicated several SNPs significantly correlated with metabolite levels, that will highly interesting and potentially valuable to follow up in the future. Growing the strain dataset would only increase the predictive capabilities of the GWAS undertaken; in any case, the metabolic data is a valuable resource that can sit alongside other NCYC information.

In this body of work, a specific tool was created called Electron Transport Miner (ETMiner). This tool, alongside other upstream pipelines, use function

and structure predicting tools (Prosite, TMBB2, BLAST) to predict cytochrome and porins within all known bacterial species. Bacteria possessing gene clusters (Operons) with specific cytochrome-porin structures (haem counts, cellular localisation signals,...) are predicted to be specific operons used by electrogenic bacteria and are, in turn, putative electrogenic bacteria themselves.

ETMiner permitted the assembling of operons, automated creation of relevant figures for the analysis of the upstream analyses (heatmaps, histograms, scatterplots). ETMiner allows the interrogation of an operon database with specific protein prediction data (haem count, β-strand number) to identify operons that fit a specific criteria including molecular weight ratios and type of operon (MtrCAB, MtrAB, Cyc2).

Subsequent studies might investigate SNPs within identified genomic regions, yet our preliminary research provides a start point for those wishing to locate specific operon structures within unknown genomes. This is a valuable dataset for anyone wishing to search

We were able to find the spread, frequency and depth of specific cytochrome operon types across the breadth of known bacterial genomes. This includes those of as-yet unknown or 'Other' structure. A vital start-point for researchers, it can provide a start point for where to attempt to find cytochrome-porins for electrogenic bacterial identification purposes. These bacteria could be useful for a range of purposes such as bioremediation (waste water treatment, heavy metal extraction), electricity generation, or chemosynthesis using electricity.

This thesis has explored many of the ways genomic data can be used as the foundation for predictions of phenotypic characteristics. These varied methods are used in different fields for different purposes but all expand our knowledge of genomic data and how to interpret it. The predictions generated and the models build a foundation for others to build upon. Whether future researchers are searching for genes related to a range of metabolite expressions levels in yeast strains, or which bacteria to investigate for commercial bioremediation niches, the body of research has been expanded appropriately.

6.1 Future work

Future work for this body of work would include the validation of SANE as a robust Q-Matrix founder-population predictor. This would aid in the next segments; performing the GWAS on a larger strain dataset to increase the statistical power of the correlations performed. A longer DE experiment would likely increase furfuraldehyde resistance better than our shorter one was able to. A CRISPR study focussed on validating identified SNPs would provide experimental evidence to the efficacy of the GWAS carried out.

Talking to brewers would be a great way to ensure the experiments and analysis were of material use to them. Maintaining a dialogue in future work would ensure that the goals of the research remained aligned with the needs of brewers.

The predictions made by ETMiner are expansive. Thousands of predicted cytochrome-porins have been identified that vastly expands the scope of potential electrogenic bacteria in nature. Experimental validation should be carried out to verify and refine the model. How many of the predicted operons are functioning as expected? Are there other functions for the protein structures as-yet not understood? Are there constraints in structure we were unaware of going into the study that should be used on the next iteration?

Nonetheless, the methodology of ETMiner (from server pipeline to desktop GUI app) can be replicated for any number of genomic structures. Improvements could be performed to enable rapid pipeline-building and database construction based on any number of genomic structures to be identified across the entire set of known bacterial genomes.

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

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

NCYC strain	Species	Sequencing plate 1	Sequencing plate 2	Sequencing plate 3
70	*Saccharomyces cerevisiae*	Plate 1		
72	*Saccharomyces cerevisiae*	Plate 1		
74	*Saccharomyces cerevisiae*	Plate 1		
76	*Saccharomyces cerevisiae*	Plate 1		
77	*Saccharomyces cerevisiae*	Plate 1		
78	*Saccharomyces cerevisiae*	Plate 1		
79	*Saccharomyces cerevisiae*	Plate 1		
80	*Saccharomyces cerevisiae*	Plate 1		
81	*Saccharomyces cerevisiae*	Plate 1		
82	*Saccharomyces cerevisiae*	Plate 1		
83	*Saccharomyces cerevisiae*	Plate 1		
84	*Saccharomyces cerevisiae*	Plate 1	Plate 10	
85	*Saccharomyces cerevisiae*	Plate 1		
86	*Saccharomyces cerevisiae*	Plate 1		
87	*Saccharomyces cerevisiae*	Plate 1	Plate 10	

NCYC strain	Species	Sequencing plate 1	Sequencing plate 2	Sequencing plate 3
88	*Saccharomyces cerevisiae*	Plate 1		
89	*Saccharomyces cerevisiae*	Plate 1		
90	*Saccharomyces cerevisiae*	Plate 1		
91	*Saccharomyces cerevisiae*	Plate 1	Plate 10	
92	*Saccharomyces cerevisiae*	Plate 1		
93	*Saccharomyces cerevisiae*	Plate 1		
95	*Saccharomyces cerevisiae*	Plate 1		
96	*Saccharomyces cerevisiae*	Plate 1		
97	*Saccharomyces cerevisiae*	Plate 1		
167	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
192	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
196	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
197	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
200	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
205	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
206	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
208	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	Plate 11
210	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
211	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
212	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
213	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
221	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
222	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
223	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
224	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
225	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
228	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
232	*Saccharomyces cerevisiae*	Plate 4		
235	*Saccharomyces cerevisiae*	Plate 4		
241	*Saccharomyces cerevisiae*	Plate 5		
356	*Saccharomyces cerevisiae*	Plate 5		
357	*Saccharomyces cerevisiae*	Plate 5		

NCYC strain	Species	Sequencing plate 1	Sequencing plate 2	Sequencing plate 3
358	*Saccharomyces cerevisiae*	Plate 5		
360	*Saccharomyces cerevisiae*	Plate 4		
361	*Saccharomyces cerevisiae*	Plate 4		
430	*Saccharomyces cerevisiae*	Plate 5		
478	*Saccharomyces cerevisiae*	Plate 5		
479	*Saccharomyces cerevisiae*	Plate 5		
482	*Saccharomyces cerevisiae*	Plate 5		
490	*Saccharomyces cerevisiae*	Plate 5		
491	*Saccharomyces cerevisiae*	Plate 5		
505	*Saccharomyces cerevisiae*	Plate 1		
609	*Saccharomyces cerevisiae*	Plate 6		
619	*Saccharomyces cerevisiae*	Plate 5		
620	*Saccharomyces cerevisiae*	Plate 5		
621	*Saccharomyces cerevisiae*	Plate 5		
667	*Saccharomyces cerevisiae*	Plate 4		
672	*Saccharomyces cerevisiae*	Plate 5		
684	*Saccharomyces cerevisiae*	Plate 5		
695	*Saccharomyces cerevisiae*	Plate 4		
816	*Saccharomyces cerevisiae*	Plate 5		
1006	*Saccharomyces cerevisiae*	Plate 1		
1026	*Saccharomyces cerevisiae*	Plate 1	Plate 10	
1064	*Saccharomyces cerevisiae*	Plate 4		
1151	*Saccharomyces cerevisiae*	Plate 4		
1228	*Saccharomyces cerevisiae*	Plate 1		
1245	*Saccharomyces cerevisiae*	Plate 1		
1315	*Saccharomyces cerevisiae*	Plate 6		
1337	*Saccharomyces cerevisiae*	Plate 4		
1406	*Saccharomyces cerevisiae*	Plate 5		
1407	*Saccharomyces cerevisiae*	Plate 5		
1408	*Saccharomyces cerevisiae*	Plate 5		
1409	*Saccharomyces cerevisiae*	Plate 5		
1413	*Saccharomyces cerevisiae*	Plate 5		

NCYC strain	Species	Sequencing plate 1	Sequencing plate 2	Sequencing plate 3
1414	*Saccharomyces cerevisiae*	Plate 5		
1415	*Saccharomyces cerevisiae*	Plate 5		
1444	*Saccharomyces cerevisiae*	Plate 4		
1529	*Saccharomyces cerevisiae*	Plate 5		
1603	*Saccharomyces cerevisiae*	Plate 4		
1681	*Saccharomyces cerevisiae*	Plate 1		
2397	*Saccharomyces cerevisiae*	Plate 4		
2401	*Saccharomyces cerevisiae*	Plate 5		
2517	*Saccharomyces cerevisiae*	Plate 5		
2592	*Saccharomyces cerevisiae*	Plate 4		
2688	*Saccharomyces cerevisiae*	Plate 5		
2733	*Saccharomyces cerevisiae*	Plate 4		
2737	*Saccharomyces cerevisiae*	Plate 4		
2776	*Saccharomyces cerevisiae*	Plate 7		
2777	*Saccharomyces cerevisiae*	Plate 7		
2778	*Saccharomyces cerevisiae*	Plate 7		
2779	*Saccharomyces cerevisiae*	Plate 7		
2780	*Saccharomyces cerevisiae*	Plate 7		
2798	*Saccharomyces cerevisiae*	Plate 7		
2826	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
2855	*Saccharomyces cerevisiae*	Plate 5		
2945	*Saccharomyces cerevisiae*	Plate 4		
2947	*Saccharomyces cerevisiae*	Plate 5		
2948	*Saccharomyces cerevisiae*	Plate 5		
2967	*Saccharomyces cerevisiae*	Plate 7		
2974	*Saccharomyces cerevisiae*	Plate 7		
3025	*Saccharomyces cerevisiae*	Plate 7		
3026	*Saccharomyces cerevisiae*	Plate 7		
3028	*Saccharomyces cerevisiae*	Plate 7		
3030	*Saccharomyces cerevisiae*	Plate 7		
3031	*Saccharomyces cerevisiae*	Plate 7		
3032	*Saccharomyces cerevisiae*	Plate 7		

NCYC strain	Species	Sequencing plate 1	Sequencing plate 2	Sequencing plate 3
3033	*Saccharomyces cerevisiae*	Plate 7		
3035	*Saccharomyces cerevisiae*	Plate 7		
3036	*Saccharomyces cerevisiae*	Plate 7		
3037	*Saccharomyces cerevisiae*	Plate 7		
3038	*Saccharomyces cerevisiae*	Plate 7		
3039	*Saccharomyces cerevisiae*	Plate 7		
3051	*Saccharomyces cerevisiae*	Plate 7		
3052	*Saccharomyces cerevisiae*	Plate 7		
3076	*Saccharomyces cerevisiae*	Plate 7		
3077	*Saccharomyces cerevisiae*	Plate 7		
3078	*Saccharomyces cerevisiae*	Plate 7		
3114	*Saccharomyces cerevisiae*	Plate 7		
3121	*Saccharomyces cerevisiae*	Plate 7		
3122	*Saccharomyces cerevisiae*	Plate 7		
3123	*Saccharomyces cerevisiae*	Plate 7		
3124	*Saccharomyces cerevisiae*	Plate 7		
3125	*Saccharomyces cerevisiae*	Plate 7		
3126	*Saccharomyces cerevisiae*	Plate 7		
3127	*Saccharomyces cerevisiae*	Plate 7		
3265	*Saccharomyces cerevisiae*	Plate 4		
3266	*Saccharomyces cerevisiae*	Plate 4		
3311	*Saccharomyces cerevisiae*	Plate 4		
3313	*Saccharomyces cerevisiae*	Plate 4		
3314	*Saccharomyces cerevisiae*	Plate 4		
3315	*Saccharomyces cerevisiae*	Plate 4		
3318	*Saccharomyces cerevisiae*	Plate 4		
3319	*Saccharomyces cerevisiae*	Plate 4		
3324	*Saccharomyces cerevisiae*	Plate 7		
3325	*Saccharomyces cerevisiae*	Plate 7		
3326	*Saccharomyces cerevisiae*	Plate 7		
3331	*Saccharomyces cerevisiae*	Plate 7		
3333	*Saccharomyces cerevisiae*	Plate 7		

NCYC strain	Species	Sequencing plate 1	Sequencing plate 2	Sequencing plate 3
3334	*Saccharomyces cerevisiae*	Plate 7		
3338	*Saccharomyces cerevisiae*	Plate 7		
3339	*Saccharomyces cerevisiae*	Plate 7		
3406	*Saccharomyces cerevisiae*	Plate 4		
3445	*Saccharomyces cerevisiae*	Plate 4		
3447	*Saccharomyces cerevisiae*	Plate 4		
3448	*Saccharomyces cerevisiae*	Plate 4		
3449	*Saccharomyces cerevisiae*	Plate 4		
3452	*Saccharomyces cerevisiae*	Plate 3	Plate 3B	
3455	*Saccharomyces cerevisiae*	Plate 4		
3456	*Saccharomyces cerevisiae*	Plate 4		
3457	*Saccharomyces cerevisiae*	Plate 4		
3458	*Saccharomyces cerevisiae*	Plate 4		
3460	*Saccharomyces cerevisiae*	Plate 4		
3461	*Saccharomyces cerevisiae*	Plate 4		
3462	*Saccharomyces cerevisiae*	Plate 4		
3467	*Saccharomyces cerevisiae*	Plate 4		
3470	*Saccharomyces cerevisiae*	Plate 4		
3471	*Saccharomyces cerevisiae*	Plate 4		
3472	*Saccharomyces cerevisiae*	Plate 4		
3486	*Saccharomyces cerevisiae*	Plate 4		
3487	*Saccharomyces cerevisiae*	Plate 4		
3557	*Saccharomyces cerevisiae*	Plate 6		
3612	*Saccharomyces cerevisiae*	Plate 4		
3630	*Saccharomyces cerevisiae*	Plate 4		

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

NCYC strain	**Species**	**YNB study**	**Malt Study**
1	*Candida famata var. famata*	YES	NO
2	*Dekkera anomala*	YES	NO
4	*Candida tropicalis*	YES	NO
6	*Candida kefyr*	YES	NO
8	*Debaryomyces hansenii*	YES	NO
9	*Debaryomyces hansenii*	YES	NO
10	*Debaryomyces hansenii*	YES	NO
16	*Pichia subpelliculosa*	YES	NO
17	*Hanseniaspora valbyensis*	YES	NO
18	*Pichia anomala*	YES	NO
20	*Pichia anomala*	YES	NO
21	*Pichia membranifaciens*	YES	NO
22	*Williopsis saturnus var. saturnus*	YES	NO
23	*Williopsis saturnus var. saturnus*	YES	NO
26	*Kloeckera africana*	YES	NO
31	*Kloeckera corticis*	YES	NO
36	*Hanseniaspora vineae*	YES	NO
39	*Candida catenulata*	YES	NO
40	*Guilliermondella selenospora*	YES	NO
43	*Candida krusei*	YES	NO
44	*Pichia membranifaciens*	YES	NO
45	*Candida krusei*	YES	NO
46	*Nadsonia fulvescens Var. fulvescens*	YES	NO
49	*Geotrichum candidum*	YES	NO
51	*Pichia membranifaciens*	YES	NO
52	*Pichia membranifaciens*	YES	NO
54	*Pichia membranifaciens*	YES	NO
55	*Pichia membranifaciens*	YES	NO
57	*Williopsis saturnus var. saturnus*	YES	NO
58	*Kloeckera africana*	YES	NO

59	*Rhodotorula glutinis var. glutinis*	YES	NO
60	*Rhodotorula glutinis var. glutinis*	YES	NO
61	*Rhodotorula glutinis var. glutinis*	YES	NO
62	*Rhodotorula minuta var. minuta*	YES	NO
63	*Rhodotorula mucilaginosa*	YES	NO
64	*Rhodotorula mucilaginosa*	YES	NO
65	*Rhodotorula mucilaginosa*	YES	NO
68	*Rhodotorula mucilaginosa*	YES	NO
70	*Saccharomyces cerevisiae*	NO	YES
71	*Candida famata var. famata*	YES	NO
72	*Saccharomyces cerevisiae*	NO	YES
74	*Saccharomyces cerevisiae*	NO	YES
76	*Saccharomyces cerevisiae*	NO	YES
77	*Saccharomyces cerevisiae*	NO	YES
78	*Saccharomyces cerevisiae*	NO	YES
79	*Saccharomyces cerevisiae*	NO	YES
80	*Saccharomyces cerevisiae*	NO	YES
81	*Saccharomyces cerevisiae*	NO	YES
82	*Saccharomyces cerevisiae*	NO	YES
83	*Saccharomyces cerevisiae*	NO	YES
84	*Saccharomyces cerevisiae*	NO	YES
85	*Saccharomyces cerevisiae*	NO	YES
86	*Saccharomyces cerevisiae*	NO	YES
87	*Saccharomyces cerevisiae*	NO	YES
88	*Saccharomyces cerevisiae*	NO	YES
89	*Saccharomyces cerevisiae*	NO	YES
90	*Saccharomyces cerevisiae*	NO	YES
91	*Saccharomyces cerevisiae*	NO	YES
92	*Saccharomyces cerevisiae*	NO	YES
93	*Saccharomyces cerevisiae*	NO	YES
95	*Saccharomyces cerevisiae*	NO	YES
96	*Saccharomyces cerevisiae*	NO	YES

97	*Saccharomyces cerevisiae*	NO	YES
100	*Kluyveromyces marxianus*	YES	NO
111	*Kluyveromyces marxianus*	YES	NO
128	*Zygosaccharomyces bailii*	YES	NO
135	*Rhodotorula mucilaginosa*	YES	NO
138	*Rhodotorula aurantiaca*	YES	NO
140	*Candida colliculosa*	YES	NO
141	*Candida colliculosa*	YES	NO
142	*Rhodotorula mucilaginosa*	YES	NO
143	*Candida kefyr*	YES	NO
147	*Torulaspora delbrueckii*	YES	NO
151	*Kluyveromyces marxianus*	YES	NO
152	*Candida kefyr*	YES	NO
154	*Rhodotorula glutinis var. glutinis*	YES	NO
155	*Rhodotorula glutinis var. glutinis*	YES	NO
158	*Rhodotorula mucilaginosa*	YES	NO
159	*Rhodotorula mucilaginosa*	YES	NO
161	*Torulaspora delbrueckii*	YES	NO
162	*Rhodotorula glutinis var. glutinis*	YES	NO
167	*Saccharomyces cerevisiae*	NO	YES
171	*Zygosaccharomyces bisporus*	YES	NO
179	*Kluyveromyces marxianus*	YES	NO
188	*Candida kefyr*	YES	NO
192	*Saccharomyces cerevisiae*	NO	YES
195	*Rhodotorula mucilaginosa*	YES	NO
196	*Saccharomyces cerevisiae*	NO	YES
197	*Saccharomyces cerevisiae*	NO	YES
200	*Saccharomyces cerevisiae*	NO	YES
205	*Saccharomyces cerevisiae*	NO	YES
206	*Saccharomyces cerevisiae*	NO	YES
208	*Saccharomyces cerevisiae*	NO	YES
210	*Saccharomyces cerevisiae*	NO	YES

211	*Saccharomyces cerevisiae*	NO	YES
212	*Saccharomyces cerevisiae*	NO	YES
213	*Saccharomyces cerevisiae*	NO	YES
221	*Saccharomyces cerevisiae*	NO	YES
222	*Saccharomyces cerevisiae*	NO	YES
223	*Saccharomyces cerevisiae*	NO	YES
224	*Saccharomyces cerevisiae*	NO	YES
225	*Saccharomyces cerevisiae*	NO	YES
228	*Saccharomyces cerevisiae*	NO	YES
232	*Saccharomyces cerevisiae*	YES	YES
235	*Saccharomyces cerevisiae*	YES	YES
241	*Saccharomyces cerevisiae*	NO	YES
243	*Kluyveromyces marxianus*	YES	NO
244	*Kluyveromyces marxianus*	YES	NO
350	*Candida glabrata*	YES	NO
356	*Saccharomyces cerevisiae*	NO	YES
357	*Saccharomyces cerevisiae*	NO	YES
358	*Saccharomyces cerevisiae*	NO	YES
360	*Saccharomyces cerevisiae*	YES	YES
361	*Saccharomyces cerevisiae*	YES	YES
371	*Metschnikowia pulcherrima*	YES	NO
372	*Metschnikowia pulcherrima*	YES	NO
373	*Metschnikowia pulcherrima*	YES	NO
377	*Rhodotorula glutinis var. glutinis*	YES	NO
385	*Zygosaccharomyces bailii*	YES	NO
388	*Candida glabrata*	YES	NO
392	*Saccharomyces pastorianus*	YES	NO
408	*Torulaspora delbrueckii*	YES	NO
416	*Kluyveromyces lactis*	YES	NO
417	*Zygosaccharomyces bailii*	YES	NO
426	*Kluyveromyces marxianus*	YES	NO
430	*Saccharomyces cerevisiae*	NO	YES

431	*Saccharomyces cerevisiae*	YES	NO
464	*Zygosaccharomyces bailii*	YES	NO
469	*Kluyveromyces lactis*	YES	NO
478	*Saccharomyces cerevisiae*	NO	YES
479	*Saccharomyces cerevisiae*	NO	YES
482	*Saccharomyces cerevisiae*	NO	YES
490	*Saccharomyces cerevisiae*	NO	YES
491	*Saccharomyces cerevisiae*	NO	YES
492	*Torulaspora delbrueckii*	YES	NO
502	*Rhodotorula graminis*	YES	NO
505	*Saccharomyces cerevisiae*	YES	YES
523	*Kluyveromyces polysporus*	YES	NO
524	*Torulaspora pretoriensis*	YES	NO
538	*Kluyveromyces dobzhanskii*	YES	NO
539	*Rhodotorula minuta var. minuta*	YES	NO
541	*Rhodotorula minuta var. minuta*	YES	NO
543	*Saccharomyces kluyveri*	YES	NO
546	*Kluyveromyces wickerhamii*	YES	NO
548	*Kluyveromyces lactis*	YES	NO
551	*Kluyveromyces lactis*	YES	NO
559	*Torulaspora delbrueckii*	YES	NO
563	*Zygosaccharomyces bailii*	YES	NO
566	*Torulaspora delbrueckii*	YES	NO
568	*Zygosaccharomyces rouxii*	YES	NO
570	*Kluyveromyces lactis*	YES	NO
571	*Kluyveromyces lactis*	YES	NO
573	*Zygosaccharomyces bailii*	YES	NO
575	*Kluyveromyces lactis*	YES	NO
580	*Zygosaccharomyces bailii*	YES	NO
582	*Torulaspora delbrueckii*	YES	NO
585	*Torulaspora delbrueckii*	YES	NO
587	*Kluyveromyces marxianus*	YES	NO

608	*Candida colliculosa*	YES	NO
609	*Saccharomyces cerevisiae*	YES	NO
619	*Saccharomyces cerevisiae*	NO	YES
620	*Saccharomyces cerevisiae*	NO	YES
621	*Saccharomyces cerevisiae*	NO	YES
667	*Saccharomyces cerevisiae*	YES	YES
672	*Saccharomyces cerevisiae*	NO	YES
677	*Torulaspora delbrueckii*	YES	NO
684	*Saccharomyces cerevisiae*	NO	YES
695	*Saccharomyces cerevisiae*	YES	YES
696	*Torulaspora delbrueckii*	YES	NO
731	*Saccharomycodes ludwigii*	YES	NO
739	*Saccharomyces cerevisiae*	YES	NO
744	*Candida kefyr*	YES	NO
745	*Metschnikowia reukaufii*	YES	NO
747	*Metschnikowia pulcherrima*	YES	NO
752	*Kluyveromyces lactis*	YES	NO
754	*Saccharomyces cerevisiae*	YES	NO
758	*Rhodotorula mucilaginosa*	YES	NO
768	*Kluyveromyces delphensis*	YES	NO
776	*Kluyveromyces lactis*	YES	NO
777	*Saccharomyces dairenensis*	YES	NO
783	*Metschnikowia zobellii*	YES	NO
794	*Metschnikowia zobellii*	YES	NO
796	*Rhodotorula mucilaginosa*	YES	NO
797	*Rhodotorula mucilaginosa*	YES	NO
807	*Saccharomyces cerevisiae*	YES	NO
814	*Saccharomyces exiguus*	YES	NO
816	*Saccharomyces cerevisiae*	NO	YES
820	*Torulaspora globosa*	YES	NO
826	*Saccharomyces cerevisiae*	YES	NO
827	*Kluyveromyces marxianus*	YES	NO

844	*Rhodotorula minuta var. minuta*	YES	NO
845	*Rhodotorula minuta var. minuta*	YES	NO
851	*Kluyveromyces marxianus*	YES	NO
894	*Metschnikowia lunata*	YES	NO
906	*Candida kefyr*	YES	NO
911	*Pseudozyma aphidis*	YES	NO
929	*Kluyveromyces lactis*	YES	NO
931	*Rhodotorula minuta var. minuta*	YES	NO
935	*Saccharomyces cerevisiae*	YES	NO
956	*Saccharomyces cerevisiae*	YES	NO
970	*Kluyveromyces marxianus*	YES	NO
971	*Saccharomyces unisporus*	YES	NO
974	*Rhodotorula glutinis var. glutinis*	YES	NO
975	*Saccharomyces pastorianus*	YES	NO
1006	*Saccharomyces cerevisiae*	YES	YES
1026	*Saccharomyces cerevisiae*	YES	YES
1063	*Saccharomyces cerevisiae*	YES	NO
1064	*Saccharomyces cerevisiae*	YES	YES
1151	*Saccharomyces cerevisiae*	YES	YES
1187	*Saccharomyces cerevisiae*	YES	NO
1228	*Saccharomyces cerevisiae*	YES	YES
1245	*Saccharomyces cerevisiae*	YES	YES
1315	*Saccharomyces cerevisiae*	YES	NO
1337	*Saccharomyces cerevisiae*	YES	YES
1368	*Kluyveromyces lactis*	YES	NO
1384	*Pseudozyma fusiformata*	YES	NO
1400	*Zygosaccharomyces bailii*	YES	NO
1406	*Saccharomyces cerevisiae*	NO	YES
1407	*Saccharomyces cerevisiae*	NO	YES
1408	*Saccharomyces cerevisiae*	NO	YES
1409	*Saccharomyces cerevisiae*	NO	YES
1413	*Saccharomyces cerevisiae*	NO	YES

1414	*Saccharomyces cerevisiae*	NO	YES
1415	*Saccharomyces cerevisiae*	NO	YES
1416	*Zygosaccharomyces bailii*	YES	NO
1417	*Kluyveromyces lodderae*	YES	NO
1424	*Kluyveromyces marxianus*	YES	NO
1425	*Kluyveromyces marxianus*	YES	NO
1426	*Kluyveromyces marxianus*	YES	NO
1429	*Kluyveromyces marxianus*	YES	NO
1441	*Candida kefyr*	YES	NO
1444	*Saccharomyces cerevisiae*	YES	YES
1449	*Candida bombicola*	YES	NO
1495	*Zygosaccharomyces bisporus*	YES	NO
1510	*Pseudozyma tsukubaensis*	YES	NO
1529	*Saccharomyces cerevisiae*	NO	YES
1603	*Saccharomyces cerevisiae*	YES	YES
1606	*Saccharomyces cerevisiae*	YES	NO
1645	*Rhodotorula mucilaginosa*	YES	NO
1646	*Rhodotorula mucilaginosa*	YES	NO
1647	*Rhodotorula mucilaginosa*	YES	NO
1649	*Rhodotorula mucilaginosa*	YES	NO
1650	*Rhodotorula mucilaginosa*	YES	NO
1651	*Rhodotorula mucilaginosa*	YES	NO
1659	*Rhodotorula mucilaginosa*	YES	NO
1660	*Rhodotorula mucilaginosa*	YES	NO
1681	*Saccharomyces cerevisiae*	NO	YES
2265	*Kluyveromyces marxianus*	YES	NO
2321	*Metschnikowia pulcherrima*	YES	NO
2322	*Metschnikowia pulcherrima*	YES	NO
2395	*Metschnikowia hawaiiensis*	YES	NO
2396	*Metschnikowia hawaiiensis*	YES	NO
2397	*Saccharomyces cerevisiae*	YES	YES
2401	*Saccharomyces cerevisiae*	NO	YES

2403	*Zygosaccharomyces mellis (20DEG!)*	YES	NO
2433	*Lachancea thermotolerans*	YES	NO
2439	*Rhodotorula glutinis*	YES	NO
2440	*Rhodotorula glutinis*	YES	NO
2449	*Kazachstania telluris*	YES	NO
2450	*Candida humilis*	YES	NO
2473	*Candida colliculosa*	YES	NO
2480	*Metschnikowia agaves*	YES	NO
2483	*Kluyveromyces piceae*	YES	NO
2486	*Metschnikowia agaves*	YES	NO
2489	*Zygotorulaspora mrakii*	YES	NO
2491	*Metschnikowia gruessii*	YES	NO
2508	*Lachancea fermentati*	YES	NO
2513	*Zygotorulaspora florentinus*	YES	NO
2517	*Saccharomyces cerevisiae*	NO	YES
2521	*Metschnikowia bicuspidata*	YES	NO
2529	*Metschnikowia bicuspidata*	YES	NO
2559	*Kluyveromyces dobzhanskii*	YES	NO
2560	*Kluyveromyces sinensis*	YES	NO
2568	*Zygosaccharomyces microellipsoides*	YES	NO
2572	*Debaryomyces hansenii var. hansenii*	YES	NO
2577	*Kazachstania servazzii*	YES	NO
2578	*Saccharomyces bayanus*	YES	NO
2580	*Metschnikowia pulcherrima*	YES	NO
2581	*Rhodotorula minuta var. minuta*	YES	NO
2592	*Saccharomyces cerevisiae*	YES	YES
2597	*Kluyveromyces marxianus var. marxianus*	YES	NO
2599	*Sporobolomyces albo-rubescens*	YES	NO
2600	*Saccharomyces paradoxus*	YES	NO
2605	*Rhodotorula vanillica*	YES	NO
2629	*Torulaspora delbrueckii*	YES	NO

2644	*Kluyveromyces waltii*	YES	NO
2666	*Rhodotorula glutinis var glutinis*	YES	NO
2675	*Kluyveromyces marxianus*	YES	NO
2688	*Saccharomyces cerevisiae*	NO	YES
2693	*Saccharomyces servazzii*	YES	NO
2701	*Kazachstania viticola*	YES	NO
2702	*Kazachstania kunashirensis*	YES	NO
2729	*Kluyveromyces africanus*	YES	NO
2733	*Saccharomyces cerevisiae*	YES	YES
2737	*Saccharomyces cerevisiae*	YES	YES
2739	*Hanseniaspora uvarum*	YES	NO
2741	*Torulaspora delbrueckii*	YES	NO
2742	*Kluyveromyces lactis*	YES	NO
2752	*Rhodotorula cresolica*	YES	NO
2753	*Metschnikowia zobellii*	YES	NO
2754	*Kluyveromyces yarrowii*	YES	NO
2775	*Saccharomyces servazzii*	YES	NO
2789	*Zygosaccharomyces lentus*	YES	NO
2790	*Zygosaccharomyces bailii*	YES	NO
2791	*Kluyveromyces marxianus*	YES	NO
2797	*Kluyveromyces lactis*	YES	NO
2804	*Saccharomyces bayanus/pastorianus*	YES	NO
2808	*Saccharomyces bayanus/pastorianus*	YES	NO
2809	*Saccharomyces bayanus/pastorianus*	YES	NO
2826	*Saccharomyces cerevisiae*	NO	YES
2827	*Saccharomyces rosinii*	YES	NO
2855	*Saccharomyces cerevisiae*	NO	YES
2864	*Rhodotorula mucilaginosa*	YES	NO
2875	*Lachancea cidri*	YES	NO
2878	*Saccharomyces barnettii*	YES	NO
2885	*Torulaspora delbrueckii*	YES	NO
2886	*Kluyveromyces marxianus*	YES	NO

2887	*Kluyveromyces marxianus*	YES	NO
2888	*Saccharomyces mikatae*	YES	NO
2889	*Saccharomyces kudriavzevii*	YES	NO
2890	*Saccharomyces cariocanus*	YES	NO
2897	*Zygosaccharomyces kombuchaensis*	YES	NO
2898	*Naumovozyma castellii*	YES	NO
2904	*Yarrowia lipolytica*	YES	NO
2907	*Kluyveromyces marxianus*	YES	NO
2908	*Starmerella bombicola*	YES	NO
2927	*Zygosaccharomyces bailii*	YES	NO
2931	*Zygosaccharomyces bailii*	YES	NO
2932	*Zygosaccharomyces bailii*	YES	NO
2933	*Zygosaccharomyces bailii*	YES	NO
2934	*Zygosaccharomyces bailii*	YES	NO
2935	*Zygosaccharomyces bisporus*	YES	NO
2945	*Saccharomyces cerevisiae*	YES	YES
2947	*Saccharomyces cerevisiae/paradoxus*	NO	YES
2948	*Saccharomyces cerevisiae/paradoxus*	YES	YES
2956	*Kluyveromyces lactis*	YES	NO
2976	*Hanseniaspora osmophila*	YES	NO
2980	*Kluyveromyces lactis var. lactis*	YES	NO
2981	*Kluyveromyces lactis var. drosophilarum*	YES	NO
2991	*Saccharomyces spencerorum*	YES	NO
2995	*Zygosaccharomyces bailii*	YES	NO
2999	*Zygosaccharomyces kombuchaensis*	YES	NO
3000	*Zygosaccharomyces kombuchaensis*	YES	NO
3001	*Zygosaccharomyces kombuchaensis*	YES	NO
3024	*Zygosaccharomyces microellipsoides*	YES	NO
3025	*Saccharomyces cerevisiae*	NO	YES
3026	*Saccharomyces cerevisiae*	NO	YES
3028	*Saccharomyces cerevisiae*	NO	YES

3030	*Saccharomyces cerevisiae*	NO	YES
3031	*Saccharomyces cerevisiae*	NO	YES
3032	*Saccharomyces cerevisiae*	NO	YES
3033	*Saccharomyces cerevisiae*	NO	YES
3034	*Saccharomyces cerevisiae*	YES	NO
3035	*Saccharomyces cerevisiae*	NO	YES
3036	*Saccharomyces cerevisiae*	NO	YES
3037	*Saccharomyces cerevisiae*	NO	YES
3038	*Saccharomyces cerevisiae*	NO	YES
3039	*Saccharomyces cerevisiae*	NO	YES
3041	*Kluyveromyces lactis*	YES	NO
3047	*Metschnikowia pulcherrima*	YES	NO
3051	*Saccharomyces cerevisiae*	NO	YES
3052	*Saccharomyces cerevisiae*	NO	YES
3053	*Saccharomyces servazzii*	YES	NO
3056	*Rhodotorula sp. nov.*	YES	NO
3057	*Rhodotorula mucilaginosa*	YES	NO
3072	*Rhodotorula laryngis*	YES	NO
3076	*Saccharomyces cerevisiae*	NO	YES
3077	*Saccharomyces cerevisiae*	NO	YES
3078	*Saccharomyces cerevisiae*	NO	YES
3090	*Zygosaccharomyces bailii*	YES	NO
3091	*Zygosaccharomyces bailii*	YES	NO
3096	*Metschnikowia fructicola*	YES	NO
3104	*Candida pseudointermedia*	YES	NO
3108	*Naumovozyma castellii*	YES	NO
3114	*Saccharomyces cerevisiae*	NO	YES
3120	*Rhodotorula phylloplana*	YES	NO
3121	*Saccharomyces cerevisiae*	NO	YES
3122	*Saccharomyces cerevisiae*	NO	YES
3123	*Saccharomyces cerevisiae*	NO	YES
3124	*Saccharomyces cerevisiae*	NO	YES

3125	*Saccharomyces cerevisiae*	NO	YES
3126	*Saccharomyces cerevisiae*	NO	YES
3127	*Saccharomyces cerevisiae*	NO	YES
3141	*Torulaspora delbrueckii*	YES	NO
3239	*Torulaspora delbrueckii*	YES	NO
3255	*Torulaspora delbrueckii*	YES	NO
3264	*Saccharomyces cerevisiae*	YES	NO
3265	*Saccharomyces cerevisiae*	YES	YES
3266	*Saccharomyces cerevisiae*	YES	YES
3267	*Pseudozyma sp.*	YES	NO
3303	*Candida glabrata*	YES	NO
3311	*Saccharomyces cerevisiae*	YES	YES
3313	*Saccharomyces cerevisiae*	YES	YES
3314	*Saccharomyces cerevisiae*	YES	YES
3315	*Saccharomyces cerevisiae*	YES	YES
3318	*Saccharomyces cerevisiae*	YES	YES
3319	*Saccharomyces cerevisiae*	YES	YES
3324	*Saccharomyces cerevisiae*	NO	YES
3325	*Saccharomyces cerevisiae*	NO	YES
3326	*Saccharomyces cerevisiae*	NO	YES
3331	*Saccharomyces cerevisiae*	NO	YES
3333	*Saccharomyces cerevisiae*	NO	YES
3334	*Saccharomyces cerevisiae*	NO	YES
3338	*Saccharomyces cerevisiae*	NO	YES
3339	*Saccharomyces cerevisiae*	NO	YES
3344	*Kluyveromyces marxianus*	YES	NO
3396	*Kluyveromyces marxianus*	YES	NO
3398	*Metschnikowia aff. fructicola*	YES	NO
3400	*Metschnikowia sp. nov.*	YES	NO
3401	*Rhodotorula graminis*	YES	NO
3406	*Saccharomyces cerevisiae*	YES	YES
3411	*Rhodotorula mucilaginosa*	YES	NO

3431	*Pseudozyma hubeiensis*	YES	NO
3444	*Rhodotorula dairenensis*	YES	NO
3445	*Saccharomyces cerevisiae*	YES	YES
3447	*Saccharomyces cerevisiae*	YES	YES
3448	*Saccharomyces cerevisiae*	YES	YES
3449	*Saccharomyces cerevisiae*	YES	YES
3451	*Saccharomyces cerevisiae*	YES	NO
3452	*Saccharomyces cerevisiae*	NO	YES
3453	*Saccharomyces cerevisiae*	YES	NO
3454	*Saccharomyces cerevisiae*	YES	NO
3455	*Saccharomyces cerevisiae*	YES	YES
3456	*Saccharomyces cerevisiae*	YES	YES
3457	*Saccharomyces cerevisiae*	YES	YES
3458	*Saccharomyces cerevisiae*	YES	YES
3460	*Saccharomyces cerevisiae*	YES	YES
3461	*Saccharomyces cerevisiae*	YES	YES
3462	*Saccharomyces cerevisiae*	YES	YES
3466	*Saccharomyces cerevisiae*	YES	NO
3467	*Saccharomyces cerevisiae*	YES	YES
3469	*Saccharomyces cerevisiae*	YES	NO
3470	*Saccharomyces cerevisiae*	YES	YES
3471	*Saccharomyces cerevisiae*	YES	YES
3472	*Saccharomyces cerevisiae*	YES	YES
3486	*Saccharomyces cerevisiae*	YES	YES
3487	*Saccharomyces cerevisiae*	YES	YES
3502	*Candida glabrata*	YES	NO
3504	*Rhodotorula mucilaginosa*	YES	NO
3506	*Torulaspora delbrueckii*	YES	NO
3519	*Candida glabrata*	YES	NO
3536	*Rhodotorula mucilaginosa*	YES	NO
3537	*Candida glabrata*	YES	NO
3612	*Saccharomyces cerevisiae*	YES	YES

3630	*Saccharomyces cerevisiae*	YES	YES
3719	*Metschnikowia sp.*	YES	NO
3721	*Rhodotorula slooffiae*	YES	NO
3722	*Rhodotorula graminis*	YES	NO
3725	*Rhodotorula glutinis var. dairenensis*	YES	NO
3735	*Rhodotorula mucilaginosa*	YES	NO
3772	*Rhodotorula mucilaginosa*	YES	NO
3775	*Rhodotorula mucilaginosa*	YES	NO
3788	*Hanseniaspora guilliermondii*	YES	NO
3792	*Metschnikowia koreensis*	YES	NO
3816	*Rhodotorula mucilaginosa*	YES	NO
3817	*Rhodotorula mucilaginosa*	YES	NO
3820	*Rhodotorula mucilaginosa*	YES	NO
3821	*Rhodotorula mucilaginosa*	YES	NO
3832	*Rhodotorula sp. nov.*	YES	NO
3833	*Rhodotorula sp. nov.*	YES	NO
3834	*Rhodotorula laryngis*	YES	NO
3835	*Rhodotorula sp. nov.*	YES	NO
3836	*Rhodotorula laryngis*	YES	NO
3837	*Rhodotorula laryngis*	YES	NO
3838	*Rhodotorula laryngis*	YES	NO
3853	*Kazachstania bulderi*	YES	NO
3867	*Rhodotorula mucilaginosa*	YES	NO
3872	*Rhodotorula mucilaginosa*	YES	NO
4000	*Kazachstania yasuniensis*	YES	NO
3455	*Saccharomyces cerevisiae*	NO	YES

Table 3: ***Yeast strains included within the metabolomics studies***
The NCYC strain designation is on the first column, species on the second, with presence (YES/NO) in Malt and YNB media in third and fourth columns, respectively.

Bibliography

[1] Nueno-Palop, Carmen et al (2021), 'National Collection of Yeast Cultures |Quadram Institute Bioscience'.

[2] Waldron, Keith W. (2010), *Bioalcohol Production : Biochemical Conversion Of Lignocellulosic Biomass.*, Elsevier Science.

[3] Geyer, Roland et al (2017), 'Production, use, and fate of all plastics ever made', *Sci. Adv.*

[4] Żymańczyk-Duda, Ewa et al (2017), Yeast as a Versatile Tool in Biotechnology, *in* 'Yeast - Ind. Appl.', InTech.

[5] Sanchez-Garcia, Laura et al (2016), 'Recombinant pharmaceuticals from microbial cells: a 2015 update.', *Microb. Cell Fact.* **15**, 33.

[6] Karathia, Hiren et al (2011), 'Saccharomyces cerevisiae as a model organism: a comparative study.', *PLoS One* **6**(2), e16015.

[7] Wang, Hanyu et al (2017), 'YKL071W from Saccharomyces cerevisiae encodes a novel aldehyde reductase for detoxification of glycolaldehyde and furfural derived from lignocellulose', *Appl. Microbiol. Biotechnol.* **101**(23-24), 8405–8418.

[8] Hagemann, Ian S. (2015), 'Overview of Technical Aspects and Chemistries of Next-Generation Sequencing', *Clin. Genomics* pp. 3–19.

[9] Wetterstrand, Kris A. (2021), 'The Cost of Sequencing a Human Genome'.

[10] Kono, Nobuaki & Arakawa, Kazuharu (2019), 'Nanopore sequencing: Review of potential applications in functional genomics', *Dev. Growth Differ.* **61**(5), 316–326.

[11] Wu, Yishuo et al (2018), 'Genome-wide Association Study (GWAS) of germline copy number variations (CNVs) reveal genetic risks of prostate cancer in Chinese population', *J. Cancer* **9**(5), 923–928.

[12] Chang, Michelle et al (2018), An overview of genome-wide association studies, *in* 'Methods Mol. Biol.', Vol. 1754, Humana Press Inc., pp. 97–108.

[13] Kryshtafovych, Andriy et al (2019), 'Critical assessment of methods of protein structure prediction (CASP)-Round XIII', *Proteins* **87**(12), 1011–1020.

[14] Consortium, The Gene Ontology et al (2000), 'Gene Ontology: tool for the unification of biology', *Nat. Genet.* **25**(1), 25.

[15] McGinnis, Scott & Madden, Thomas L (2004), 'BLAST: at the core of a powerful and diverse set of sequence analysis tools.', *Nucleic Acids Res.* **32**(Web Server issue), W20–5.

[16] Tsirigos, Konstantinos D et al (2016), 'PRED-TMBB2: improved topology prediction and detection of beta-barrel outer membrane proteins', *Bioinformatics* **32**(17), i665–i671.

[17] Hulo, Nicolas et al (2006), 'The PROSITE database', *Nucleic Acids Res.* **34**(Database issue), D227.

[18] I, Letunic & P, Bork (2021), 'Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation', *Nucleic Acids Res.* **49**(W1), W293–W296.

[19] Cooper, Geoffrey M (2000), *The Cell: A Molecular Approach. 2nd edition. Sunderland* , 2 edn, Sinauer Associates.

[20] Garcia-Albornoz, M. et al (2020), 'A proteome-integrated, carbon source dependent genetic regulatory network in: Saccharomyces cerevisiae', *Mol. Omi.* **16**(1), 59–72.

[21] Neiman, Aaron M. (2005), 'Ascospore Formation in the Yeast Saccharomyces cerevisiae', *Microbiol. Mol. Biol. Rev.* **69**(4), 565–584.

[22] Peter, Jackson et al (2018), 'Genome evolution across 1,011 Saccharomyces cerevisiae isolates', *Nature* **556**(7701), 339–344.

[23] Tretter, Laszlo et al (2016), 'Succinate, an intermediate in metabolism, signal transduction, ROS, hypoxia, and tumorigenesis', *Biochim. Biophys. Acta - Bioenerg.* **1857**(8), 1086–1101.

[24] Narancic, Tanja et al (2020), 'Recent advances in bioplastics: Application and biodegradation', *Polymers (Basel).*

[25] Lee, Yong Jae & Jeong, Ki Jun (2015), 'Challenges to production of antibodies in bacteria and yeast', *J. Biosci. Bioeng.* **120**(5), 483–490.

[26] Connors, Jessica et al (2019), 'The role of succinate in the regulation of intestinal inflammation', *Nutrients*.

[27] Maicas, Sergi (2020), 'The role of yeasts in fermentation processes', *Microorganisms* **8**(8), 1–8.

[28] Parapouli, Maria et al (2020), 'Saccharomyces cerevisiae and its industrial applications', *AIMS Microbiol.* **6**(1), 1–31.

[29] Beopoulos, Athanasios & Nicaud, Jean-Marc (2012), 'Yeast: A new oil producer?', *Oléagineux, Corps gras, Lipides* **19**(1), 22–28.

[30] Mussatto, Solange I. et al (2010), 'Technological trends, global market, and challenges of bio-ethanol production', *Biotechnol. Adv.* **28**(6), 817–830.

[31] Wang, Qing et al (2014), 'Pretreating lignocellulosic biomass by the concentrated phosphoric acid plus hydrogen peroxide (PHP) for enzymatic hydrolysis: Evaluating the pretreatment flexibility on feedstocks and particle sizes', *Bioresour. Technol.* **166**, 420–428.

[32] Banerjee, Nirupama et al (1981), 'Inhibition of glycolysis by furfural in Saccharomyces cerevisiae', *Eur. J. Appl. Microbiol. Biotechnol.* **11**(4), 226–228.

[33] Alabduladhem, Tamim O. & Bordoni, Bruno (2020), *Physiology, Krebs Cycle*, StatPearls Publishing.

[34] Moriya, Hisao (2015), 'Quantitative nature of overexpression experiments', *Mol. Biol. Cell* **26**(22), 3932–3939.

[35] Palmqvist, Eva et al (1999), 'Influence of furfural on anaerobic glycolytic kinetics of saccharomyces cerevisiae in batch culture', *Biotechnol. Bioeng.* **62**(4), 447–454.

[36] Horváth, Ilona Sárvári et al (2001), 'Effects of furfural on anaerobic continuous cultivation of Saccharomyces cerevisiae', *Biotechnol. Bioeng.* **75**(5), 540–549.

[37] Imura, Makoto et al (2018), 'Metabolomics approach to reduce the Crabtree effect in continuous culture of Saccharomyces cerevisiae', *J. Biosci. Bioeng.* **126**(2), 183–188.

[38] Kelly, David J. et al (2001), 'Microaerobic Physiology: Aerobic Respiration, Anaerobic Respiration, and Carbon Dioxide Metabolism', *Helicobacter pylori* pp. 111–124.

[39] Lewis Liu, Z. et al (2008), 'Multiple gene-mediated NAD(P)H-dependent aldehyde reduction is a mechanism of in situ detoxification of furfural and 5-hydroxymethylfurfural by Saccharomyces cerevisiae', *Appl. Microbiol. Biotechnol.* **81**(4), 743–753.

[40] Yang, Bin & Wyman, Charles E. (2008), 'Pretreatment: The key to unlocking low-cost cellulosic ethanol', *Biofuels, Bioprod. Biorefining* **2**(1), 26–40.

[41] Irwin, Scott (2017), 'The Profitability of Ethanol Production in 2016', *Farmdoc Dly.*

[42] Elliston, Adam et al (2015), 'Effect of steam explosion on waste copier paper alone and in a mixed lignocellulosic substrate on saccharification and fermentation', *Bioresour. Technol.* **187**, 136–143.

[43] Nevin, Kelly P. et al (2010), 'Microbial Electrosynthesis: Feeding Microbes Electricity To Convert Carbon Dioxide and Water to Multicarbon Extracellular Organic Compounds', *MBio*.

[44] McCarthy, Charley G.P. & Fitzpatrick, David A. (2019), 'Pan-genome analyses of model fungal species', *Microb. Genomics*.

[45] Heitman, Joseph et al (2020), 'Advances in understanding the evolution of fungal genome architecture', *F1000Research*.

[46] Gilchrist, Ciaran & Stelkens, Rike (2019), 'Aneuploidy in yeast: Segregation error or adaptation mechanism?'.

[47] Large, Christopher et al (2020), 'Genomic stability and adaptation of beer brewing yeasts during serial repitching in the brewery', *bioRxiv* p. 2020.06.26.166157.

[48] Fay, Justin C. et al (2019), 'A polyploid admixed origin of beer yeasts derived from European and Asian wine populations', *PLOS Biol.* **17**(3), e3000147.

[49] Gerstein, Aleeza C. et al (2006), 'Genomic convergence toward diploidy in Saccharomyces cerevisiae', *PLoS Genet.* **2**(9), 1396–1401.

[50] Nishant, K T et al (2010), 'The Baker's Yeast Diploid Genome Is Remarkably Stable in Vegetative Growth and Meiosis', *PLoS Genet* **6**(9), 1001109.

[51] Selmecki, Anna M. et al (2015), 'Polyploidy can drive rapid adaptation in yeast', *Nature* **519**(7543), 349–351.

[52] Scott, Amber L. et al (2017), 'The Influence of Polyploidy on the Evolution of Yeast Grown in a Sub-Optimal Carbon Source', *Mol. Biol. Evol.* **34**(10), 2690–2703.

[53] Richardson, David J. (2000), 'Bacterial respiration: a flexible process for a changing environment1999 Fleming Lecture (Delivered at the 144th meeting of the Society for General Microbiology, 8 September 1999)', *Microbiology* **146**(3), 551–571.

[54] Vetriani, Costantino et al (2005), 'Mercury adaptation among bacteria from a deep-sea hydrothermal vent', *Appl. Environ. Microbiol.* **71**(1), 220–226.

[55] Vogler, K. G. (1942), 'STUDIES ON THE METABOLISM OF AUTOTROPHIC BACTERIA : II. THE NATURE OF THE CHEMOSYNTHETIC REACTION', *J. Gen. Physiol.* **26**(1), 103.

[56] Edwards, Marcus J. et al (2020), 'The Crystal Structure of a Biological Insulated Transmembrane Molecular Wire', *Cell* **181**(3), 665–673.e10.

[57] Lockwood, Colin et al (2018), 'Membrane-Spanning Electron Transfer Proteins from Electrogenic Bacteria: Production and Investigation', *Methods Enzymol.* **613**, 257–275.

[58] Land, Miriam et al (2015), 'Insights from 20 years of bacterial genome sequencing', *Funct. Integr. genomics* **15**(2), 141–161.

[59] Bobay, Louis-Marie & Ochman, Oward (2017), 'The Evolution of Bacterial Genome Architecture', *Front. Genet.*

[60] DiCenzo, George C. & Finan, Turlough M. (2017), 'The Divided Bacterial Genome: Structure, Function, and Evolution', *Microbiol. Mol. Biol. Rev.*

[61] S, Ballouz et al (2010), 'Conditions for the evolution of gene clusters in bacterial genomes', *PLoS Comput. Biol.*

[62] Fang, Gang et al (2008), 'Persistence drives gene clustering in bacterial genomes', *BMC Genomics* **9**, 4.

[63] D, Cook & L, Sequeira (1991), 'Genetic and biochemical characterization of a Pseudomonas solanacearum gene cluster required for extracellular polysaccharide production and for virulence', *J. Bacteriol.* **173**(5), 1654–1662.

[64] O'Leary, Nuala A. et al (2016), 'Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation', *Nucleic Acids Res.* **44**(D1), D733–D745.

[65] Li, H. et al (2009), 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics* **25**(16), 2078–2079.

[66] Sandmann, Sarah et al (2017), 'Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data', *Sci. Rep.*

[67] Li, Heng & Durbin, Richard (2009), 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics* **25**(14), 1754–1760.

[68] et al, Van Rossum (1995), 'Python reference manual'.

[69] RStudio Team (2020) (2015), 'RStudio: Integrated Development for R. RStudio, PBC, Boston'.

[70] Fuente-Hernández, Ariadna et al (2017), 'Reduction of furfural to furfuryl alcohol in liquid phase over a biochar-supported platinum catalyst', *Energies* **10**(3), 286.

[71] Wahlbom, C Fredrik & Hahn-Hägerdal, Bärbel (2002), 'Furfural, 5-hydroxymethyl furfural, and acetoin act as external electron acceptors during anaerobic fermentation of xylose in recombinant Saccharomyces cerevisiae.', *Biotechnol. Bioeng.* **78**(2), 172–8.

[72] Liu, Lu et al (2018), 'Furfural production from biomass pretreatment hydrolysate using vapor-releasing reactor system', *Bioresour. Technol.* **252**, 165–171.

[73] Ran, Hong et al (2014), 'Analysis of biodegradation performance of furfural and 5- hydroxymethylfurfural by Amorphotheca resinae ZN1', *Biotechnol. Biofuels* **7**(1), 51.

[74] Field, Sarah J. et al (2015), 'Identification of furfural resistant strains of Saccharomyces cerevisiae and Saccharomyces paradoxus from a collection of environmental and industrial isolates', *Biotechnol. Biofuels* **8**(1), 33–33.

[75] Slatkin, Montgomery (2008), 'Linkage disequilibrium–understanding the evolutionary past and mapping the medical future.', *Nat. Rev. Genet.* **9**(6), 477–85.

[76] Bush, William S. & Moore, Jason H. (2012), 'Chapter 11: Genome-Wide Association Studies', *PLoS Comput. Biol.*

[77] Popescu, Andrei Alin et al (2014), 'A novel and fast approach for population structure inference using Kernel-PCA and optimization', *Genetics* **198**(4), 1421–1431.

[78] Jolliffe, Ian T. & Cadima, Jorge (2016), 'Principal component analysis: a review and recent developments', *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **374**(2065), 20150202.

[79] Kristensen, M O et al (1993), 'Serum cobalamin and methylmalonic acid in Alzheimer dementia.', *Acta Neurol. Scand.* **87**(6), 475–81.

[80] Alexander, David H et al (2009), 'Fast model-based estimation of ancestry in unrelated individuals.', *Genome Res.* **19**(9), 1655–64.

[81] Otero, José Manuel et al (2013), 'Industrial Systems Biology of Saccharomyces cerevisiae Enables Novel Succinic Acid Cell Factory', *PLoS One* **8**(1), e54144.

[82] Sharmaa, Aditi et al (2015), 'Stories and challenges of genome wide association studies in livestock - a review', *Asian-Australasian J. Anim. Sci.* **28**(10), 1371–1379.

[83] Basler, Georg et al (2018), 'Advances in metabolic flux analysis toward genome-scale profiling of higher organisms', *Biosci. Rep.* **38**(6), 20170224.

[84] Popescu, Andrei-Alin et al (2014), 'A Novel and Fast Approach for Population Structure Inference Using Kernel-PCA and Optimization', *Genetics* **198**(4), 1421–1431.

[85] Gagnaire, Pierre Alexandre et al (2015), 'Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era'.

[86] Harper, Andrea L. et al (2020), 'Validation of an Associative Transcriptomics platform in the polyploid crop species Brassica juncea by dissection of the genetic architecture of agronomic and quality traits', *Plant J.* **103**(5), 1885–1893.

[87] Bolger, Anthony M. et al (2014), 'Trimmo1. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170matic: A flexible trimmer for Illumina sequence data', *Bioinformatics* **30**(15), 2114–2120.

[88] Bushnell, Brian et al (2017), 'BBMerge – Accurate paired shotgun read merging via overlap', *PLoS One*.

[89] Emwas, Abdul Hamid M. (2015), 'The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research', *Methods Mol. Biol.* **1277**, 161–193.

[90] Bušić, Arijana et al (2018), 'Bioethanol production from renewable raw materials and its separation and purification: A review'.

[91] Medeiros Garcia Alcântara, João et al (2020), 'Current trends in the production of biodegradable bioplastics: The case of polyhydroxyalkanoates', *Biotechnol. Adv.* **42**, 107582.

[92] Hannon, Michael et al (2010), 'Biofuels from algae: challenges and potential.', *Biofuels* **1**(5), 763–784.

[93] Aro, Eva-Mari (2016), 'From first generation biofuels to advanced solar biofuels.', *Ambio* **45 Suppl 1**(Suppl 1), S24–31.

[94] Tenenbaum, David J (2008), 'Food vs. fuel: diversion of crops could cause more hunger.', *Environ. Health Perspect.* **116**(6), A254–7.

[95] Kohli, Kirtika et al (2019), 'Bio-based chemicals from renewable biomass for integrated biorefineries', *Energies*.

[96] Kucharska, Karolina et al (2018), 'Pretreatment of Lignocellulosic Materials as Substrates for Fermentation Processes.', *Molecules*.

[97] Jönsson, Leif J. & Martín, Carlos (2016), 'Pretreatment of lignocellulose: Formation of inhibitory by-products and strategies for minimizing their effects', *Bioresour. Technol.* **199**, 103–112.

[98] Godinho, Cláudia P. et al (2017), 'Yeast response and tolerance to benzoic acid involves the Gcn4- and Stp1-regulated multidrug/multixenobiotic resistance transporter Tpo1', *Appl. Microbiol. Biotechnol.* **101**(12), 5005–5018.

[99] Hope, C. F.A. (1987), 'CINNAMIC ACID AS THE BASIS OF A MEDIUM FOR THE DETECTION OF WILD YEASTS', *J. Inst. Brew.* **93**(3), 213–215.

[100] Fletcher, Eugene et al (2019), 'Yeast chemogenomic screen identifies distinct metabolic pathways required to tolerate exposure to phenolic fermentation inhibitors ferulic acid, 4-hydroxybenzoic acid and coniferyl aldehyde', *Metab. Eng.* **52**, 98–109.

[101] Field, Sarah J et al (2015), 'Identification of furfural resistant strains of Saccharomyces cerevisiae and Saccharomyces paradoxus from a collection of environmental and industrial isolates', *Biotechnol. Biofuels* **8**(1), 33.

[102] Heer, Dominik et al (2009), 'Resistance of Saccharomyces cerevisiae to high concentrations of furfural is based on NADPH-dependent reduction by at least two oxireductases', *Appl. Environ. Microbiol.* **75**(24), 7631–7638.

[103] Lin, Feng-Ming et al (2009), 'Comparative proteomic analysis of tolerance and adaptation of ethanologenic Saccharomyces cerevisiae to furfural, a lignocellulosic inhibitory compound.', *Appl. Environ. Microbiol.* **75**(11), 3765–76.

[104] Allen, Sandra A et al (2010), 'Furfural induces reactive oxygen species accumulation and cellular damage in Saccharomyces cerevisiae', *Biotechnol. Biofuels* **3**(1), 2.

[105] Taherzadeh, Mohammad J. et al (1999), 'Conversion of furfural in aerobic and anaerobic batch fermentation of glucose by Saccharomyces cerevisiae', *J. Biosci. Bioeng.* **87**(2), 169–174.

[106] Gorsich, S. W. et al (2006), 'Tolerance to furfural-induced stress is associated with pentose phosphate pathway genes ZWF1, GND1, RPE1, and TKL1 in Saccharomyces cerevisiae', *Appl. Microbiol. Biotechnol.* **71**(3), 339–349.

[107] Boyer, L.J. et al (1992), 'The effects of furfural on ethanol production by saccharomyces cereyisiae in batch culture', *Biomass and Bioenergy* **3**(1), 41–48.

[108] Steinley, Douglas. (2006), 'K-means clustering: A half-century synthesis', *Br. J. Math. Stat. Psychol.* **59**(1), 1–34.

[109] Tibshirani, Robert et al (2001), 'Estimating the number of clusters in a data set via the gap statistic', *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **63**(2), 411–423.

[110] Tamura, K & Nei, M (1993), 'Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.', *Mol. Biol. Evol.* **10**(3), 512–526.

[111] Ryden, Peter et al (2017), 'Bioethanol production from spent mushroom compost derived from chaff of millet and sorghum', *Biotechnol. Biofuels* **10**(1), 195.

[112] Wu, Jia et al (2017), 'Yeast diversity in relation to the production of fuels and chemicals', *Sci. Rep.* **7**(1), 14259–14259.

[113] Elliston, Adam et al (2013), 'High concentrations of cellulosic ethanol achieved by fed batch semi simultaneous saccharification and fermentation of waste-paper', *Bioresour. Technol.* **134**, 117–126.

[114] Van Ommen, G. J.B. et al (1979), 'Transcription maps of mtdnas of two strains of saccharomyces: transcription of strain-specific insertions; complex rna maturation and splicing', *Cell* **18**(2), 511–523.

[115] Tropea, Alessia et al (2016), 'Development of minimal fermentation media supplementation for ethanol production using two Saccharomyces cerevisiae strains', *Nat. Prod. Res.* **30**(9), 1009–1016.

[116] Bokulich, Nicholas A. & Bamforth, Charles W. (2013), 'The Microbiology of Malting and Brewing', *Microbiol. Mol. Biol. Rev.* **77**(2), 157.

[117] Engel, Stacia R. et al (2014), 'The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now', *G3 Genes, Genomes, Genet.* **4**(3), 389–398.

[118] Diniz-Filho, José Alexandre F. et al (2013), 'Mantel test in population genetics'.

[119] Onishi, Hiroshi et al (2018), 'A genome-wide association study identifies three novel genetic markers for response to tamoxifen: A prospective multicenter study', *PLoS One*.

[120] Verna, James et al (1997), 'A family of genes required for maintenance of cell wall integrity and for the stress response in Saccharomyces cerevisiae', *Proc. Natl. Acad. Sci. U. S. A.* **94**(25), 13804.

[121] Anderson, James B et al (2003), 'Mode of selection and experimental evolution of antifungal drug resistance in Saccharomyces cerevisiae.', *Genetics* **163**(4), 1287.

[122] Alvaro, David et al (2007), 'Genome-Wide Analysis of Rad52 Foci Reveals Diverse Mechanisms Impacting Recombination', *PLoS Genet.* **3**(12), 2439–2449.

[123] Nekrasov, Vladimir S. et al (2003), 'Interactions between Centromere Complexes in Saccharomyces cerevisiae', *Mol. Biol. Cell* **14**(12), 4931.

[124] Loewith, Robbie et al (2002), 'Two TOR complexes, only one of which is rapamycin sensitive, have distinct roles in cell growth control', *Mol. Cell* **10**(3), 457–468.

[125] Lin, Coney Pei-Chen et al (2013), 'A Highly Redundant Gene Network Controls Assembly of the Outer Spore Wall in S. cerevisiae', *PLoS Genet.*

[126] Enyenihi, Akon H & Saunders, William S (2003), 'Large-scale functional genomic analysis of sporulation and meiosis in Saccharomyces cerevisiae.', *Genetics* **163**(1), 47.

[127] Bolesm, Eckhard et al (1996), 'Cloning of a second gene encoding 6-phosphofructo-2-kinase in yeast, and characterization of mutant strains without fructose-2,6-bisphosphate', *Mol. Microbiol.* **20**(1), 65–76.

[128] Minard, K I & McAlister-Henn, L (1991), 'Isolation, nucleotide sequence analysis, and disruption of the MDH2 gene from Saccharomyces cerevisiae: evidence for three isozymes of yeast malate dehydrogenase.', *Mol. Cell. Biol.* **11**(1), 370.

[129] Wang, Hanyu et al (2017), 'YKL071W from Saccharomyces cerevisiae encodes a novel aldehyde reductase for detoxification of glycolaldehyde and furfural derived from lignocellulose', *Appl. Microbiol. Biotechnol.* **101**(23-24), 8405–8418.

[130] Zumstein, E et al (1995), 'A 29.425 kb segment on the left arm of yeast chromosome XV contains more than twice as many unknown as known open reading frames', *Yeast* **11**(10), 975–986.

[131] VALENS, MICHELE et al (1998), 'The Sequence of a 54·7 kb Fragment of Yeast Chromosome XV Reveals the Presence of Two tRNAs and 24 New Open Reading Frames - VALENS - 1997 - Yeast - Wiley Online Library'.

[132] Shi, Xunwun et al (2016), 'Large-Scale Deletion of Non-Essential Genes Region of Chromosome XV in Saccharomyces Cerevisiae', *IOSR J. Pharm. Biol. Sci. (IOSR-JPBS* **11**(3), 46–50.

[133] Blagosklonny, Mikhail V. (2019), 'Rapamycin for longevity: opinion article', *Aging (Albany NY)* **11**(19), 8048.

[134] Bušić, Arijana et al (2018), 'Bioethanol production from renewable raw materials and its separation and purification: A review'.

[135] Wi, Seung Gon et al (2013), 'Bioethanol production from rice straw by popping pretreatment', *Biotechnol. Biofuels* **6**(1), 166.

[136] Cao, Yujin et al (2013), 'Fermentative Succinate Production: An Emerging Technology to Replace the Traditional Petrochemical Processes', *Biomed Res. Int.* **2013**, 1–12.

[137] McKinlay, James B. et al (2007), 'Prospects for a bio-based succinate industry', *Appl. Microbiol. Biotechnol.* **76**(4), 727–740.

[138] Louis F Hartman and A. L. Oppenheim (1950), *On Beer and Brewing Techniques in Ancient Mesopotamia*, Journal of the American Oriental Society.

[139] Gallone, Brigida et al (2016), 'Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts.', *Cell* **166**(6), 1397–1410.e16.

[140] Sicard, Delphine & Legras, Jean Luc (2011), 'Bread, beer and wine: Yeast domestication in the Saccharomyces sensu stricto complex', *Comptes Rendus - Biol.* **334**(3), 229–236.

[141] Jeong, Ki Jun et al (2011), 'Recombinant antibodies: Engineering and production in yeast and bacterial hosts', *Biotechnol. J.* **6**(1), 16–27.

[142] Ito, M. et al (2008), 'Antigen-specific antibody production of human B cells in NOG mice reconstituted with the human immune system'.

[143] Zhang, Richard Yi & Shen, Wenyan David (2012), 'Monoclonal antibody expression in mammalian cells', *Methods Mol. Biol.* **907**, 341–358.

[144] Roser, Hannah Ritchie & Max (2018), 'Alcohol Consumption', *Our World Data*.

[145] Shaw, Lucy (2019), ''Sober curious' movement gathering speed in US'.

[146] Cuello, Raúl Andrés et al (2017), 'Construction of low-ethanol–wine yeasts through partial deletion of the Saccharomyces cerevisiae PDC2 gene', *AMB Express* **7**(1), 67.

[147] Wu, Jia et al (2017), 'Yeast diversity in relation to the production of fuels and chemicals', *Sci. Reports 2017 71* **7**(1), 1–11.

[148] Badgujar, Kirtikumar C. & Bhanage, Bhalchandra M. (2018), 'Dedicated and Waste Feedstocks for Biorefinery: An Approach to Develop a Sustainable Society', *Waste Biorefinery Potential Perspect.* pp. 3–38.

[149] Shastri, Yogendra et al (2011), 'Impact of Weather on Biomass Feedstock Harvest System Operations and Cost', *Am. Soc. Agric. Biol. Eng. Annu. Int. Meet. 2011, ASABE 2011* **1**, 1–.

[150] Williams, C. Luke et al (2015), 'Sources of Biomass Feedstock Variability and the Potential Impact on Biofuels Production', *BioEnergy Res. 2015 91* **9**(1), 1–14.

[151] Cristóbal, Jorge et al (2018), 'Techno-economic and profitability analysis of food waste biorefineries at European level', *Bioresour. Technol.* **259**, 244–252.

[152] Quispe, César A.G. et al (2013), 'Glycerol: Production, consumption, prices, characterization and new trends in combustion', *Renew. Sustain. Energy Rev.* **27**, 475–493.

[153] Galafassi, Silvia et al (2012), 'Lipid production for second generation biodiesel by the oleaginous yeast Rhodotorula graminis', *Bioresour. Technol.* **111**, 398–403.

[154] G, Fia et al (2005), 'Study of beta-glucosidase production by wine-related yeasts during alcoholic fermentation. A new rapid fluorimetric method to determine enzymatic activity', *J. Appl. Microbiol.* **99**(3), 509–517.

[155] Segal, David S. & Stockwell, Tim (2009), 'Low alcohol alternatives: A promising strategy for reducing alcohol related harm', *Int. J. Drug Policy* **20**(2), 183–187.

[156] Belal, Elsayed B. (2013), 'Bioethanol production from rice straw residues', *Brazilian J. Microbiol.* **44**(1), 225–234.

[157] De Witt, R. N. et al (2019), 'QTL analysis of natural Saccharomyces cerevisiae isolates reveals unique alleles involved in lignocellulosic inhibitor tolerance', *FEMS Yeast Res.*

[158] Naumoff, D. G. & Naumov, G. I. (2010), 'Discovery of a novel family of α-glucosidase IMA genes in yeast Saccharomyces cerevisiae', *Dokl. Biochem. Biophys. 2010 4321* **432**(1), 114–116.

[159] Needleman, R. B. et al (1984), 'MAL6 of Saccharomyces: a complex genetic locus containing three genes required for maltose fermentation.', *Proc. Natl. Acad. Sci. U. S. A.* **81**(9), 2811.

[160] Wang, Danny L. & Reitz, Ronald C. (1983), 'Ethanol ingestion and polyunsaturated fatty acids: effects on the acyl-CoA desaturases', *Alcohol. Clin. Exp. Res.* **7**(2), 220–226.

[161] BoucheriÃ©, Helian et al (1995), 'Differential synthesis of glyceraldehyde-3-phosphate dehydrogenase polypeptides in stressed yeast cells', *FEMS Microbiol. Lett.* **125**(2-3), 127–133.

[162] National Center for Biotechnology Information (2021), 'PubChem Compound Summary for CID 160419, Succinate'.

[163] Huang, Shaobai & Millar, A. Harvey (2013), 'Succinate dehydrogenase: The complex roles of a simple enzyme'.

[164] Kratzer, Steffen & Schüller, Hans Joachim (1995), 'Carbon source-dependent regulation of the acetyl-coenzyme A synthetase-encoding gene ACS1 from Saccharomyces cerevisiae', *Gene* **161**(1), 75–79.

[165] Bakker, B. M. et al (2000), 'The mitochondrial alcohol dehydrogenase Adh3p is involved in a redox shuttle in Saccharomyces cerevisiae.', *J. Bacteriol.* **182**(17), 4730–4737.

[166] Jordan, Paulina et al (2016), 'Hxt13, Hxt15, Hxt16 and Hxt17 from Saccharomyces cerevisiae represent a novel type of polyol transporters', *Sci. Rep.*

[167] Liu, Jian Ming et al (2020), 'From Waste to Taste - Efficient Production of the Butter Aroma Compound Acetoin from Low-Value Dairy Side Streams Using a Natural (Nonengineered) Lactococcus lactis Dairy Isolate', *J. Agric. Food Chem.* **68**(21), 5891–5899.

[168] Allen, Joseph G. et al (2016), 'Flavoring chemicals in e-cigarettes: Diacetyl, 2,3-pentanedione, and acetoin in a sample of 51 products, including fruit-, candy-, and cocktail-flavored e-cigarettes', *Environ. Health Perspect.* **124**(6), 733–739.

[169] Gonzalez, E. et al (2000), 'Characterization of a (2R,3R)-2,3-butanediol dehydrogenase as the Saccharomyces cerevisiae YAL060W gene product. Disruption and induction of the gene', *J. Biol. Chem.* **275**(46), 35876–35885.

[170] White, G. F. et al (2016), 'Mechanisms of Bacterial Extracellular Electron Exchange', *Adv. Microb. Physiol.* **68**, 87–138.

[171] Tahernia, Mehdi et al (2020), 'Characterization of Electrogenic Gut Bacteria', *ACS Omega* **5**(45), 29439.

[172] Puig, Sebastià et al (2021), 'Editorial: Microbial Electrogenesis, Microbial Electrosynthesis, and Electro-bioremediation', *Front. Microbiol.* **12**, 742479.

[173] Williams, Kenneth H et al (2010), 'Electrodic voltages accompanying stimulated bioremediation of a uranium-contaminated aquifer', *J. Geophys. Res* **115**, 0–05.

[174] Szydlowski, Lukasz et al (2020), 'Metabolic engineering of a novel strain of electrogenic bacterium Arcobacter butzleri to create a platform for single analyte detection using a microbial fuel cell', *Enzyme Microb. Technol.*

[175] Breuer, Marian et al (2015), 'Multi-haem cytochromes in Shewanella oneidensis MR-1: Structures, functions and opportunities', *J. R. Soc. Interface*.

[176] Clarke, Thomas A. et al (2008), The role of multihaem cytochromes in the respiration of nitrite in Escherichia coli and Fe(III) in Shewanella oneidensis, *in* 'Biochem. Soc. Trans.', Vol. 36, Biochem Soc Trans, pp. 1005–1010.

[177] Holmes, Dawn E. et al (2004), 'Potential Role of a Novel Psychrotolerant Member of the Family Geobacteraceae, Geopsychrobacter electrodiphilus gen. nov., sp. nov., in Electricity Production by a Marine Sediment Fuel Cell', *Appl. Environ. Microbiol.* **70**(10), 6023.

[178] Garber, Arkadiy I. et al (2020), 'FeGenie: A Comprehensive Tool for the Identification of Iron Genes and Iron Gene Neighborhoods in Genome and Metagenome Assemblies', *Front. Microbiol.*

[179] Gagkaeva, Z. V. et al (2018), 'Terahertz-infrared spectroscopy of Shewanella oneidensis MR-1 extracellular matrix', *J. Biol. Phys.* **44**(3), 401–417.

[180] Coppi, Maddalena V. et al (2001), 'Development of a genetic system for Geobacter sulfurreducens', *Appl. Environ. Microbiol.* **67**(7), 3180–3187.

[181] Hädrich, Anke et al (2019), 'Microbial Fe(II) oxidation by Sideroxydans lithotrophicus ES-1 in the presence of Schlöppnerbrunnen fen-derived humic acids', *FEMS Microbiol. Ecol.*

[182] Barbeyron, T. et al (2001), 'Zobellia galactanovorans gen. nov., sp. nov., a marine species of Flavobacteriaceae isolated from a red alga, and classification of [Cytophaga] uliginosa (ZoBell and Upham 1944) Reichenbach 1989 as Zobellia uliginosa gen. nov., comb. nov.', *Int. J. Syst. Evol. Microbiol.* **51**(3), 985–997.

[183] Fu, Jia Cheng et al (2019), 'Dyella amyloliquefaciens sp. Nov., isolated from forest soil', *Int. J. Syst. Evol. Microbiol.* **69**(11), 3560–3566.

[184] Thomas, François et al (2017), 'Gene expression analysis of Zobellia galactanivorans during the degradation of algal polysaccharides reveals both substrate-specific and shared transcriptome-wide responses', *Front. Microbiol.*

[185] Mathur, Sunil & Sutton, Joseph (2017), 'Personalized medicine could transform healthcare (Review)'.

[186] Simcoe, Mark et al (2021), 'Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color.', *Sci. Adv.* **7**(11), 24.

[187] Ding, Wentao et al (2020), 'Development and Application of CRISPR/Cas in Microbial Biotechnology'.

[188] Sampson, Joshua N. et al (2011), 'Selecting SNPs to Identify Ancestry', *Ann. Hum. Genet.* **75**(4), 539–553.

[189] Keinan, Alon et al (2007), 'Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans', *Nat. Genet.* **39**(10), 1251–1255.

[190] Bansal, Vikas & Libiger, Ondrej (2015), 'Fast individual ancestry inference from DNA sequence data leveraging allele frequencies for multiple populations', *BMC Bioinformatics* **16**(1), 4.