# An Integrated Machine Learning and Experimental Approach to Uncover Ageing-Associated Processes in Fission Yeast

**Olivia Valerie Hillson**

Submitted in partial fulfilment of the requirements for the degree of

**Doctor of Philosophy**

University College London

**2023**

I, Olivia Valerie Hillson, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

# Acknowledgements

I would like to acknowledge and thank Prof. Jürg Bähler and everyone at the Bählerlab (past and present) for their help and collaboration throughout this process and my second supervisor Prof. Christine Orengo and her lab for theirs.

Extra special thanks to María Rodríguez-López for holding my hand through everything I had to learn, to Shaimaa Hassan for being ‘in this together’ with me, and to Babis Rallis and Martina Neville, without whom I would never have even applied to do a PhD.

Eternal gratitude to Jolanta Beinaroviča who forged through her own PhD one step ahead, dragging me along through mine in her wake, kicking and screaming. Thank you (I think?) for getting me here – you’re the best.

Big love to my family for always supporting me, no matter how crazy this journey seemed to you. And to Jack, thank you for going with me to the end, into the very fires of Mordor. Most importantly of all, thank you to my Mum, Producer of Doctors – everything I have and everything I am, I owe to you. This is no exception.

And finally, something I never imagined I would have to write. I would like to dedicate my thesis to the memory of StJohn Townsend. He was very special to us all, and I wouldn’t have finished this without his constant teaching, support, and friendship. He is sorely missed.

*“When you are a Bear of Very Little Brain, and you Think of Things, you find sometimes that a Thing which seemed very Thingish inside you is quite different when it gets out into the open and has other people looking at it.”*

*-A.A. Milne, Winnie the Pooh*

# Abstract

This work attempts to bring together knowledge of different pathways associated with cellular ageing and create connections between them using both machine learning and experimental methods. Initially, I describe the development of a novel proxy for chronological lifespan as part of the analysis pipeline of a high-throughput chronological lifespan assay in fission yeast. I then use this technique to go on to develop novel machine learning models that can predict lifespan, a complex phenotype, from simple traits, and identify ageing-associated phenotypes in fission yeast.

Complementary to this, I investigate a transcription factor of interest, Hsr1, for its involvement in cellular ageing and ageing-associated processes. I describe direct regulatory targets and how it forms a network with at least four other ageing-associated transcription factors which bridges the gaps between models of ageing, and suggest mechanisms for these interactions.

In this way, this work provides novel links between cellular ageing mechanisms and ageing-associated processes from both machine learning and experimental sources.

# Impact Statement

Ageing is one of the biggest socioeconomic challenges faced by the developed world today. Interventions for health decline in ageing populations and ageing-related diseases, such as cancer and Alzheimer's, can benefit from deeper understanding of the mechanisms which cause and control the ageing process. Current models of such mechanisms are widely disputed and generally thought to only tell part of the story. This project aims to develop machine learning approaches, integrated with experimental techniques, to give insights into the overarching mechanisms of the ageing process and attempt to elucidate how known pathways are linked.

This thesis describes the application of high-throughput methods to determine the lifespan of natural strains of fission yeast for a machine learning model which predicts the chronological lifespan of non-dividing cells from simple phenotypes in response to multiple environmental conditions. The model aims to objectively identify which phenotypes are most predictive of ageing and, therefore, which molecular pathways are most involved with the ageing process. In contrast to more traditional methods, this approach might provide a broader understanding of multi-pathway involvement in ageing.

Additionally, this project contributes to developing insights into the transcriptional control of ageing cells in fission yeast. The transcription factor Hsr1 is shown to have a resistant phenotype in the ageing-associated stress of caffeine and rapamycin and is then further characterised including binding targets, binding motifs, and genome-wide genetic interactions.

Reference:
Romila, A., Townsend, S., Malecki, M., Kamrad, S., Rodríguez-López, M., Hillson, O., Cotobal, C., Ralser, M., Bähler, J. (2021). Barcode sequencing and a high-throughput assay for chronological lifespan uncover ageing-associated genes in fission yeast. Microb Cell, 8, 146-160.

# Contents

# Table of figures

# Preface

Ageing is a multifactorial process spanning cellular damage, loss of biological function, disease, and death (Fontana et al., 2010). The overarching goal of any treatment is healthy lifespan increase without reduction in the quality of life. The UK is faced with an ageing population and many of its major medical challenges are ageing-related diseases such as dementia, diabetes, and cancer. While lifespan has been increasing in the UK, healthspan has not been increasing at the same rate which creates the medical and socioeconomic challenge of a large population left in a prolonged state of declining health (Olshansky, 2018). As a result, research which develops an understanding of the elusive mechanisms underpinning this process is of great relevance to many of the medical and socioeconomic challenges faced by the UK today. Lifelong health is also a key challenge area for the funding body BBSRC, the sponsor of my studentship.

When it comes to determining the causes of ageing, there are a range of models in use today, all with evidence for and against their accuracy. The evolutionary model of ageing is a popular and widely accepted theory which postulates that the force of natural selection is stronger in younger, reproductively active individuals than in older individuals. This means that

genes expressed when an individual is younger are more subject to selective pressures than those expressed when an individual is older (Hughes and Reynolds, 2005). This theory, however, does not constitute a mechanistic explanation for the process of ageing, rather an explanation as to why the process of ageing is occurring. Being non-mechanistic means the theory provides little scope for ageing interventions based on it, which is the focus of today's ageing research.

Since ageing is a multifactorial process, mechanistic models developed to describe it have thus far had serious failings. They have also been wide reaching, involving entirely different pathways to each other (Davidovic et al., 2010). Current mechanistic theories can be split loosely into two groups, models of programmed ageing and models of ageing caused by damage and error. The programmed models deal with the idea of a program of changes taking place to cause senescence such as the control of genes or programmed endocrine and immunological changes. The models involving damage and error include theories surrounding the damage to cells from 'wearing parts', metabolic speed, free radicals, and DNA damage (Jin, 2010).

Programmed theories of ageing are an intuitive way of describing what is often seen as an unavoidable and inexorable decline towards death. They include theories of programmed genetic changes, hormonal control, and immune decline to describe an ordered series of changes in an organism which leads to senescence (Cornelius, 1972, Hayflick, 2007, van Heemst, 2010). However, as more and more evidence accumulates to describe the effect of random damage events on lifespan, it seems likely that the idea of programmed ageing is a theory which is predicated on an comforting but oversimplified narrative that ageing is an ordered and controlled process (Hayflick, 2007).

The wearing parts theory is an idea that certain cells and tissues have components which wear and tear over time with repeated use in a similar way to wearing parts of a car (Jin, 2010). However, this theory is countered by evidence showing that organisms protected from these damages show no difference in lifespan (Park and Yeo, 2013).

Arguably, one of the most popular mechanistic models is the free radical model of ageing. Developed in the 1950s, this theory suggests that ageing is caused by the damage accumulated from reactive oxygen species. This theory has some success in accurately representing the strong impact of oxidative stress on ageing: a decrease in reactive oxygen species production has been shown to increase lifespan in numerous studies across multiple model organisms (Gladyshev, 2014). However, a study in *C. elegans* showed that increasing oxidative stress within an organism did not necessarily decrease its lifespan (Van Raamsdonk and Hekimi, 2009). This kind of evidence suggests that oxidative stress accumulation can be a life-limiting factor in some circumstances but that it is not a simple cause-and-effect relationship.

The rate of living theory is closely related to the free radical theory. This model suggests that an organism’s lifespan is inversely correlated to its rate of metabolism (Jin, 2010). Increased rate of metabolism would increase the rate of cellular processes and occurrence of damage, thus shortening the lifespan. This theory seems to integrate oxidative stress pathways and the target of rapamycin nutrient response pathway which is another popular candidate for ageing involvement (Brys et al., 2007, Rollo, 2010). The rate of living theory has evidence to suggest a strong correlational relationship but is lacking in evidence for a direct causal link (Park and Yeo, 2013). It is possible that this theory is

more descriptive of an increased rate of life and ageing as opposed to a mechanistic model.

DNA damage theory is a strongly evidenced theory of ageing. Multiple studies have shown that increased DNA damage shortens lifespan and increased DNA damage repair extends it (Park and Yeo, 2013). Interestingly a recent study into the transcriptional profile of the ‘immortal jellyfish’ *Turritopsis dohrnii* at different life cycle stages strongly supports this theory. As the only known organism to ‘reverse’ it’s ageing process, transcriptional changes in *T. dohrnii* offers a unique perspective on transcriptional changes for ageing. It has been shown that genes involved in DNA integration, transposition and repair, and telomere maintenance and organisation are strongly upregulated in the life cycle stage during which the organism metamorphoses back into the preceding lifecycle stage (Matsumoto et al., 2019). However, DNA damage could also be contributing to ageing by the damage of ageing related genes specifically as opposed to just the process itself.

The combination of contradictory and complementary theories and evidence around ageing models suggests a far more complex interplay between pathways and processes than has currently been described. This idea is also intuitive since ageing-associated diseases involve a wide range of organs and processes from diabetes to cardiovascular issues to neurodegeneration (Jaul and Barron, 2017). Combined, the current evidence is pointing to a need for a mechanistic model which envelops these global processes and describes how they interact to contribute to ageing.

The PhD project aims to investigate molecular mechanisms involved in cellular ageing in *S. pombe*. To identify the candidates, a machine learning tool which

predicts the complex phenotype of lifespan from simple, easily screenable phenotypes has been developed. This will help to elucidate which phenotypes are heavily predictive of lifespan and therefore more strongly involved in ageing. The project will also provide insight into the ageing-related transcription factor Hsr1, including its specific binding sites and genetic interactions, as well as its place in a network of ageing related transcription factors.

Chapter 1 characterises the training datasets developed and repurposed for use in a machine learning tool to predict lifespan, as well as detailing methodology for a lifespan proxy developed for this work. Chapter 2 then describes the process of building machine learning models to predict lifespan from simple phenotypes and identify the most predictive phenotypes. Chapter 3 explores the role of the transcription factor Hsr1 in cellular ageing and ageing-associated processes, including characterisation of Hsr1 binding and genome-wide genetic interactions of *hsr1*.

# 1 Characterising the Training Datasets for Prediction of Lifespan from Simple Phenotypes in Yeast

## 1.1 Introduction

### 1.1.1 Fission yeast as an ageing model

Yeasts have long been used as models for other eukaryotes, including human cells, and are a proven, long-standing and reliable system due to the presence of highly conserved cellular ageing pathways in higher eukaryotes. *Schizosaccharomyces pombe*, or fission yeast, has been a popular yeast ageing model, second only to the traditional use of *Saccharomyces cerevisiae*, or budding yeast. Fission yeast is an appealing model organism due to some conserved processes not found in *S. cerevisiae* such as mRNA splicing (Lin and Austriaco, 2014). As an ageing model, fission yeast has the advantage of a short generation time of only 2.5 hours, along with being an easily manipulated organism with a commercially available deletion library of all non-essential genes. These benefits make it an excellent model choice over multicellular ageing models such as mice and zebrafish, whose generation time is many months.

### 1.1.2 Lifespan measurement in fission yeast

Ageing research in fission yeast usually involves measuring the yeast cellular lifespan. This measurement can be considered a complex phenotype, affected by a range of variables, many of which are poorly understood. The two main ways to measure lifespan in fission yeast are chronological lifespan and replicative lifespan.

The replicative lifespan measures the number of divisions a cell has before death and is widely used in similar models such as *S. cerevisiae* but is not a good measure of ageing in *S. pombe* as their ageing and lifespan does not seem to be connected to their replicative ability (Spivey et al., 2017). Alternatively, the chronological lifespan (CLS) assay is a technique for analysing cellular lifespan by measuring the loss of cell viability over time. CLS assays can be carried out in multiple model organisms and are a well-established technique for fission yeast (Chen and Runge, 2009). During a CLS experiment, cells are grown to stationary phase and a measurement of colony forming units (CFUs) is taken at timepoint 0 and subsequent timepoints until the number of CFUs is <0.1% of the initial cell survival at timepoint 0 (Rallis et al., 2013).

In budding yeast, the CLS survival curve shows an initial loss of viability followed by a regrowth and death cycle which can continue for months, caused by the budding of new cells. However, in fission strains which cannot mate (*h*- or *h*+ strains) this is not possible and a smooth decline in viability is observed. This makes fission yeast ideal for determining chronological lifespan without the interference of regrowth (Runge and Zhang, 2018).

Despite the gold standard nature of CLS assays, they are not without limitations. Importantly, CLS experiments show a significant inter- and intra-experimental variability, possibly due to the large number of variables which need to be controlled, some of which remain poorly understood e.g. flask size (Lithgow et al., 2017). In addition, CLS is measuring the strain's ability to survive under high stress and low nutrient conditions, and this ability to survive could be biologically distinct from the yeast's natural lifespan. For example, fission yeast could enter stationary phase and decline in a nutrient starved environment but then regrow from only one cell back into a large competitive population if nutrients became available. Because of this, it is possible that measuring lifespan only until 1% of the starting cell density is reached would be more appropriate for estimating cellular ageing, and thus more applicable to other organisms such as humans (Runge and Zhang, 2018). Despite these limitations, in the absence of a viable alternative, CLS is currently the standard method of measuring lifespan.

Since one of the major advantages of using fission yeast as a model is the ability to design experiments to screen large strain collections, the time-consuming nature of CLS has previously held back research from larger lifespan screens. However, recent research by StJohn Townsend in the lab has developed higher-throughput CLS methods including a high-throughput Bar-seq screen to identify altered CLS and a medium-throughput colony forming unit (CFU) assay (Romila et al., 2021). The Bar-seq screen, a high-throughput method capable of screening entire collections at once, was shown to be able to identify long- and short-lived mutants. This presents a novel opportunity for quick identification of strains of interest to ageing for future experiments (Romila et al., 2021).

The medium-throughput CFU assay (also referred to in this thesis as the high-throughput CLS assay) reported in this paper, offers a method to collect more traditional CLS data with a much higher capacity with the use of automated robotics and modelling. In a traditional CLS method, each sample would be plated to three solid media plates, this method allows for 8 samples to be plated to only one solid media plate, as well as utilising a RoToR HDA pinning robot (Singer Instruments) and a liquid handling robot, further reducing the workload per sample.

### 1.1.3 Flocculation and longevity

When measuring the lifespan of fission yeast via CLS, the potential for flocculation must be considered. Flocculation is a natural process of yeasts, including *S. pombe* where the cells clump together to form what is known as a floc (Kwon et al., 2012). This clumping is achieved through cell-to-cell adhesion where cell surface glycoproteins, known as flocculins, bind to cell surface carbohydrates, quickly creating larger structures of many cells, which precipitate out of the liquid media (Soares, 2011, Stratford, 1989).

This precipitation creates a fundamental challenge for the CLS protocol, increasing variability in the step of collecting samples for the cell viability assay. If the yeast cells have flocculated, they may not be easily re-homogenised within the media, meaning that the sample taken would not be representative of the population of viable cells in the culture as a whole. Additionally, performing mixing to re-homogenise the culture could result in added stress to the cells which may also affect cell viability. While flocculation has not been observed during CLS with the fission yeast *972 h-* lab strain, it is possible, or even likely, that it could be observed in other strains, especially wild type strains which

arguably retain more natural *S. pombe* phenotypes (Kwon et al., 2012, Jeffares et al., 2015).

Investigation of flocculating *S. pombe* mutants has shown that similarly to other yeasts, fission yeast flocculation is calcium dependant but, differing from *S. cerevisiae*, cell adhesion relies on galactose residues rather than mannose or glucose (Tanaka et al., 1999). In *S. pombe*, it is thought there is a network of transcriptional control for flocculation affecting both when flocculation occurs and the size of the flocs. This network has been shown to include the transcription factors Mbx2 and Rfl1 as key regulators of flocculation alongside several other genes, pointing towards multifaceted and nuanced transcriptional control (Kwon et al., 2012).

Though the transcriptional control of flocculation has not been fully elucidated in *S. pombe*, it is understood that yeast form flocs in response to environmental stresses as the cells within the floc are protected by a barrier of the external cells (Smukalla et al., 2008). With this in mind, we can theorise that flocculation may occur during a CLS assay when the cells experience the stress of low nutrients, and that this flocculation could lead to longer lifespan measurements with the internal floc cells being protected from the environmental stress, increasing their longevity.

### 1.1.4 Ploxine B as a novel lifespan measure

While CLS is considered the gold standard lifespan measure it cannot be considered the most natural. In the wild, fission yeast would not live in shaking liquid media and has been shown to survive for longer when able to form colony structures. It has even been suggested that fission yeast colonies are so beneficial they form an almost multicellular-like structure. Cell viability assays

such as the one used for CLS would not be possible from a colony so a different measure of cell death would be needed.

Recently the dead-cell stain phloxine B has been used with success to measure cell death in colonies within a high-throughput screen context (Kamrad et al., 2020). Phloxine B is a chemical which only stains dead *S. pombe* cells red. If phloxine B is added to solid media, it is actively pumped out of live cells but accumulates in dead cells meaning that 'redness' can be used as a measure of the proportion of live cells (Kwolek-Mirek and Zadrag-Tecza, 2014). The 'redness' measure of a colony is shown to be directly proportional to the number of dead cells. In this way it could be possible to use redness over time in the same way as cell viability over time is used to create a lifespan curve. Measuring cell survival within a colony could be a novel, more intuitive method of measuring lifespan.

## 1.2 Methodology

### 1.2.1 Design principles

This research aims to develop a machine learning model which predicts the complex phenotype of lifespan from simple phenotypes in fission yeast. Insights gained from this model could then be applied to higher eukaryotes due to the large number of conserved pathways in fission yeast. In this chapter, the training data chosen for creating this model is characterised and discussed to demonstrate that the research questions of the model are being appropriately addressed.

The first step in any modelling requires a sound and reliable training dataset. As discussed in the introduction, the training dataset defines what is ‘true’ for the model and determines what any output is able to tell us. Therefore, it is important to ensure that the training data provides a strong foundation for the model which is specific to answering the research questions we have set out:

1. Can the complex phenotype of lifespan be predicted from simple phenotypes?
2. Which simple phenotypes are most predictive of lifespan?
3. Based on these phenotypes, which cellular processes are most predictive of ageing?

In this case, as with many ageing models, the ‘true’ data input that must be clearly defined and specific to the research questions is the ‘lifespan’. The training dataset for lifespan will define to the model what ‘lifespan’ and ‘ageing’ are in our research questions, so we must ensure that the data is specific for this purpose.

To achieve this, the methodology for this research uses a collection of wild-type yeast strains, sampled from over 20 countries across the globe, to give a normally distributed dataset which is representative of fission yeast as a species

(Jeffares et al., 2015). A more commonly used, and much larger, dataset would be the commercially available Bioneer deletion library. However, deletion mutants often have unusual lifespans, often extremely short lived, so this kind of data may not be appropriate for machine learning applications.

As well as this, keeping in mind the specific questions the model is designed to answer, wild type yeast is the better choice for training. The lifespan of a deletion mutant is often directly impacted by the gene deletion, and this would interfere with forming connections between the lifespan and more subtle phenotypes. Wild type yeast is less likely to have any extreme phenotypes so offers a greater chance of identifying these subtle connections.

The Bioneer deletion library also does not offer any natural genetic diversity, as the library is all transformed from a single strain. Using a genetically diverse collection of wild type yeast allows the addition of more species representative data into the model. Consequently, while the Bioneer deletion library would offer a significantly larger dataset, which would be better for modelling, the collection cannot address the research questions in the same way the wild type collection can.

Chronological lifespan assay (CLS) was then chosen to define the complex phenotype of lifespan, due to its long-standing and reliable history in both fission yeast and the field of ageing research. The availability of a high-throughput, robot based, method for CLS assays within the lab was a major advantage for the method, as it allows for the collection of the quantity of data necessary for modelling.

### 1.2.2 Wild type fission yeast strain collection

The collection used in this research contains a total of 161 *S. pombe* isolates, published by Jeffares et al. (2015). The strains were collected from more than 20 different countries during the last 100 years. Most of the strains were collected from cultivated fruit and fermentations and once analysed, the 161 samples had 57 non-clonal strains. These 57 strains have ≥1,900 different SNPs to each other which accounts for 99.6% of the SNPs found in all 161 strains in the original collection (Jeffares et al., 2015). The collection has also been shown to have large structural variations which contribute to the collection's rich phenotypic diversity (Jeffares et al., 2017). This collection of strains presents an opportunity to investigate the genetic and phenotypical differences between natural fission yeast strains and Jeffares et al. showed that this collection had rich genetic and phenotypic variations which could be further investigated.

### 1.2.3 Chronological lifespan testing of *h*+, *h*- and *h90* strains

An initial problem with the wild type collection was that many of the strains had a homothallic mating type of *h90* rather than *h*+ or *h*- meaning it was possible for them to mate during the lifespan and seriously affect the results. Even though the CLS experiment would be carried out using nutrient-rich YES media there was a possibility of the *h90* strains mating once they were under stress. To test this, the lifespans of three lab strains, a *h*+ strain (JB32), a *h*- strain (JB972) and a *h90* strain (JB50), were measured.

Two replicates of the strains were woken up from cryostock by streaking on to YES agar and grown at 32˚C for two days. Single colonies were then picked from each plate and resuspended in 1ml YES in a UV-sterilised cuvette. The

ODs of each colony were taken and 10ml OD corrected YES precultures were set in 25ml conical flasks and left shaking at 32˚C overnight. YES cultures (10ml) of OD 0.002 were then set in 25ml conical flasks and left shaking to grow for two days at 32˚C. After 48 hours, or at 'Day 0' of the lifespan assay, 150µl of each culture was serially diluted in YES and spotted four times on to a YES plate. These plates were scanned once they were grown, and in-house R scripts were used to calculate the number of colony forming units in each culture. This process was repeated until Day 7.

### 1.2.4 Chronological lifespan data for wild type strains

Lifespan data of the wild type strains was collected in three batches to allow for three complete, independent biological repeats and high-throughput CLS was performed as follows:

57 wild type strains were woken up from cryostock by spotting on to a YES agar PlusPlate using a RoToR HDA robot (Singer Instruments) and long-pin 96-density pads. These spots were left to grow for 2 days and then used to set 57 precultures in 10ml YES media in 25ml volumetric flasks which were left to grow overnight, shaking, at 32˚C. Cultures were then set from in 10ml YES in 25ml volumetric flasks, corrected to OD 0.002 and then placed in the shaking incubator at 32˚C.

After two days of growth, when the cultures had reached stationary phase, the reading for Day0 was taken. For this, a 150µl sample from each culture was transferred to the first column of a 96-well plate and a serial dilution in YES was performed using an Integra Assist automated multichannel pipette (Integra Biosciences Ltd.). The dilutions were then spotted in quadruplicate on to a YES agar PlusPlate using the RoToR HDA robot (Singer Instruments) and long-pin

96-density pads, making sure to revisit the source plate before each pin. The plates were incubated at 32˚C for 2-4 days, until suitable growth was seen.

Once grown, stress plates were imaged using a conventional scanner and a custom Unix script within the lab to crop (figure 1A). Our R package *DeadOrAlive* (Romila et al., 2021) was used to process the spot plates into lifespan curves. Image analysis, based on the R package, grids the images to show every available location for a spot. Locations containing a spot were marked in red and those not containing a spot were marked in blue. This blue-red step, shown in figure 1B, is useful for quality control to ensure that the spots have been correctly identified. The number of repeat pins (0-4) which show a spot is then calculated as shown in figure 1C. The number of spots out of the four repeat pins for each dilution factor creates an array which predicts the most likely number of colony forming units (CFUs) per droplet which would make that array. It also allows the computer to perform some quality control by excluding spots which occur in a serial dilution after several dilutions of no spots such as the point highlighted in yellow in figure 1B and C.

During the second biological repeat of the full lifespan set, a sample of each culture was checked under the microscope on each read day to look for evidence of mating in the form of asci or spores. Qualitative flocculation data was also recorded during the second repeat.

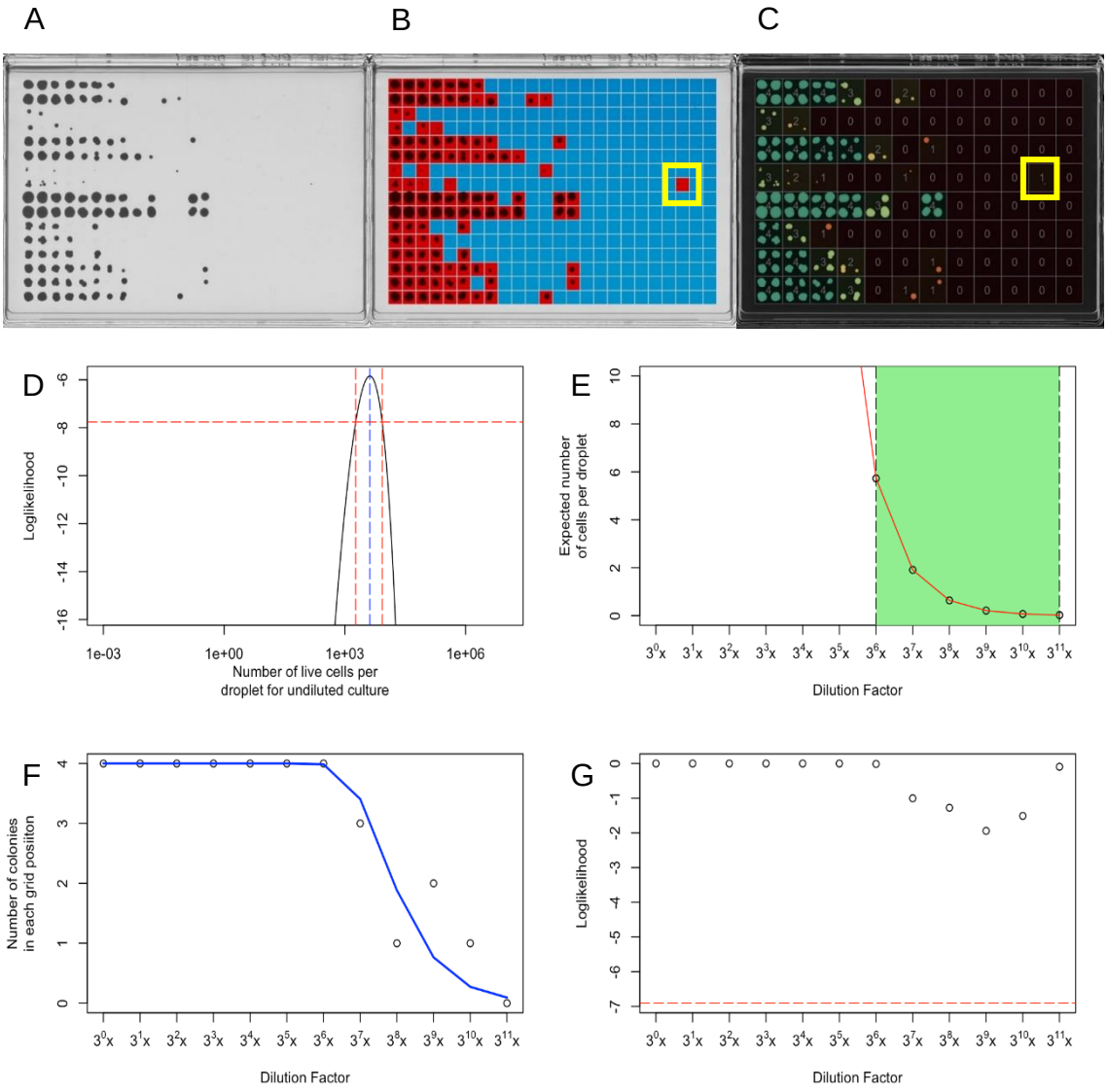



Figure 1: Example of image processing and maximum likelihood estimation for the wild type lifespan data

A) The serial dilution spot plate is scanned; colonies appear dark on a light background. B) The plate image is overlaid with a grid over each potential spot on the plate, these are then coloured red if they contain a spot on the image and blue if they do not. C) This image is then processed to show how many of the four replicate spots has grown. These number arrays are then used to calculate the number of CFUs per droplet expected to create this array pattern. The spot highlighted in yellow in B and C was excluded data. D) The likelihood function for JB22, day 0, repeat 1: blue dashed lines indicate the maximum likelihood, and red dashed lines show the bounds of the confidence interval for the desired probability. E) The expected number of cells per droplet at each dilution based on the maximum likelihood estimate for JB22, day 0, repeat 1. The shaded green area shows the expected informative region (the region in which there is reasonable probability that some positions will contain a colony and some will not). F) The pattern of observed colonies for JB22, day 0, repeat 1: the blue line indicates the expected distribution based on the maximum likelihood estimate. G) The likelihoods of observing each particular data point for the maximum likelihood estimate for JB22, day 0, repeat 1. The red dashed line shows the tolerance: if there any data points for which the probability of observing them is less than the tolerance, then it is assumed that something has gone wrong. In this case, the most troublesome data point will be excluded, and the maximum likelihood estimation will be performed again.

### 1.2.5 Cell adhesion phenotype data for wild type strains

Cell adhesion phenotype data was produced by Bence Kover (2023, UCL, personal communication).

The 57 non-clonal wild type strain collection was woken up from cryostock onto a YES PlusPlate using the RoToR HDA robot and long-pin 96-density pads (Singer Instruments). The cells were incubated for 3 days at 32°C. Cells were then resuspended in 100µl of EMM + nitrogen media in a 96 well plate using the RoToR HDA robot and long-pin 96-density pads and ensuring mixing in both the source and target plates. The liquid resuspension was pinned on to EMM + nitrogen agar in 96 format using the RoToR HDA robot and long-pin 96-density pads. Each sample is pinned in a 7x7 square to allow for wide adhesion to the agar, ensuring the source plate was revisited before each pin, and then incubated for 4 days at 32°C.

Plates were imaged using a conventional scanner and a custom Unix script to crop. The cells were then washed from the plate using water with a constant flow rate of 35ml/sec for 1 second on each 7x7 square. Following the washing they were imaged a second time. The ratio of pixel intensity between the washed and unwashed plates is the cell adhesion score, measuring cell adhesion to the agar. Pixel intensity was processed with a custom python-based pipeline making use of scikit-image.

### 1.2.6 Lifespan proxies for wild type strains

Lifespan proxies were calculated using the *DeadOrAlive* package (Chen and Runge, 2009, Rallis et al., 2014, Roux et al., 2006). The package defines the proxy as the square root of the number of days to 5% viability, but to achieve the wider spread of data preferred by machine learning models, simply the

number of days to 5% viability was used in this research. The proxy was checked for reproducibility using ANOVA and Tukey's post hoc analysis, performed in R.

Because this proxy is a new protocol it was important to check for the intuitiveness of this novel method. To achieve that, a proxy was also calculated in a way to avoid the spline fitting, the most likely area to introduce bias. Here a linear regression line was constructed between the closest datapoints above and below 5% viability and the proxy was calculated along this regression line.

The lifespan proxy data was also used to create a categorised dataset by binning the proxy data into the three categories around the average of the dataset in R.

### 1.2.7 Phloxine B lifespan as a novel lifespan measure

57 strains were woken up from cryostock by spotting on to a YES agar PlusPlate using a RoToR HDA robot and long-pin 96-density pads (Singer Instruments). The plates were incubated for three days at 32˚C and then pinned on to YES (0.1% glucose) and phloxine b agar PlusPlates in quadruplicate (384 format) using a RoToR HDA robot and long-pin 96-density pads. The plates were incubated at 32˚C and imaged for redness at days 2, 4, 7, 9 and 11. Imaging was performed using a conventional scanner and a custom Unix script to crop. *Pyphe* (Kamrad et al., 2020) was then used to quantify the redness of each colony.

To normalise the data and fully capture the meaning of the redness scores the redness scores were divided by the mean redness score for each timepoint. This allows the data to represent higher or lower redness than the average. The

relative change in redness between day 2 and day 11 was calculated and used as the proxy score for lifespan.

## 1.3 Results

### 1.3.1 Chronological lifespans of wild type strains

During the test of *h*+ *h*- an *h90* strain lifespans, the cells of both *h90* repeats showed some evidence of mating under the microscope in the form of a few spores. There was little evidence of mating and the *h90* strain lifespan showed no significant difference from the *h*- strain, suggesting that the mating had no significant effect on the overall lifespan. The *h*+ strain, JB32, showed a much-decreased lifespan in both biological replicates but this was attributed to a strain characteristic since the *h90* strain was not different from the *h*- strain.

Three full biological repeat lifespans of all 57 strains in the collection were generated. Some scans resulted in the analysis script calculating infinite CFUs per droplet for a small number of time points. At this point the number of CFUs cannot be calculated by the program as they are effectively out of the range of the experiment. The issue was worked around by removing infinite values and imputing the missing data with k-nearest neighbour.

Figure 2 shows the reproducibility of the lifespan between repeats for each strain with real data shown in grey and imputed datapoints shown in red. This figure also highlights that the k-nearest neighbour imputed values lay intuitively in the lifespans and do not introduce any outlying datapoints.

Qualitative observation data showed that only JB871 had any evidence of mating during the lifespan. The evidence was limited with only a few spores and asci observed in the final days of the lifespan. Strains 32, 34 and 36 all flocculated during the lifespans, with 36 flocculating enough to form one solid mass in visibly clear media towards the end of its lifespan. All flocculation was carefully homogenised by gentle pipetting before the sample was taken.

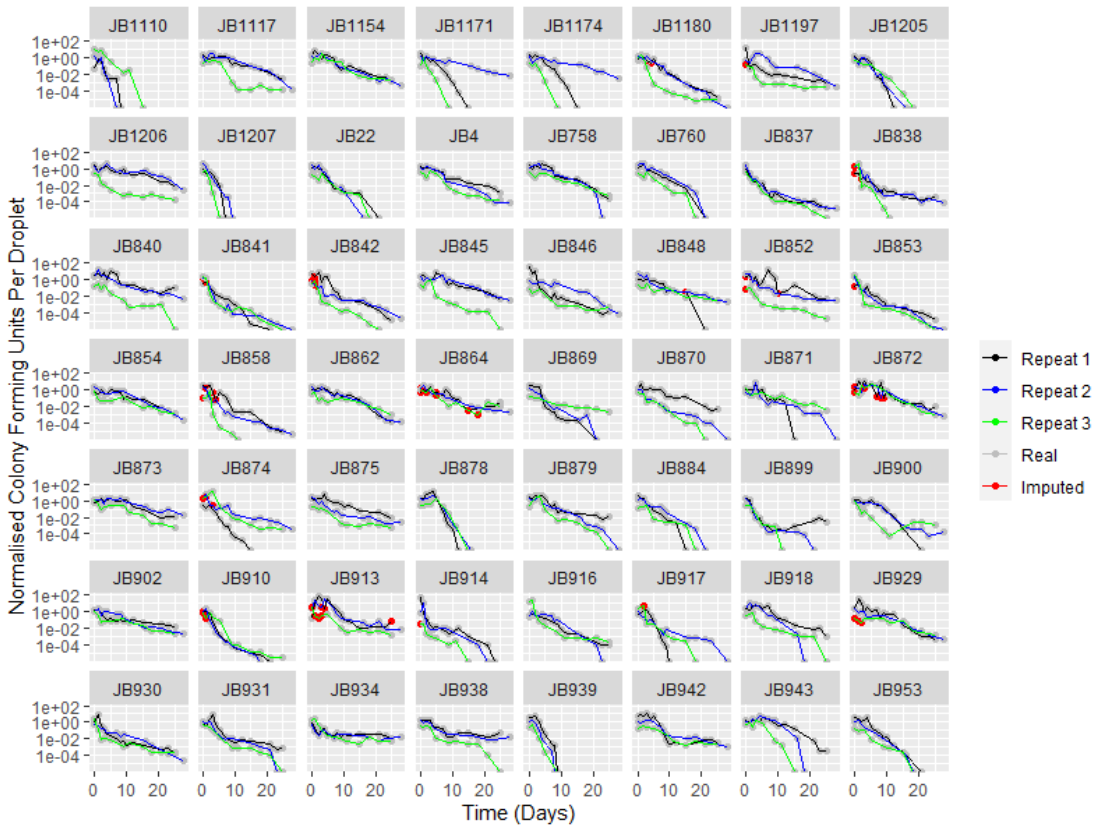



Figure 2: Graphs to show 57 wild type stain lifespans in triplicate.

Graphs showing normalised colony forming units per droplet over time in days for all 57 strains. Repeats are indicated in black, blue and green respectively and values imputed using k-nearest-neighbour imputation are shown in red.

## 1.3.2 Lifespan proxies for wild type strains

Initial experimentation with alternative proxy calculations yielded limited results for the purposes of this study. The implementation of area under the curve (AUC) calculations is shown in figure 3. While the AUC proxies to both 5% and 50% viability using the trapezoid and spline fitting methods were highly correlated between repeats, the data was clustered towards small proxies and lacked an even spread from short to long lived.

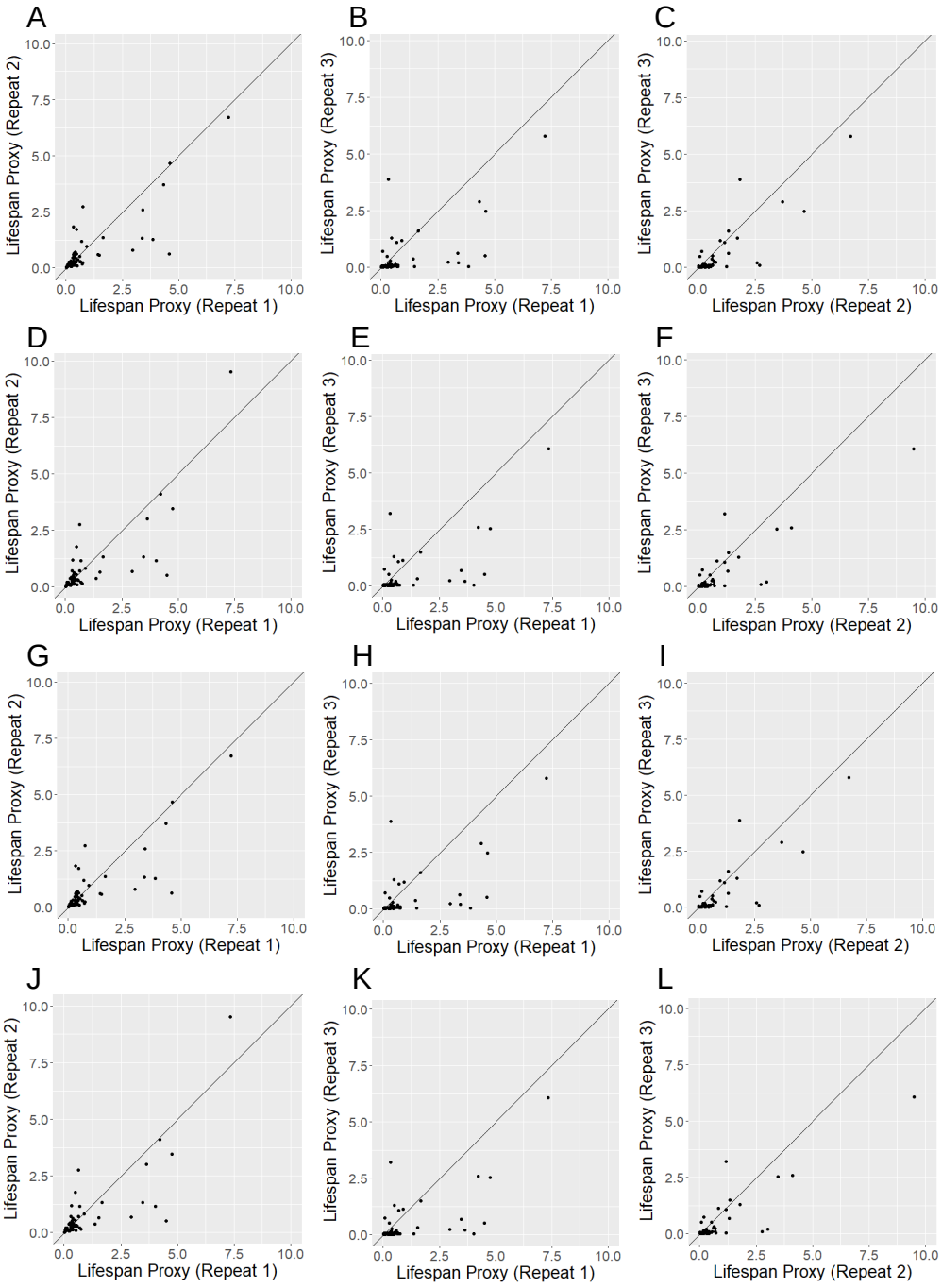



Figure 3: Pairwise correlation scatterplots for area under the curve (AUC) based proxies.

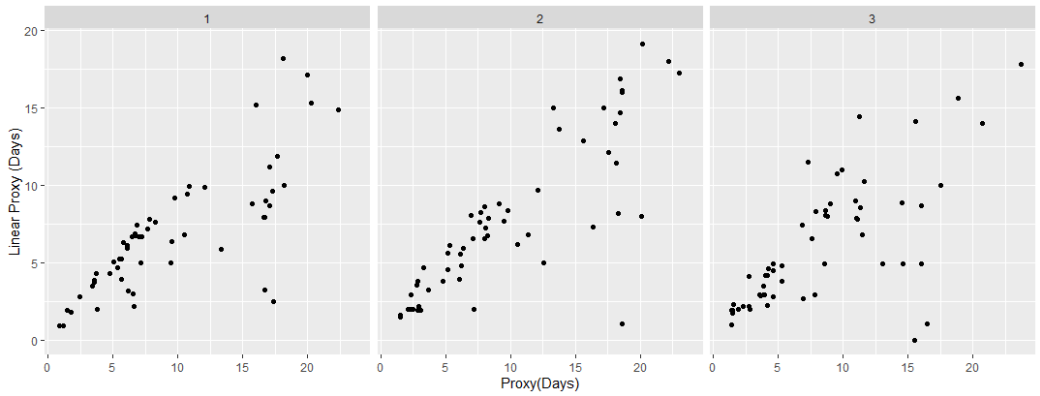
AUC to 5% viability correlations fitted with trapezoid method between (A) repeat 1 and repeat 2 (coefficient = 0.81***), (B) repeat 1 and repeat 3 (coefficient = 0.64***), and (C) repeat 2 and repeat 3 (coefficient = 0.84***). AUC to 5% viability correlations fitted with spline method between (D) repeat 1 and repeat 2 (coefficient = 0.80***), (E) repeat 1 and repeat 3 (coefficient = 0.67***), and (F) repeat 2 and repeat 3 (coefficient = 0.86***). AUC to 50% viability correlations fitted with trapezoid method between (G) repeat 1 and repeat 2 (coefficient = 0.81***), (H) repeat 1 and repeat 3 (coefficient = 0.64***), and (I) repeat 2 and repeat 3 (coefficient = 0.84***). AUC to 50% viability correlations fitted with spline method between (J) repeat 1 and repeat 2 (coefficient = 0.80***), (K) repeat 1 and repeat 3 (coefficient = 0.67***), and (L) repeat 2 and repeat 3 (coefficient = 0.86***). ***p-value<0.001.

The lifespan proxy from our *DeadOrAlive* package and the linear proxy were performed on all three repeats of the lifespans, and these were plotted against each other. Figure 4A shows how the proxy developed for *DeadOrAlive* correlates strongly with the linear proxy for all three repeats of the lifespan. A histogram of the proxy results for all three repeats (figure 4B) combined was also produced and shows the dataset has a normal distribution of lifespans with a slight skew towards short lived strains.

ANOVA analysis of the lifespan proxies showed significant difference between strains ($p=<0.0001$) and no significant difference between repeats ($p=0.334$). This was followed by Tukey's multiple comparison post hoc analysis showing no statistically significant difference between repeats for all comparisons ($p=0.94$, $p=0.52$, $p=0.33$).

The mean *DeadOrAlive* lifespan proxy was plotted in a pairwise correlation with the cell adhesion score from the data generated by Bence Kover (supplementary figure 1). This plot showed that there was no correlation between the lifespan and the cell adhesion score.

A




B

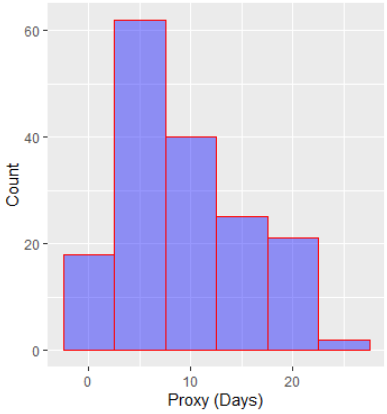



Figure 4: Quality control for lifespan proxy training data

A) Scatter plots to show the strong positive correlation between the linear proxy and our proxy for repeats 1, 2 and 3 respectively. B) Histogram to show the spread of the training data lifespan proxies. There is a normal distribution skewed slightly towards the shorter-lived strains.

### 1.3.3 Phloxine B as a novel lifespan measure

The phloxine B data was incomplete once the scans had been processed in *Pyphe* as the program failed to identify many colonies leading to lots of missing datapoints. The data also suffered from a lot of noise and irreproducibility problems between repeats. After normalisation and the calculation of rate of redness, this rate of redress lifespan proxy was plotted in a pairwise correlation with the *DeadOrAlive* lifespan proxy mean of the wild type lifespans (figure 5). This shows a negative correlation, as rate of redness increases the lifespan of the strain decreases. The correlation has a small negative coefficient and a relatively high p-value, but it does indicate that there may be a negative correlation with less noisy and more reproducible data.

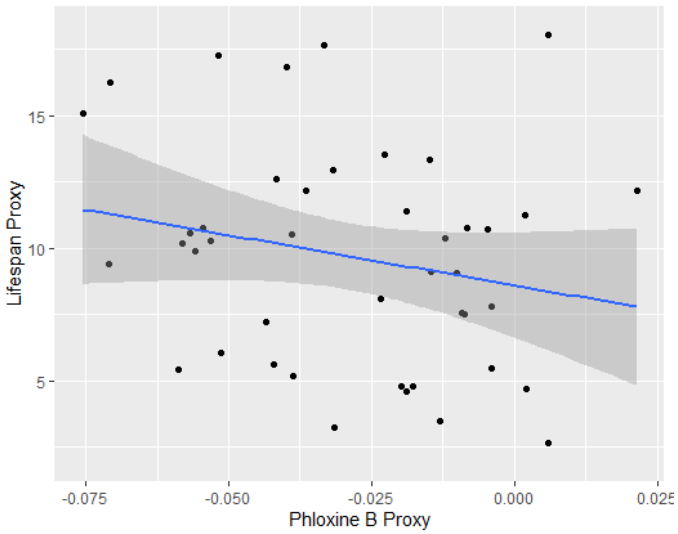



Figure 5: Preliminary insights into the use of Phloxin B as a lifespan measure.

Correlation of rate of redness over time proxy with the mean of the lifespan proxies (coefficient = -0.22, p-value = 0.16), showing redness increases as lifespan decreases.

## 1.4 Discussion

### 1.4.1 CLS and lifespan proxies of wild type strains

The lifespans of the 57 wild type yeast strains show an expected level of noise and despite this are highly reproducible, especially considering that irreproducibility issues are commonplace in this technique. When plotted (figure 2), it can be visualised that the lifespans are similar to each other between repeats, and statistical analysis showed that the repeat proxies were reproducible to within a 99% confidence limit.

While some data needed to be imputed due to infinite predictions, this occurred in only a small number of datapoints and visualisation of the lifespans (figure 2) showed that these imputed datapoints sat intuitively within the lifespan curve and did not introduce outliers to the data. In data such as lifespans where the trend is highly predictable a knn imputation is often a reliable form of imputation.

The infinite values were, on two occasions produced by bacterial contamination of the plate, and on other occasions likely due to either high cell density or cell stickiness. There is a large range of cell morphology within the wild strain collection, owing to its generic diversity, this could lead to strains with smaller cells having much higher cell densities than the JB22 strain on which the high-throughput CLS assay was designed. Several of the strains in the wild type collection flocculate indicating that they have high levels of sell stickiness. This stickiness could cause the strains to stick to the pipette tips during the serial dilution, introducing systematic error.

This theory of systematic error causing infinite values highlights a concern that cell stickiness could also be causing artificially long lifespan results in the CLS by the same mechanism. To account for this, observational data and

quantitative data from a cell stickiness assay was correlated to the lifespan data to look for a relationship between stickiness and longer lifespan. None of the strains which were observed to flocculate during the lifespan were especially long lived, and a pairwise correlation of the cell stickiness score to the mean lifespan proxy did not show any relationship (supplementary figure 1). In this way we can be confident that flocculation and cell stickiness did not significantly affect the reliability of the lifespan data.

Despite the positive indication from the *h*+, *h*- and *h90* test lifespans, during the second biological repeat of lifespans, qualitative data of mating was recorded. JB871 showed some evidence of mating towards the ends of the lifespan in the form of small number of spores and asci seen under the microscope. However, there were very few spores and asci observed and the strain was not especially long-lived, so it is not thought that enough mating took place to alter the chronological lifespan results in any meaningful way.

### 1.4.2 Lifespan proxies for wild type strains

For use in machine learning the lifespan curve needed to be translated into a representative single number proxy. Although there are standardised ways of achieving this with lifespans of multicellular organisms, within *S. pombe* or yeasts overall the methods used for this in the literature are sparse and extremely variable. This meant that it was necessary to develop a novel, intuitive method. Much of the literature used numbers of days as the final single number description of a lifespan, often by simply discussing the difference in the graph at a certain number of days or the number of days to a percentage viability, e.g. 50% or maximal lifespan (Chen and Runge, 2009, Rallis et al., 2014, Roux et al., 2006).

In line with the literature, we developed a proxy based on the number of days to maximal lifespan. Due to the nature of the high-throughput CLS analysis, at the end of a lifespan, cell survival can be calculated as less than one colony forming unit per droplet. Since this a measurement of less than a single cell the data is more robust before these miniscule final calculations. To account for this issue, maximal lifespan was defined as 5% of the maximum cell survival (day 0) ensuring that cell survival rates were always above one colony forming unit per droplet.

Since the literature sets a precedence for variable percentage survivals to be used as the final output of a CLS (Kalita et al., 2021, Rodríguez-López et al., 2020, Lee et al., 2021), a range of percentage survivals were trialled as the proxy for this research:

- 5% survival (representing maximal lifespan for this data)
- 10% survival (representing close to maximal lifespan but with the very end of the curve not considered)
- 50% (representing only the first half of the lifespan curve)
- 70% (representing only the beginning of the lifespan curve)

Each percentage survival has its own advantages and disadvantages due to what it represents. While 5% viability (in this case maximal lifespan) represents the entire lifespan and therefore takes all the data into account there are disadvantages to its use. Since the measure includes the entire lifespan curve, it can be skewed by any noise at the end of the lifespan. By using 5% rather than the absolute maximal lifespan, we are effectively trimming data from very small populations of long-lived cells which will help reduce the impact of end of lifespan noise but there is still a risk. A 10% viability could be used in place of a 5% viability to give a larger margin around any potential end of lifespan noise in

the data and more reliably eliminate it. However, taking more from the end of the data risks loss of insights found at the end of the lifespan curve.

Along with maximal viability, 50% viability is a commonly used metric within the literature (Kalita et al., 2021, Rodríguez-López et al., 2020, Lee et al., 2021). This survivability measure only represents the first half of the lifespan curve, but this is generally considered to be the part of the lifespan curve which is most rich in biological insights. Many lifespans will diverge most noticeably in the first half of the lifespan due to the initial drop in viability which occurs immediately after the cells exit stationary phase. This initial fall in viability slows as the lifespan continues before levelling off by the end, so the differences in lifespans can often be captured by the initial drop alone. To capture only the initial drop in viability 70% viability can be used, this measure works similarly to the 50% viability measure but has less risk of being affected by any noise after the initial drop in viability.

In the literature, often more than one metric is used at the same time to describe a lifespan e.g., 50% and maximal viability (Rallis et al., 2021, Mirzaei et al., 2014). However, for the purpose of the proxy it was necessary to use only one. The best measure was chosen based on the appropriateness of the output data as training data measured by:

- Reproducibility – If a measure is particularly affected by noise in this lifespan data, then the proxies will be less reproducible between the lifespans.
- Spread of datapoints – A measure is more appropriate for machine learning modelling if it can be representative of the large range of lifespans in this data and capture where the lifespan curves diverge.

Experimentation with using 70%, 50% and 10% survival did not produce proxies with as high reproducibility or the desired spread of datapoints and so 5% viability was decided on as the best metric.

From here there are two main options for lifespan calculation with precedence in the literature:

- A simple measure of the number of days it takes to reach 5% viability.
- A calculation of the area under the lifespan curve (AUC) to 5% viability.

Calculating the AUC as opposed to simply reading the x axis for a defined y value can often create a more robust description of the data by taking the entire lifespan curve until this point into account. However, for the same reason AUC measurements can be greatly affected by noisy data within the curve. With this data in particular, the AUC proxies did not perform as well, with clustering of datapoints in smaller proxies and a lack of an even spread of proxies for short to long lived strains (figure 3). This spread of data was seen in AUC proxies using both trapezoid and spline fitting methods, calculated to both 5% and 50% viability, and renders the proxy inappropriate for machine learning applications.

The resultant proxy, available in the *DeadOrAlive* R package and published in a previous paper from the lab (Romila et al., 2021) draws a smooth spline through the lifespan data to create a lifespan curve and then uses this curve to calculate the number of days taken to reach 5% viability. Several spline fitting methods were trialled during the proxy's development including second- and third-degree polynomials to ensure the lifespan curve struck the balance of being true to the original datapoints while eliminating noise which would skew the proxy calculation.

Initially, polynomials were fitted to the lifespan curve but neither second- nor third-degree polynomials fitted the data as closely as would be necessary and so attention turned to spline fitting packages. Ideally, the spline needs to be smooth enough to be unaffected by any noise in the data but also fit the data tightly enough that it is fully representative of the lifespan. Attempting to fit

defined shapes on the data such as sigmoid curves also failed to be properly representative. One of the major obstacles in fitting the spline was the noise at the beginning of the lifespan which often led to fitted splines going up at the start of the lifespan when we know, intuitively, that the viability always decreases.

The R package *Cobs* (Ng and Maechler, 2007) computes constrained quantile curves using linear or quadratic splines. The constraint options in this package allowed a spline to be fitted with the constraint ‘decrease’ which meant that the spline will always decrease as we know the lifespan does, in spite of noise. It outputs a median spline which is a robust smoother and sits intuitively over the lifespan data (figure 6).

For the purpose of the previous research paper (Romila et al., 2021), the *DeadOrAlive* package defines the proxy as the square root of the number of days to 5% viability, but to achieve the wider spread of data preferred by machine learning models, simply the number of days to 5% viability was used in this research. The proxy was checked for reproducibility using ANOVA and Tukey’s post hoc analysis, performed in R.

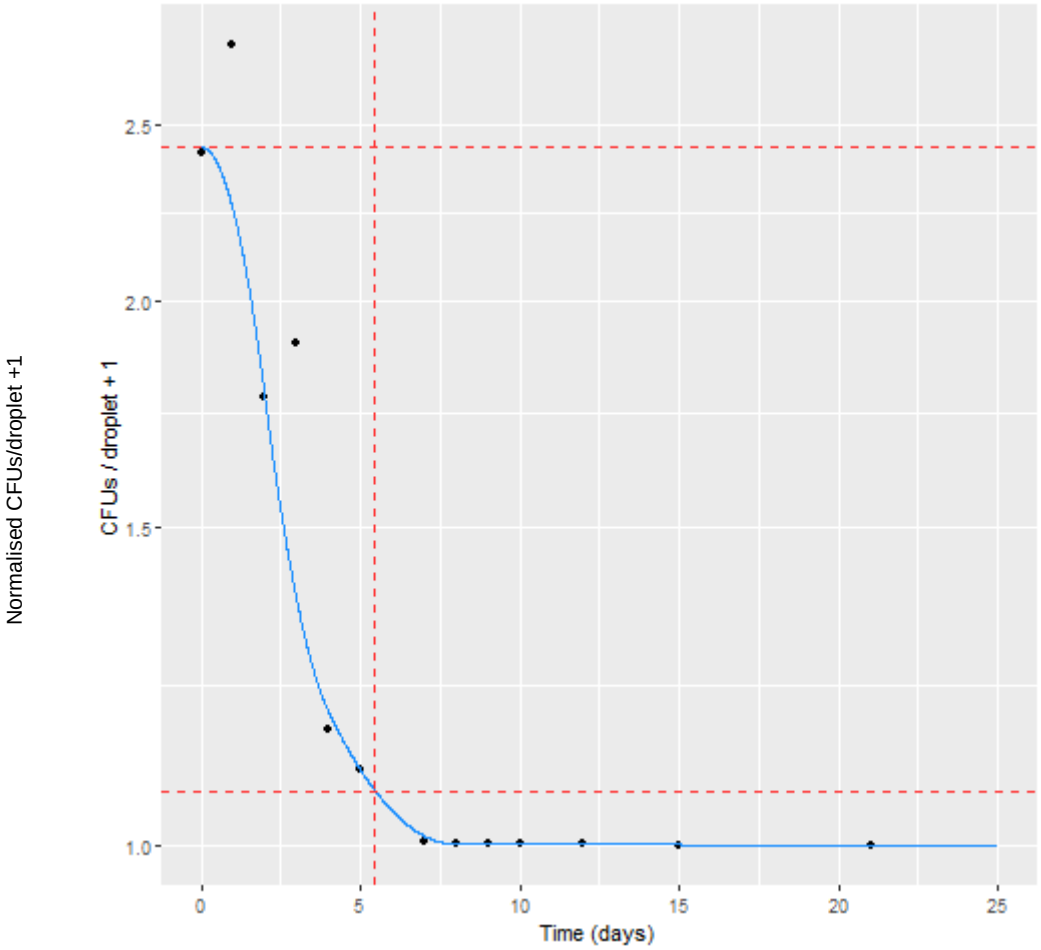



Figure 6: Example of spline fitting for lifespan proxy

Graph showing the actual data points for the lifespan of JB22 in the first repeat (black points), the spline fitted by the 'cobs' package with the constraint that it must always decrease (blue line), and the maximum prediction of the number of days to 5% viability (red dotted lines). These figures can be generated in the DeadOrAlive package by using the plotProxy() function.

Because this proxy is a new protocol, it was important to check for the intuitiveness of this novel method. To achieve that, a proxy was also calculated in a way to avoid the spline fitting, the most likely area to introduce bias. Here a linear regression line was constructed between the closest datapoints above and below 5% viability and the proxy was calculated along this regression line. This method would be more vulnerable to the effects of any noise in the data but should correlate well with the spline fitted proxy if it is intuitive. As seen in figure 4A our proxy strongly correlates with the linear proxy (with some noisy outliers) showing the accuracy of our proxy calculation.

After showing that the proxy calculation is an intuitive method, attention turned to the quality of the lifespan training data. During the proxy calculation one strain was found to be so long lived that it did not reach 5% viability in any of the repeats. This strain was removed as a quality control both because it was not possible to calculate the proxy and because the strain was such an outlier that it would skew the training data. As shown in figure 4B, the data has a relatively normal distribution, this is necessary for training data as a skewed dataset will usually preferentially predict the extreme it is skewed towards. The training data has a slight skew towards shorter lived strains, but it is very minor and therefore unlikely to affect the performance of any machine learning models trained on it.

### 1.4.3 Phloxine B as a novel lifespan measure

The initial data shows that this phloxine B measure of lifespan is correlated with the traditional measure, but the data was very noisy and not that reproducible. While the measure shows promise, the assay would need to be repeated and optimised to create more reliable and reproducible data. Future experiments should focus on increasing the quality of the scanned images so that less data is lost during the image analysis pipeline as well as introducing a gridded plate layout similar to the one used in the phenotype screen of the wild type strains. This would allow for downstream normalisation to positional bias and likely reduce noise and increase reproducibility.

It is, however, important to remember that the phloxine B assay is a different measure of lifespan compared to CLS data. Since the assay measures lifespan within a colony on solid media as opposed to within a liquid culture, it may not correlate directly to CLS data and this would not necessarily be a measure of the assay’s validity. Arguably, the assay has the potential to be a more natural measure of strain lifespan which could be used for further insights in the future.

### 1.4.4 Training datasets

Due to the level of noise and irreproducibility issues in the phloxine B data, this dataset was excluded from model building at this time. If the experiment was to be repeated to produce a more reliable dataset, this could be used in modelling in the future, but in its current state it is likely to decrease the performance of the model due to the quantity of outlying datapoints it contains. The chronological lifespan proxy data and the phenotype data were used to produce training datasets for modelling.

The lifespan proxies were used to produce two sets of training data:

1. Mean lifespan for the 56 strains (56 incidences)
2. Individual inputs of each repeat for the 56 strains (168 incidences)

Due to the relatively small dataset of lifespans available, the mean lifespan might not be the best way to input this data into a model. Therefore, a dataset where the repeats were used as individual datapoints within the models rather than averaged was produced. This has two potential advantages. Firstly, it triples the number of examples available to input into a machine learning model mitigating the issue that this dataset is small. Secondly, the aim of this modelling is to find insights into how the stress response phenotypes of a strain are related to its lifespan. In this instance all the lifespans recorded, even ones that vary more from their replicates, are lifespans of the strain and carry biological meaning. All the lifespans and phenotypes recorded carry their own biological significance for this question, perhaps most importantly in their variation from their replicates. By using each datapoint as an example instead of as a repeat, we are capturing and showing the model the biological range involved in this question rather than eliminating it.

As well as this, it is interesting to consider how we usually talk about lifespan data for *S. pombe* in the literature. For discussions surrounding lifespan experiments as well as descriptions of strain characteristics, commonly lifespan is talked about in terms of strains being short- or long-lived. The exact measurement of the lifespan carries less interest for ageing scientists than the meaning of that measurement. Considering that this distinction is important when we look at lifespan data, it would follow that the distinction may also be important to the model. Because of this, two more training datasets were produced by categorising the lifespan proxy datasets into ‘short’, ‘average’, and ‘long’ lived.

The lifespan data was categorised using R by binning the proxy data into the three categories around the average of the dataset. Initially, it was considered that the lifespans could be binned around the JB22 standard lab strain, but this produced skewed training data which would not have been appropriate for model building. As well as this, the JB22 strain does not necessarily have an ‘average’ lifespan: according to this wild type data set it has either an ‘average’ or a ‘short’ one with a mean on the shorter end of the ‘average’ bin. While JB22 is the appropriate lifespan standard control for lifespans of deletions in the JB22 strain such as the deletion library, it does not follow that it is the appropriate standard for ‘average’ lifespan in fission yeast as a species.

To achieve a more appropriate average, with specific relevance to this dataset, the average and range of the dataset was used to define three equally spaced bins into which the datapoints were sorted. The data produced is slightly uneven, with 19 short, 25 average, and 12 long in the mean lifespan proxy dataset, but this spread is representative of the skew towards short-lived strains we can see in the continuous lifespan dataset (figure 4B)

For the phenotype data, three training datasets of phenotypes were produced:

1. Mean phenotype for the 55 strains with three repeats (55 incidences)
2. Individual inputs of each repeat for the 55 strains with three repeats (165 incidences)
3. Individual inputs for all strains where 55 strains have three repeats and JB942 has two repeats (167 incidences)

These datasets consisted of continuous colony size ratio data which can be noisy and arguably is not the best input for this model. If we keep in mind the way in which we usually talk about colony size phenotypes in yeast, the language we use is often categorical. For this assay, it is usual to discuss results in terms of smaller colonies indicating a sensitive phenotype and larger colonies indicating a resistant phenotype, without too much weight being given to exact measurements and numbers. Because of this, three more training datasets were produced, identical to the original three but this time with the data categorised into sensitive, resistant or no change.

These training datasets are the final output which will be used to train the models detailed in chapter 2.

# 2 Developing Models to Predict Lifespan from Simple Phenotypes and Identify Ageing-Associated Processes in Yeast

## 2.1 Introduction

### 2.1.1 The rise of machine learning in ageing research

One way researchers can attempt to tackle overarching biological questions is the use of machine learning. In almost a complete U-turn from studying the minutiae of how individual biological pathways are involved in ageing, machine learning can be used as a kind of top-down method to look at ageing as a whole. Machine learning and artificially intelligent models are used to search data for patterns which are not accessible by traditional analysis methods, as such they can be applied to ageing to find connections in data without needing the exact biological mechanisms.

The use of machine learning models in ageing has gained huge popularity in recent years. Since the inception of the DNA methylation ageing prediction model, the Horvath Clock, in 2013, the use of machine learning to predict ageing or lifespan has snowballed. The attraction is clear, if we can predict ageing and lifespans through these models, we can simultaneously create

useful tools and identify what data we should be looking at for the mechanisms (Horvath, 2013).

With the Horvath Clock in particular, it has use as both a predictive tool and as evidence that DNA methylation was directly involved in the ageing process, since it was possible to predict lifespan from DNA methylation. The Horvath Clock also highlighted another reason machine learning is so popular within ageing, in fact within biology as a whole: it can make new use of pre-existing datasets. The model leveraged large pre-existing datasets of 8000 DNA methylation samples which allowed him to identify 353 CpG regions that could be used to predict chronological age (Horvath, 2013).

This early successful attempt at ageing prediction relied on linear regression techniques, looking for linear relationships between DNA methylation and chronological age (Horvath, 2013), but the field of machine learning and AI has advanced at an unprecedented rate in the last 20 years and many more sophisticated ageing prediction models have been built. In 2021, the DeepMAge clock was published. This was a first attempt at using the more sophisticated modelling technique of neural networks to create a DNA methylation ageing clock (Galkin et al., 2021).

DeepMAge was able to more accurately predict chronological age from DNA methylation data than the Horvath Clock, and with the addition of further data it was also able to account for the effect of diseases on chronological age, including ovarian cancer and multiple sclerosis (Galkin et al., 2021). This is an important step in the modelling of ageing since accounting for how diseases affect the relationship between DNA methylation and chronological age means

that the model is beginning to gain insights into the differences between chronological age and biological age.

These clocks can also be put to use in other areas of ageing research. The CellAgeClock, currently in preprint, aims to harness the power of the CpG methylation clocks to measure ageing of human cells in vitro and facilitate anti-ageing drug development. In their paper, the authors demonstrate how CellAgeClock validated known anti-ageing drugs and discovered novel anti-ageing drugs when used as a measure of their pharmacological effects in vitro. The identified novel anti-ageing drugs were then validated in vivo, completing this accelerated drug development pipeline (Lujan et al., 2020).

With constant development into machine learning models in ageing research, the area presents an exciting opportunity to gain new insights into the ageing process. Novel applications of this technology have the potential to facilitate leaps forward in our understanding of ageing and ageing related diseases.

However, despite the many advantages of the use of machine learning in the ageing field, it is important to be careful of exactly how the modelling is constructed. When a lifespan or ageing machine learning model is trained, the ‘true’ training data used is most often chronological age. The performance of the ageing model is measured based on its ability to accurately predict biological age as close as possible to the ‘true’ chronological age. Since we would not consider all humans of the same chronological age to be suffering equally from the ageing process or ageing related diseases, this is potentially not the most biologically informative design. Arguably, the nuances between subjects of the same chronological age contain the most valuable biological information.

When the success of biological age predictions is measured by how closely they align to the measurable chronological age, models are actively trained to ignore the biological differences between people of the same chronological age as opposed to highlighting these insights. Any modelling results will be directly impacted by the ability of the training data to fully address the core scientific questions, therefore, it is important to ensure that training data are based upon as few assumptions as possible and that it directly addresses the research questions.

### 2.1.2 Linear regression

Linear regression is a popular first step in machine learning modelling due to its simplicity. At its core, linear regression investigates data for a linear relationship and then makes predictions based on this (Casson and Farmer, 2014). Because of this, linear regression requires there to be a linear relationship within the data to make any predictions and prediction accuracy is directly impacted by the strength of that linear relationship. Lasso (least absolute shrinkage and selection operator) regression is a popular kind of regularised linear regression, lasso regression still assumes a linear relationship between inputs but includes a penalty which shrinks coefficients that do not contribute to the prediction (Brownlee, 2021). It allows all the phenotypes to be inputted into the model without skewing the results as non-predictive phenotypes can have their coefficients shrunk. If the model deems it necessary, they can even be shrunk all the way to zero, effectively removing them from the model. In this way lasso regression can be thought of as an automatic feature selection process for linear regression which will make the models more reliable and allow them to perform better. In this way, LASSO regression can be used on the

training data in this model without concern that any potentially non-predictive phenotypes are preventing the model from building.

### 2.1.3 Random Forest

Decision tree classifier models are the classic initial step for predicting categorised data which are built upon to create random forest models. A decision tree is an intuitive form of classification where the data is put through a set by step process, where each feature contributes to a yes or no decision to of the data belonging to a specific classification. In such data, each step on the decision tree would contribute to an overall decision of the data belonging to a category (Song and Lu, 2015). As it trains, the model will also weight each feature depending on how much that feature should contribute to the overall decision. Random forest models build on this concept by creating a collection of decision trees which all work together on the same decision, arrived at by committee. This means that one, or even multiple, individual trees can be wrong but overall, the committee decision is still correct. In this way, random forest can be far more accurate than a decision tree and can unearth connections in the data missed by them (Yiu, 2019). Since random forest is a classification model, it is ideal for predicting the categorised lifespan data. It is able to do this most straightforwardly from the categorised phenotype data, but predicting from continuous phenotype data is also possible.

### 2.1.4 Neural Networks

Neural networks are a highly sophisticated form of machine learning model which are loosely based on the structure of the human brain (Choi et al., 2020, Schmidhuber, 2015). They are structured as layers of nodes where connections are built between the data within each layer. These weighted connections are

improved with each pass of the model creating the concept of network learning. The networks have architecture of an input layer with a number of nodes defined by the shape of the training data, an output layer with a number of nodes defined by the shape of the prediction data and any number of hidden layers with any number of hidden nodes.

At each layer, an activation function is used to make connections in the data, in this work the activation function of rectified linear unit (ReLU) was used. ReLU is a piecewise linear function which outputs the input directly if it is positive or will output zero (Brownlee, 2020a). This makes it an almost linear function as it works in a linear manner for any input greater than 0, but it is non-linear as it reverts any negative input to 0. It is a highly popular activation function for most types of neural networks due to its computational simplicity, linear behaviour, and its ability to output zero unlike some other non-linear functions such as sigmoid functions.

Neural networks train by repeated passes through the network, during each pass the attributes and weights of the network are fractionally changed to improve prediction by reducing loss. Each of these passes through the network is known as an epoch. Epochs are a hyperparameter which need tuning since too few epochs will mean the network never trains fully and too many epochs will mean the network overtrains. Optimisation algorithms (optimisers) change attributes of the network to reduce loss over each pass. For epoch optimisation in this work, two different optimisers were used to be compared: stochastic gradient descent and adaptive moment estimation.

Stochastic gradient descent (SGD) is one of the most commonly used optimisers for neural network regression problems, which utilises both gradient

descent and momentum algorithms. It is the fastest of the gradient descent optimisers but can run the risk of becoming stuck in local minima instead of steadily decreasing loss. Adaptive moment estimation (Adam) is also commonly used for regression problems and utilises momentum and adaptive learning rate algorithms. It is generally considered to optimise faster than SGD but can be computationally intensive, slowing the model down. However, for our small dataset there should be no issue with speed of the model (Doshi, 2019).

To further decrease the network's error rate and the likelihood of overfitting it is necessary to optimise the learning rate. The learning rate, also known as the step size, is the amount by which the weights can be updated by the optimiser during each epoch. It is usually a small number between 0 and 1, and can be optimised anywhere within this to create a smooth loss and error reduction over the epochs (Brownlee, 2020c). A combination of tuning all the hyperparameters of a neural network can help to achieve high performance and accurate predictions.

### 2.1.5 Feature Selection

Once a model has been built, or during the building process, feature selection can be used to eliminate non-predictive features which reduces the computational intensity and often the performance of the model (Brownlee, 2019). In this work, feature selection would also identify the most predictive phenotypes and therefore the phenotypes most involved in cellular ageing and lifespan. One of the most popular forms of feature selection is recursive feature elimination (RFE) (Brownlee, 2020b). RFE functions by repeatedly training the model and recording the predictive importance of each feature. It then uses these feature importances to select the least predictive feature in the model and remove it. This process is repeated until a defined number of features or model

accuracy has been reached. RFE can be used on both linear regression and random forest models but not neural networks. The hidden layers of a neural network have been historically treated as a ‘black box’ in machine learning whose mechanisms are not to be understood. However, with the upsurge in neural network use, feature selection methods are beginning to be developed but are not yet widely used (Luíza da Costa et al., 2021, Figueroa Barraza et al., 2021).

## 2.2 Methodology

To create a model for predicting lifespan, there are several different avenues to explore. Since the generated training data can be both continuous and categorical in nature, both linear and classification machine learning modelling options can be explored. Starting from initial linear regression models, this chapter will go on to explore random forest classifier models and neural networks before addressing the question of predictive phenotype identification using feature selection.

One key advantage of machine learning techniques is that it often provides researchers with the opportunity to repurpose already existing datasets. This saves the time and money necessary for new lab-based experiments and helps researchers to make the fullest possible use of the data they have collected. Breathing new life into old data is a common application for machine learning and in this work, it was possible to make use of a pre-existing dataset of yeast phenotypes.

### 2.2.1 Colony size phenotype data for wild type strains

The dataset was collected from a high-throughput colony size assay showing how the growth of the collection of 161 wild type strains from the Jeffares collection alters under 82 different stress conditions. The dataset was available within the lab and was generated by Gorjan Stojanovski (2018, UCL, personal communication).

The strains were randomised and pinned in triplicate across two YES PlusPlates (SINGER Instruments) in 384 format using a RoToR HDA robot (SINGER Instruments). This 384 format also contained a grid of 96 JB22 control strains across the plates to allow for positional bias to be corrected. In addition

to the JB22 control grid, the plates also contained random strains along the bottom and right edges to allow for correction of edge bias, as well as blank positions used for ease of identification.

These plates were pinned on to stress plates for growth using a RoToR HDA robot. The stress conditions used in the experiment can be found in table 1. Stress plates were based in either YES or EMM media and a plain YES or EMM plate was used as the control plate for each experiment, respectively. For alternative carbon/nitrogen sources, glucose/nitrogen in the YES/EMM media was replaced with the alternative carbon/nitrogen source.

Exhausted media stress plates were prepared by growing the respective strain in YES liquid media for either 2 or 7 days at 32°C. The cells were removed from the media by centrifugation (5min at 500g) and filtering the liquid through a 0.22µm filter. 3% glucose was added back into the broth and 3x agar was added in a 2:1 ratio before the media was poured into solid plates. The strains used to exhaust the media differ in growth rate and species with *S. cerevisiae* S288C and *S. pombe* strains JB762 (fast growing), JB889 (slow growing), JB1197 (similar to JB22) and JB22 (standard lab strain).

Pinned stress plates were incubated for 2 days at 32°C before imaging.

Table 1: Summary of stress conditions used in the phenotype screen of the wild type strains.

| Condition Type | Compounds | Concentration |
|---|---|---|
| **Alternative Nitrogen Sources** | Glycine | 20mM |
| | Isoleucine | 20mM |
| | Lysine | 20mM |
| | Proline | 20mM |
| | Arginine | 20mM |
| | Serine | 20mM |
| | Aspartate | 20mM |
| | Glycine + Isoleucine | 20mM + 20mM |
| | Proline + Lysine | 20mM + 20mM |
| | Aspartate + Serine | 20mM + 20mM |
| | Arginine + Serine | 20mM + 20mM |
| | Aspartate + Lysine | 20mM + 20mM |

| | | |
|---|---|---|
| **Alternative Carbon Sources** | Galactose | 2% |
| | Sucrose | 2% |
| | Maltose | 2% |
| | Fructose | 2% |
| | Glycerol + Glucose | 2% + 0.01% |
| | Galactose + Glycerol | 2% + 3% |
| | Maltose + Sucrose | 2% + 3% |
| | Ethanol | 2/10 % |
| | Glycerol + NaAc | 2% + 2.5g/l |
| **Salt Stressors** | LiCl | 5mM |
| | NaCl | 150mM |
| | $MgCl_2$ | 200mM |
| | KCl | 0.6M |
| | LiCl + NaCl | 5mM + 150mM |
| | LiCl + $MgCl_2$ | 5mM + 200mM |
| | NaCl + $MgCl_2$ | 150mM + 200mM |
| | KCl + $MgCl_2$ | 0.6M + 200mM |
| **Oxidative Stressors** | $H_2O_2$ | 2/3/4.5/6mM |
| | Oligomycin | 250/500/1000µg/l |
| | Antimycin | 250/500/1000µg/l |
| | *tert*-Butyl hydroperoxide (TBH) | 1/1.5mM |
| **Exhausted Media** | *S. cerevisiae* | 2/7days |
| | JB22 | 2/7days |
| | JB769 | 2/7days |
| | JB889 | 2/7days |
| | JB1197 | 2/7days |
| **Additional Stressors** | Hydroxyurea (HU) | 10mM |
| | Methyl methanesulfanoate (MMS) | 0.0025%/0.0075% |
| | Caffeine | 10mM |
| | Rapamycin | 100ng/ml |
| | Calcofluor | 2/10µg/ml |
| | $NaN_3$ | 0.00025% |
| | LiCl + Calcofluor | 5mM + 2/10µg/ml |
| | NaCl + Calcofluor | 150mM + 2/10µg/ml |
| | $MgCl_2$ + Calcofluor | 200mM + 2/10µg/ml |
| | HU + Calcofluor | 10mM + 10µg/ml |
| | HU + MMS | 10mM + 0.0025%/0.0075% |
| | HU + LiCl | 10mM + 5mM |
| | NaCl + MMS | 150mM + 0.0025%/0.0075% |
| | Caffeine + Rapamycin | 10mM + 100ng/ml |
| | Caffeine + LiCl | 10mM + 5mM |
| | Maltose + Rapamycin | 3% + 100ng/ml |
| | Sucrose + Rapamycin | 3% + 100ng/ml |
| | Glycerol + MMS | 3% + 0.0075% |
| | Glycerol + NaAc + Arginine | 2% + 2.5g/l + 20mM |
| | Antimycin + Arginine | 10x + 5.7mM |
| | EMM | Low agar |

Once grown, stress plates were imaged using a conventional scanner and a custom Unix script within the lab to crop. Solid media growth was measured by

colony size, determined by number of pixels. Quantification was achieved through the R package *gitter* (Wagih and Parts, 2014), similarly to the high-throughput lifespan method. After quantification, the colony size data was normalised for positional bias using each plate's JB22 grid. The grid was used to interpolate a 'growth' value for each position on the plate and all colony sizes were divided by their positional 'growth' values. Quality control was then performed to remove all colony sizes with a circularity of <0.8 or >1.1, and all those with pixel sizes <100, to remove data from colonies with abnormal growth, and bubbles within the media. Colony size values were then normalised for plate differences by dividing by the plate median. The final output of this dataset is the colony size ratio between the stress plate and the control plate, with >1 being a larger colony than the control and <1 being a smaller colony than the control.

As is expected with repurposing datasets, the data required trimming and formatting to suit the needs of this model. After this process the phenotype data consisted of 76 phenotypes for all 57 strains of the wild type yeast used in this research. 56 of these strains had three repeat readings and one strain (JB942) had two repeat readings.

These datasets consisted of continuous colony size ratio data which was converted into ternary encoding using a median phenotypic value threshold: phenotypes showing a reduction of ≥10% on the phenotypic score were coded as -1, those showing an increase of ≥10% were coded as +1, and the weaker phenotypes in between were coded as 0.

### 2.2.2 Correlations and clustering

Pairwise correlations and heatmapping were performed in R using custom scripts. Heatmapping made use of the R package *heatmap2* and used Euclidean distance.

### 2.2.3 Linear regression

Linear regression was performed by a custom script in python version 3 using *train_test_split*, and lasso regression from scikit learn (Pedregosa et al., 2011). Training datasets can be found summarised in table 2 and from these, training and test data was defined using *train_test_split* at 80/20.

Models were assessed using root mean squared error and r-squared score metrics as well as plotting lifespans predicted by the model against the actual recorded values for both seen training data and unseen test data.

Table 2: Summary of training datasets used for linear regression models.

Including descriptions of which lifespan and phenotype datasets were used to create the training dataset and how many instances there are within it.

| | **Lifespan Data** | **Phenotype Data** | **Training Dataset Instances** |
|---|---|---|---|
| **LR1** | Mean lifespan (days) for the 56 strains (56 incidences) | Mean phenotype (colony size) for the 55 strains with three repeats (55 incidences) | 55 |
| **LR2** | Lifespan (days) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (colony size) as individual inputs of each repeat for the 55 strains with three repeats (165 incidences) | 165 |
| **LR3** | Lifespan (days) as individual inputs of each | Phenotype (colony size) as individual inputs for all strains | 167 |

| | | | |
|---|---|---|---|
| | repeat for the 56 strains (168 incidences) | where 56 strains have three repeats and JB942 has two repeats (167 incidences) | |
| **LR4** | Mean lifespan (days) for the 56 strains (56 incidences) | Mean phenotype (categorised) for the 55 strains with three repeats (55 incidences) | 55 |
| **LR5** | Lifespan (days) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (categorised) as individual inputs of each repeat for the 55 strains with three repeats (165 incidences) | 165 |
| **LR6** | Lifespan (days) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (categorised) as individual inputs for all strains where 56 strains have three repeats and JB942 has two repeats (167 incidences) | 170 |

### 2.2.4 Random forest

Random forest was performed by a custom script in python version 3 using *RandomForestClassifier* from scikitlearn (Pedregosa et al., 2011). Training datasets can be found summarised in table 3 and from these, training and test data was defined using *train_test_split* at 80/20.

Hyperparameters of *max_features* and *n_estimators* were tuned for the model using a 10-fold cross validated grid search. *max_features* defines the number of features which should be considered at each split, there are 77 features

available within the training data and at each split, the number of features defined by *max_features* will be randomly selected from these 77. n_estimators defines the number of trees in the forest. For the grid search, the models are built on a range of *max_features* and *n_estimators* and the most parameters producing the most accurate model were chosen. The *max_features* searched were 7, 17, 27, 37, 47, 57, 67, and 77 and the *n_estimators* searched were 30, 40, 50, 100, 200, 300, 400 and 500.

Models were assessed using the out of the bag error score and the accuracy score metrics as well as visualising accuracy using confusion matrices.

Table 3: Summary of training datasets used for random forest models.

Including descriptions of which lifespan and phenotype datasets were used to create the training dataset and how many instances there are within it.

| | **Lifespan Data** | **Phenotype Data** | **Training Dataset Instances** |
|---|---|---|---|
| **RF1** | Mean lifespan (categorised) for the 56 strains (56 incidences) | Mean phenotype (colony size) for the 55 strains with three repeats (55 incidences) | 55 |
| **RF2** | Lifespan (categorised) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (colony size) as individual inputs of each repeat for the 55 strains with three repeats (165 incidences) | 165 |
| **RF3** | Lifespan (categorised) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (colony size) as individual inputs for all strains where 56 strains have three repeats and JB942 has two repeats (167 incidences) | 167 |

| | | | |
|---|---|---|---|
| **RF4** | Mean lifespan (categorised) for the 56 strains (56 incidences) | Mean phenotype (categorised) for the 55 strains with three repeats (55 incidences) | 55 |
| **RF5** | Lifespan (categorised) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (categorised) as individual inputs of each repeat for the 55 strains with three repeats (165 incidences) | 165 |
| **RF6** | Lifespan (categorised) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (categorised) as individual inputs for all strains where 56 strains have three repeats and JB942 has two repeats (167 incidences) | 167 |

### 2.2.5 Neural networks

All neural networks were performed by a custom script in python version 3 using *tensorflow* and *keras*. The architecture for the neural network was chosen based on the constraints of the shape of the training data and the convention that hidden layer size should be between input and output layer size. The input layer has 77 nodes to correspond with the 77 features of X and the output layer has 1 node corresponding to the single feature of y. The convention for smaller datasets and simpler problems is to use only one hidden layer with a number of nodes between the input and output, after testing a range, 40 nodes was chosen. Sigmoid, linear and rectified linear unit (ReLU) activation functions were initially tested in combinations with a network only using ReLU activation performing best.

The optimisers adaptive moment estimation (Adam) and stochastic gradient descent (SGD) were used to build networks at a range of learning rates (0.0001, 0.0003, 0.0005, 0.0007 and 0.001) and the number of epochs was optimised on a model-by-model basis. For training neural networks, training datasets using continuous lifespan data and both continuous and categorised phenotype data were developed, summarised in table 4. From these, training and test data was defined using *train_test_split* from *scikitlearn* at 80/20. Models were assessed using root mean squared error.

Table 4: Summary of training datasets used for neural network models.

Including descriptions of which lifespan and phenotype datasets were used to create the training dataset and how many instances there are within it.

| | **Lifespan Data** | **Phenotype Data** | **Training Dataset Instances** |
|---|---|---|---|
| **NN1** | Mean lifespan (days) for the 56 strains (56 incidences) | Mean phenotype (colony size) for the 55 strains with three repeats (55 incidences) | 55 |
| **NN2** | Lifespan (days) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (colony size) as individual inputs of each repeat for the 55 strains with three repeats (165 incidences) | 165 |
| **NN3** | Lifespan (days) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (colony size) as individual inputs for all strains where 56 strains have three repeats and JB942 has two repeats (167 incidences) | 167 |

| | | | |
|---|---|---|---|
| **NN4** | Mean lifespan (days) for the 56 strains (56 incidences) | Mean phenotype (categorised) for the 55 strains with three repeats (55 incidences) | 55 |
| **NN5** | Lifespan (days) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (categorised) as individual inputs of each repeat for the 55 strains with three repeats (165 incidences) | 165 |
| **NN6** | Lifespan (days) as individual inputs of each repeat for the 56 strains (168 incidences) | Phenotype (categorised) as individual inputs for all strains where 56 strains have three repeats and JB942 has two repeats (167 incidences) | 170 |

### 2.2.6 Feature selection

Feature elimination was performed by a custom script in python version 3 using k-fold cross validation and *tensorflow*. The script was designed to allow for building up features one by one and selecting the feature which created the most accurate model at each step. The training data, architecture and hyperparameters of the neural networks used in feature selection was identical to that of the best neural network from 2.3.4 Neural networks (NN1, optimiser = Adam, learning rate = 0.0005), except the input layer which was changed to match the shape of the data each time.

Initially, k-fold cross validation was used to split the training data into 10 training/test datasets, where the test data is a new subset of the full dataset

each time. These subsets were then used to train neural networks on each of the single features. The root mean squared error of the test predictions for each of the ten models for each feature was averaged and the feature with the lowest average root mean squared error was chosen as the ‘best feature’. This feature was then added to a dataset of ‘chosen’ features and the remaining features created a dataset of ‘leftover’ features.

Next, one feature from the ‘leftover’ features was added to the ‘chosen’ features dataset at a time, creating datasets of the ‘chosen’ feature plus one of the ‘leftover’ features for each ‘leftover’ feature. These datasets were again split into 10 training/test datasets using k-fold cross validation to train neural networks. Again, the scores from the cross validation were averaged and the ‘best feature’ selected as the feature from the ‘leftover’ features in the model with the lowest average root mean squared error. This feature was then removed from the ‘leftover’ features dataset and added to the ‘chosen’ features dataset.

The process was repeated until all the features had been selected and added to the ‘chosen’ dataset, at each point the best feature was recorded with their rank, feature name, and average root mean squared error of the 10-fold cross validated networks. In this way it was possible to select the most predictive phenotypes within the neural network one by one as well as record how adding features affected the model’s root mean squared error.

## 2.3 Results

### 2.3.1 Correlations

Before building machine learning models, it was necessary to search the lifespan and phenotype data for any structure which is visible without machine learning. This would help inform the selection and design of the machine learning models.

Initially, the data was searched for any simple pairwise correlations between chronological lifespan (proxy) and each of the simple phenotypes. The pairwise correlation plots (figure 7) show that no single phenotype has any significant correlation with the lifespan data, but there is some structure which suggests a more complicated non-linear relationship.

To further search for any structure within the data, clustering was used, to cluster the phenotypes against the lifespans in order from short- to long-lived. Clustering the phenotypes against the chronologically ranked lifespans (figure 8) produced horizontal structure in the heatmap showing that some phenotype datasets are similar to each other and can cluster together which is to be expected. However, the heatmap doesn’t show any horizontal structure. This shows that there is no evident structure in the data even when considering the relationship between lifespan and multiple phenotypes.

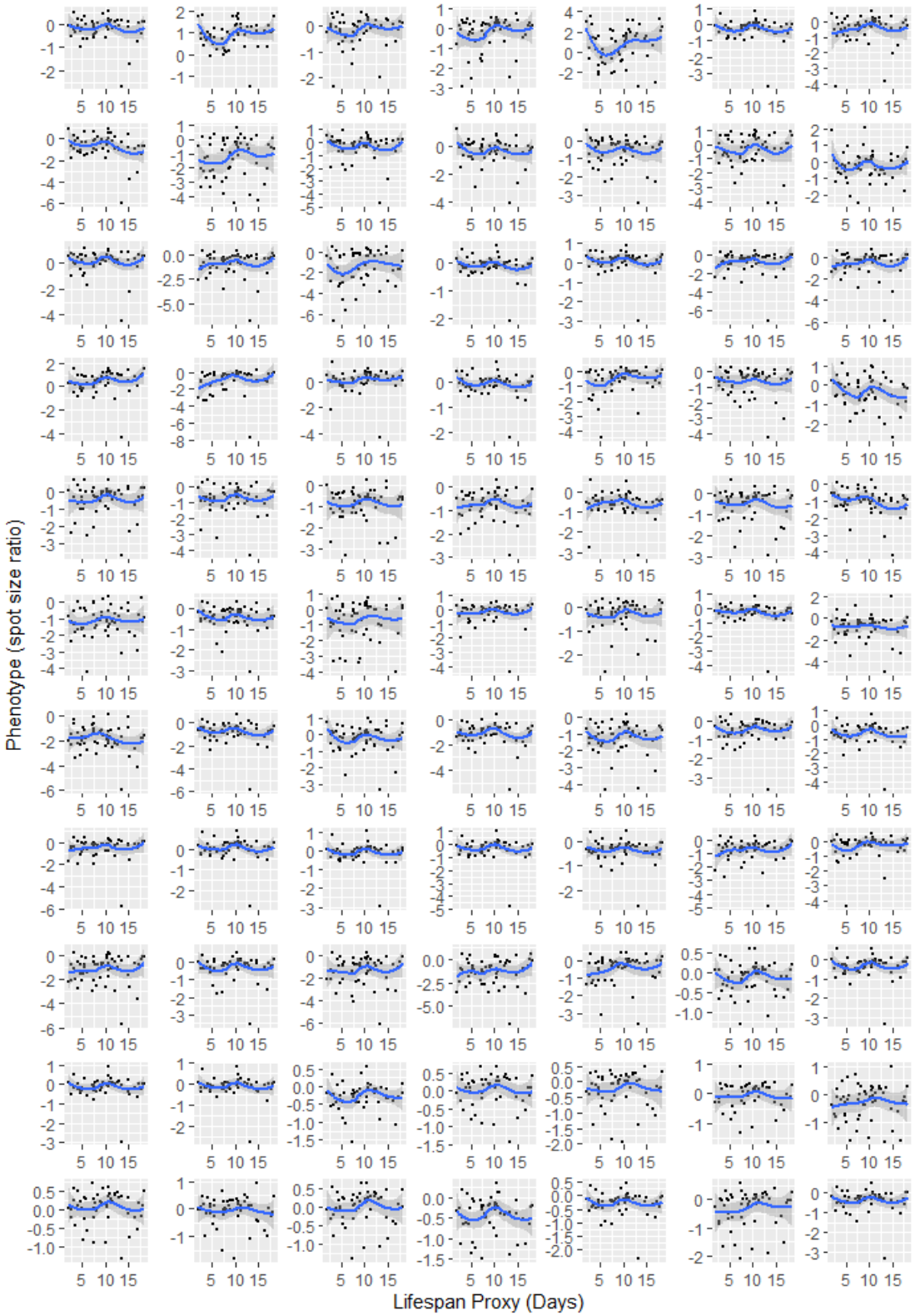



Figure 7: Pairwise plots of mean lifespan proxy data against mean phenotype data.

The scatter plots with fitted smooth spline show there are no obvious correlations between the mean lifespan data and the mean phenotype data for any phenotype measured, across the 56 available strains.

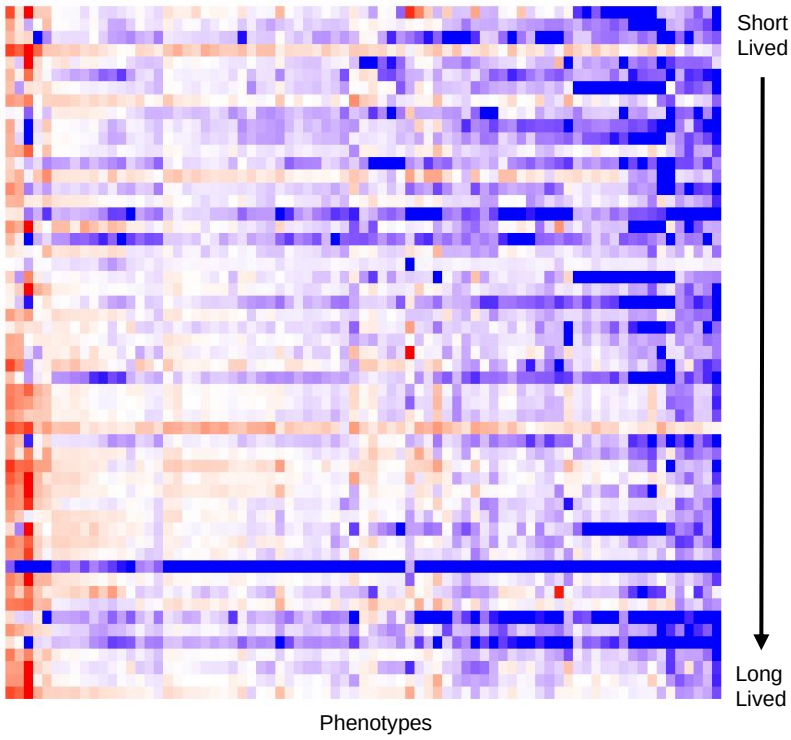



Figure 8: Heatmap of clustered phenotype datasets against lifespan data.

The heatmap shows horizontal structure with clusters of similar phenotypes but no vertical structure of clusters found within strains of particular lifespans.

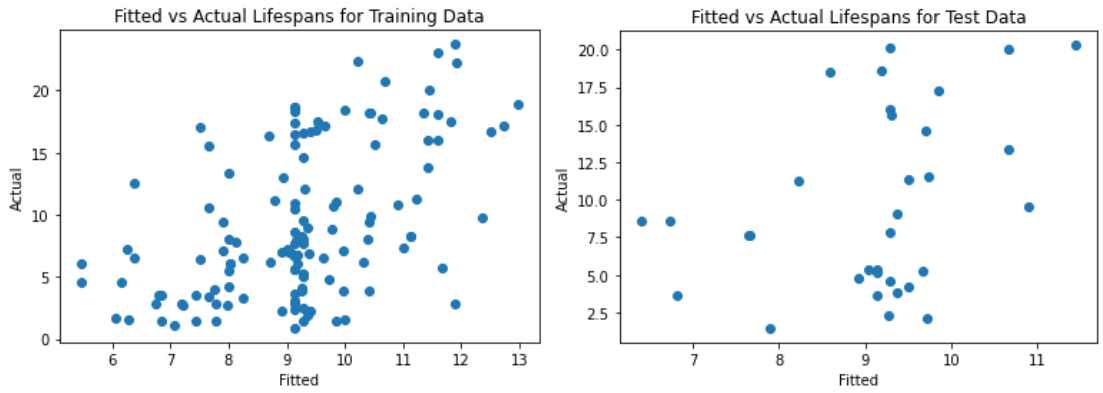
### 2.3.2 Linear regression

Initially, I used LASSO regression to ensure that there was no artificial signal from the pairing of a particular lifespan repeat with a particular phenotype repeat in the data where the lifespans and phenotypes were used as individual inputs (training datasets LR2, 3, 5 and 6).

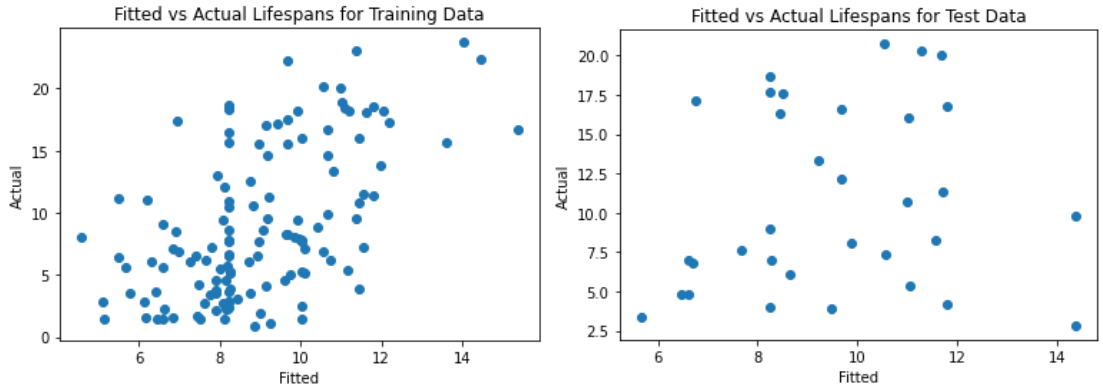
For this, training dataset LR5, a dataset with individual inputs of all strains with three repeats of continuous lifespan and categorised phenotypes, three combinations were created: each lifespan repeat paired with each phenotype repeat. The three data combinations were used to produce LASSO regression models. As shown by the almost identical metrics (table 4) and scatterplots (figure 9), there was no difference between models built with the different combinations.

The metrics in table 5 also show that these regression models had high root mean error and low r-squared scored for predicting both seen training and unseen test data, with unseen test data performing even worse. This means that none of the models successfully predicted lifespan from phenotypes.

A




B




C

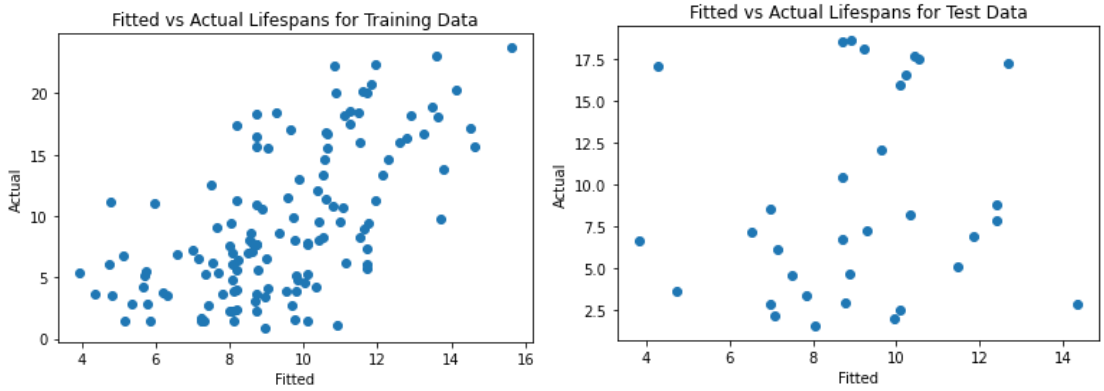



Figure 9: Scatterplots to show actual vs fitted data for training and testing of each data combination.

Plots show that the model trains similarly for combination 1(A), 2(B) and 3(C) of the training data with poor but better than random performance on the seen training data but no significant performance on the unseen test data.

Table 5: Metrics for training and testing LASSO regression models on data in the three different combinations for training dataset 5.

| Data Combination | Root Mean Squared Error | | R Squared Score | |
|---|---|---|---|---|
| | Train | Test | Train | Test |
| **1** | 5.37 | 5.55 | 0.20 | 0.09 |
| **2** | 5.17 | 5.94 | 0.25 | -0.09 |
| **3** | 4.88 | 6.00 | 0.34 | 0.06 |

The six training datasets (defined in table 2 in *2.2.2 Linear regression*) were then used to create LASSO regression models. The metrics of root mean squared error and r squared score can be found in table 6. These metrics showed that all the models had high root mean squared error and low r squared scores, meaning that they do not predict lifespan from simple phenotypes well enough to be considered successful and there is not a simple linear relationship in the data.

However, the models built on training datasets using the mean lifespan and phenotypes for both continuous (LR1) and categorised (LR4) phenotype data did clearly outperform those built on training datasets with individual inputs of the repeats, with much lower root mean squared error for predicting both seen training and unseen test data.

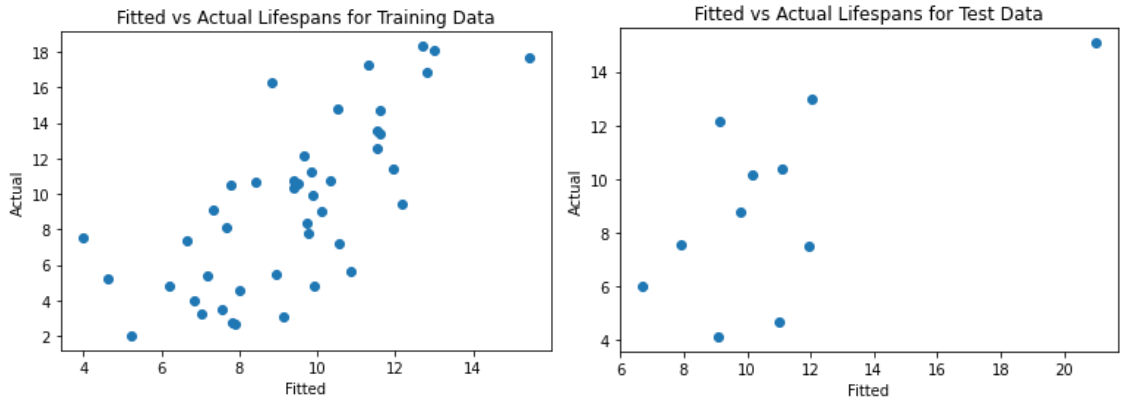
Table 6: Metrics for training and testing LASSO regression models on all 6 training datasets. The models made with training datasets LR1 and LR4 are the most accurate and these are highlighted in grey.

| Training Dataset | Root Mean Squared Error | | R Squared Score | |
|---|---|---|---|---|
| | Train | Test | Train | Test |
| **LR1** | 3.36 | 3.45 | 0.48 | -0.08 |
| **LR2** | 5.46 | 5.97 | 0.17 | -0.18 |
| **LR3** | 5.64 | 5.47 | 0.12 | 0.02 |
| **LR4** | 2.85 | 3.45 | 0.61 | -0.08 |

| | | | | |
|---|---|---|---|---|
| **LR5** | 5.37 | 5.55 | 0.20 | 0.09 |
| **LR6** | 5.11 | 6.02 | 0.23 | 0.10 |

The scatterplots in figure 10 show the lifespans fitted by the LASSO regression models built on training datasets LR1 (A) and 4 (B) vs the actual lifespans for seen training data and unseen test data. Both models performed better on the seen training data but still better than random on the unseen test data, despite the weak linear relationship shown by their low r-squared scores.

A




B

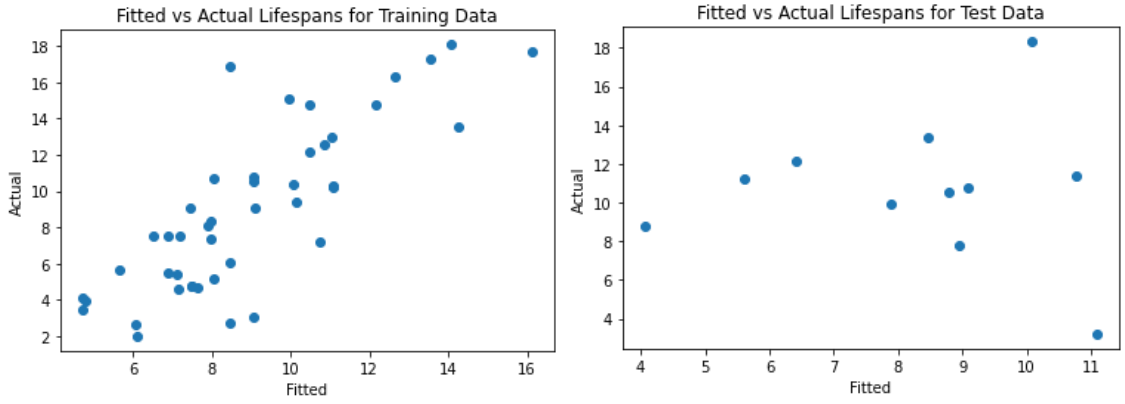



Figure 10: Scatterplots show actual vs fitted data for training and testing of training datasets LR1 and LR4.

The plots show the linear regression models for dataset 1(A) and dataset (4) have some correlation between the fitted and actual lifespans in both seen training data and unseen test data. Both datasets LR1 (colony size phenotype) and LR4 (categorised phenotype) show higher performance on seen data but they perform better than random on unseen data.

### 2.3.3 Random forest

Moving on from linear regression models, the training datasets using categorised lifespan data (defined in table 3 in *2.2.3 Random forest*) were used to train random forest models. Initially the best hyperparameters of *max_features* and *n_estimators* were tuned for the model using a 10-fold cross validated grid search. The best parameters selected by this grid search for each training dataset can be found in table 7, and these parameters were used in the final random forest models for each training dataset.

Table 7: Summary of best hyperparameters selected by cross validation parameter search for the random forest models for each training dataset.

| Training Dataset | *max_features* | *n_estimators* |
|---|---|---|
| **RF1** | 47 | 40 |
| **RF2** | 17 | 30 |
| **RF3** | 7 | 40 |
| **RF4** | 57 | 50 |
| **RF5** | 67 | 50 |
| **RF6** | 7 | 200 |

During training, the out of the bag error (OOB) score was calculated for each model and after testing on unseen data, the accuracy score was calculated for each model. These scores were used as metrics to define the success of the model. All the models predicted better than random with accuracy scores of above 33% and therefore were all able to successfully categorise lifespan based on phenotypes. There appeared to be little difference in error or accuracy between the models using continuous colony size phenotype data and those using categorised phenotypes, but the models built on training datasets using the mean lifespan and phenotype data (RF1 and RF4) were the most

successful models with accuracy of above 50%. The model built on RF4 (categorised mean phenotypes) was able to correctly categorise lifespan 64% of the time.

Table 8: Out of the bag error and accuracy score metrics for random forest models built on each training dataset. The models made with RF1 and RF4 are the most accurate and have the lowest OOB error and these are highlighted in grey.

| Training Dataset | OOB Score | Accuracy Score |
|---|---|---|
| **RF1** | 0.39 | 55% |
| **RF2** | 0.56 | 42% |
| **RF3** | 0.51 | 50% |
| **RF4** | 0.34 | 64% |
| **RF5** | 0.47 | 46% |
| **RF6** | 0.55 | 47% |

Figure 11 further analyses the models built on RF1 and RF4 using confusion matrices. Here, the model built on RF1 predicted 100% of the seen training data correctly (figure 11A) but correctly categorised only 6/11 of the unseen test data (figure 11B). In one case it misidentified, a long-lived strain as short and in another a short-lived strain as long which are complete opposite classes.

The model built on RF4 misclassified 2 of the seen training data (figure 11C) but went on to correctly classify 7/11 of the unseen test data (figure 11D). Its misclassifications were classifying a long-lived strain as average in 3/4 cases and misclassifying one average strain as short.

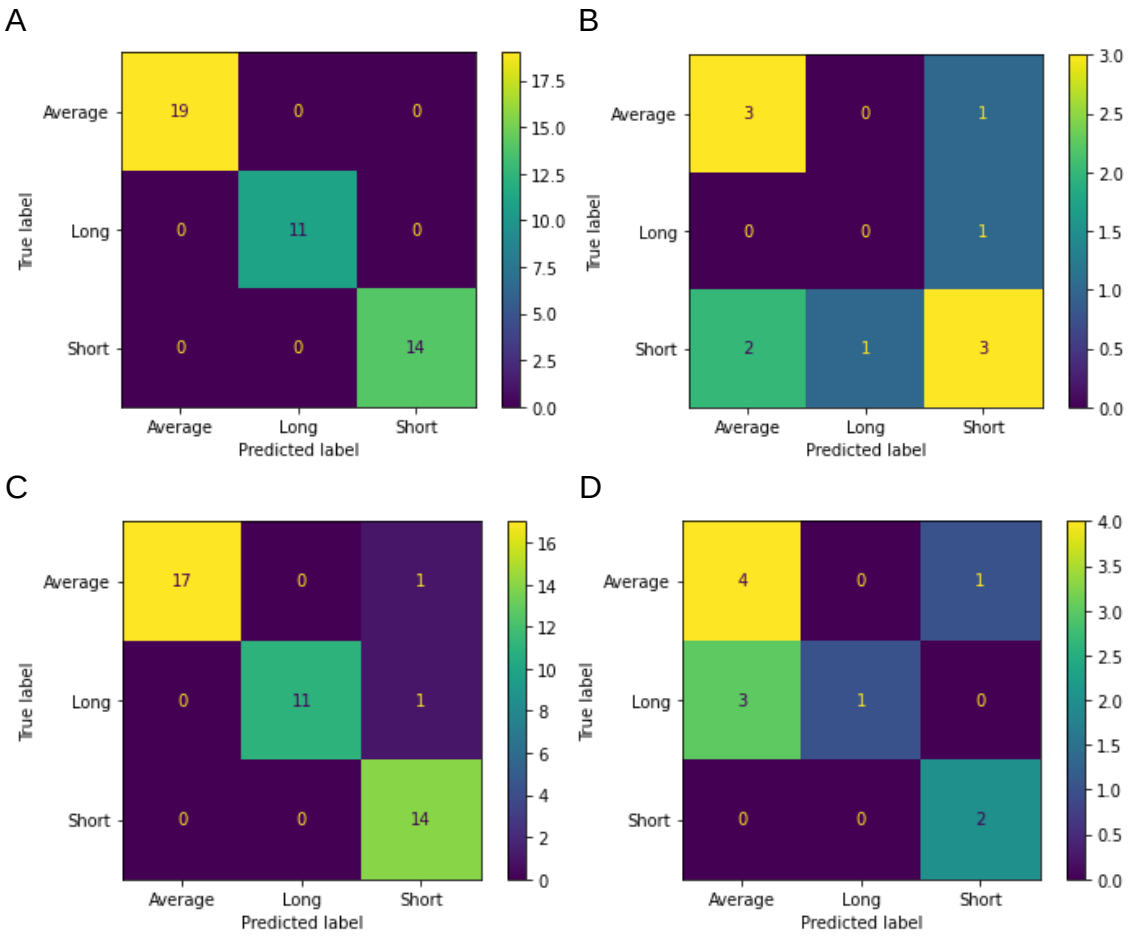



Figure 11: Confusion matrices for training and testing of random forest models on training datasets RF1 and RF4.

The model on training dataset RF1 (categorised mean lifespans and mean colony size phenotypes) shows high performance on the seen training data (A) and lower performance by still better than random performance on unseen test data (B). The model on training dataset RF4 (categorised mean lifespans and categorised mean colony size phenotypes) shows slightly less high performance in the seen training data than the model built on training dataset RF1 (C) but slightly higher performance on unseen test data than the model built on training dataset RF1 (D).

### 2.3.4 Neural networks

The neural networks were initially built to test the performance of each training dataset (defined in table 3 in *2.2.4 Neural networks*) across three learning rates and two different optimisers Adam (adaptive moment estimation) and SGD (stochastic gradient descent). The epochs were adjusted on a model-by-model basis to ensure that the model had enough passes to fully train but not enough to overfit, the epochs for each model were then recorded in tables 9 and 10.

The models were assessed based on the root mean squared error of the network, which shows how far predicted values fall from measured true values, The summaries of these models can be found in table 9 for models using the Adam optimiser and table 10 for models using the SGD optimiser.

During this initial modelling, most of the neural networks were able to predict lifespan from simple phenotypes but using both the Adam and the SGD optimisers, training dataset NN1 using a learning rate of 0.0005 performed best, with the lowest root mean squared error as highlighted in tables 9 and 10.

Table 9: Summary of neural networks trained on the training datasets with the use of the optimiser Adam and learning rates of 0.0001, 0.0005 and 0.001. The most accurate model with the least error is highlighted in grey.

| Training Dataset | Learning Rate | Epochs | Root Mean Squared Error |
|---|---|---|---|
| **NN1** | 0.0001 | 350 | 4.23 |
| | 0.0005 | 100 | 3.64 |
| | 0.001 | 50 | 4.03 |
| **NN2** | 0.0001 | 250 | 5.76 |
| | 0.0005 | 60 | 5.48 |
| | 0.001 | 25 | 5.64 |
| **NN3** | 0.0001 | 200 | 6.03 |
| | 0.0005 | 50 | 6.01 |
| | 0.001 | 20 | 6.59 |
| **NN4** | 0.0001 | 250 | 4.45 |
| | 0.0005 | 60 | 5.10 |
| | 0.001 | 25 | 5.54 |
| **NN5** | 0.0001 | 150 | 6.36 |
| | 0.0005 | 40 | 6.17 |
| | 0.001 | 20 | 6.62 |
| **NN6** | 0.0001 | 200 | 4.89 |
| | 0.0005 | 50 | 4.51 |

| | | | |
|---|---|---|---|
| | 0.001 | 20 | 4.98 |

Table 10: Summary of neural networks trained on the training datasets with the use of the optimiser SGD and learning rates of 0.0001, 0.0005 and 0.001. The most accurate model with the least error is highlighted in grey.

| Training Dataset | Learning Rate | Epochs | Root Mean Squared Error |
|---|---|---|---|
| **NN1** | 0.0001 | 200 | 4.19 |
| | 0.0005 | 50 | 3.52 |
| | 0.001 | 20 | 3.61 |
| **NN2** | 0.0001 | 150 | 6.11 |
| | 0.0005 | 40 | 6.09 |
| | 0.001 | 20 | 5.88 |
| **NN3** | 0.0001 | 150 | 5.90 |
| | 0.0005 | 40 | 5.79 |
| | 0.001 | 20 | 5.79 |
| **NN4** | 0.0001 | 200 | 5.49 |
| | 0.0005 | 60 | 5.41 |
| | 0.001 | 20 | 5.31 |
| **NN5** | 0.0001 | 150 | 6.31 |
| | 0.0005 | 40 | 6.13 |
| | 0.001 | 20 | 6.29 |
| **NN6** | 0.0001 | 150 | 6.41 |
| | 0.0005 | 40 | 6.30 |
| | 0.001 | 20 | 6.41 |

With networks for both optimisers performing best when trained on NN1 (continuous mean lifespan data and continuous mean phenotype data), the networks were fine-tuned by training with NN1 using Adam and SGD optimisers at learning rates of 0.0003, 0.0005 and 0.0007. The epochs were informed by the pervious networks but were still adjusted on a model-by-model basis to

ensure that the model had enough passes to fully train but not enough to overfit. Table 11 summarises these models, including the error metrics.

Table 11: Summary of fine-tuning neural networks trained on the training dataset NN1 with the use of the optimisers Adam and SGD and learning rates of 0.0003, 0.0005 and 0.0007. The most accurate model with the least error is highlighted in grey.

| | Learning Rate | Epochs | Root Mean Squared Error |
|---|---|---|---|
| **Adam** | 0.0003 | 100 | 2.84 |
| | 0.0005 | 100 | 2.83 |
| | 0.0007 | 100 | 3.13 |
| **SGD** | 0.0003 | 50 | 4.90 |
| | 0.0005 | 50 | 4.11 |
| | 0.0007 | 50 | 4.66 |

The most accurate network, with the lowest root mean squared error and mean squared error loss, is a network trained on the dataset NN1 using the optimiser Adam, at a learning rate of 0.0005 over 100 epochs. This model can predict lifespan to within 2.83 days.

To further assess these fine-tuning models, the loss and error curves and the fitted vs actual lifespans were plotted for each model. The loss and error curves for all models built with the Adam optimiser (figure 12A, C and E) and the models built with the SGD optimiser with a learning rate of 0.0003 and 0.0005 (figure 13A and C) show that the model trained smoothly. They also show that the number of epochs allows for the model to fully train without overfitting. An example of overfitting can be found in supplementary figure 2. The loss and error curves for the model built with the SGD optimiser with a learning rate of 0.0007 (figure 13E) does not have a smooth test data curve, suggesting that the model got stuck in local minima, and shows some evidence of overfitting.

The predicted lifespans vs actual lifespans scatterplots for all models (figure 12B, D, E, figure 13B, D and E) show good correlations and no evident overfitting. Figure 12D shows the predicted vs actual lifespans for the for the most accurate network. This scatterplot shows a strong correlation between predicted vs actual lifespans for both the seen training data and the unseen test data with a low spread of data and very few outliers.

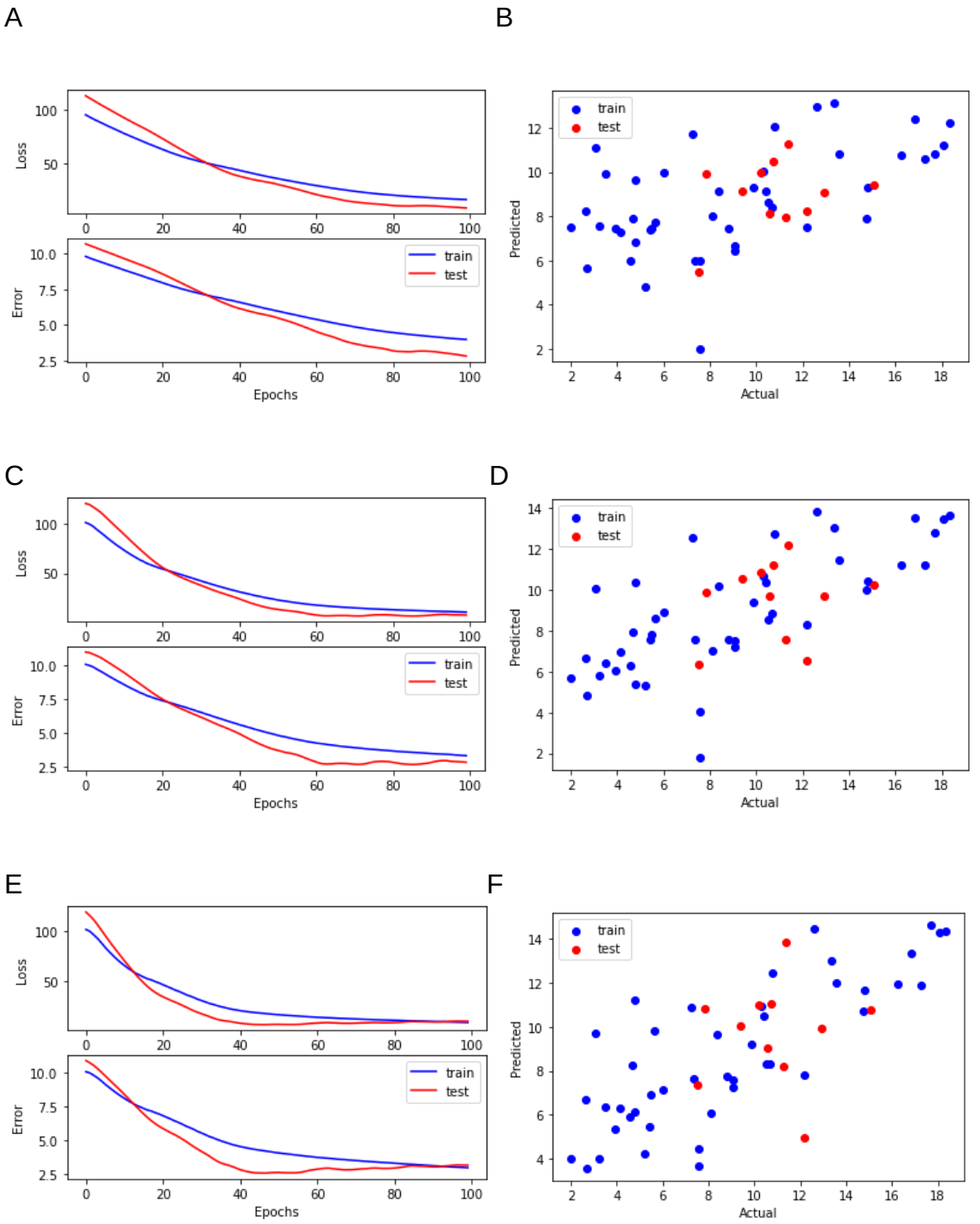



Figure 12: Graphs to show the learning rate fine tuning of a neural network built on training dataset NN1 using the Adam optimiser.

For a network built with a learning rate of 0.0003 the graphs show smooth loss and error curves (A) and predicted lifespans for seen training and unseen test data with a strong correlation to the actual lifespans (B) both without evidence of overfitting. For a network built with a learning rate of 0.0005 the graphs show smooth loss and error curves (C) and predicted lifespans for seen training and unseen test data with a strong correlation to the actual lifespans (D) both without evidence of overfitting. This network makes the most accurate predictions and has the lowest loss. For a network built with a learning rate of 0.0007 the graphs show smooth loss and error curves (E) and predicted lifespans for seen training and unseen test data with a strong correlation to the actual lifespans (F) both without evidence of overfitting.

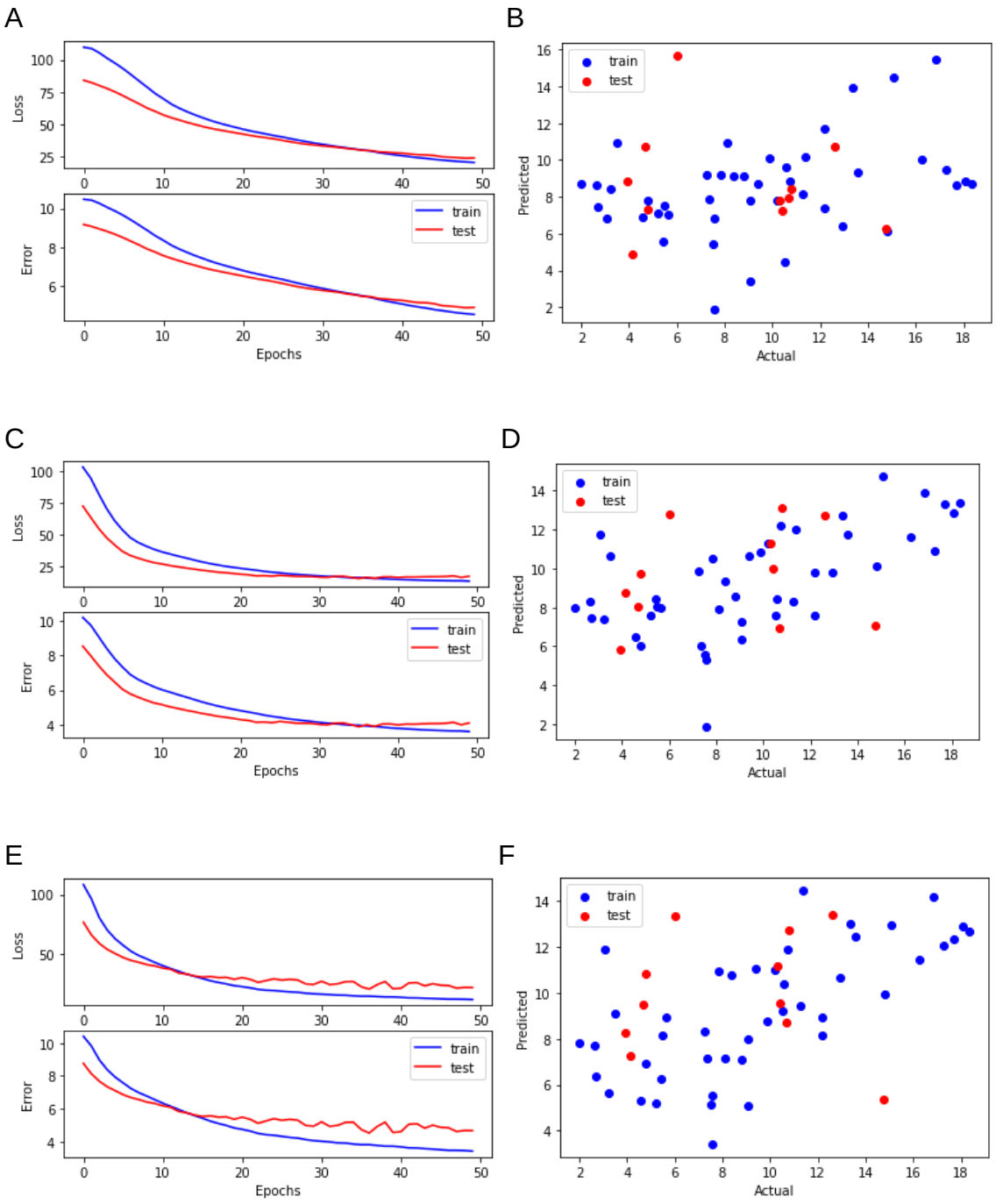



Figure 13: Graphs to show the learning rate fine tuning of a neural network built on training dataset NN1 using the SGD optimiser.

For a network built with a learning rate of 0.0003 the graphs show smooth loss and error curves (A) and predicted lifespans for seen training and unseen test data with a strong correlation to the actual lifespans (B) both without evidence of overfitting. For a network built with a learning rate of 0.0005 the graphs show smooth loss and error curves (C) and predicted lifespans for seen training and unseen test data with a strong correlation to the actual lifespans (D) both without evidence of overfitting. For a network built with a learning rate of 0.0007 the graphs show loss and error curves with some fluctuations suggesting that the model became stuck in local minima (E) and predicted lifespans for seen training and unseen test data with a strong correlation to the actual lifespans (F) both without evidence of overfitting.

### 2.3.5 Feature selection

The feature selection program creates an output file containing the selected feature name and the average root mean squared error (RMSE) of the predictions from unseen test data from the 10-fold cross validation networks for that feature for each rank. By plotting the rank against the RMSE, we can visualise how the number of features effects the model performance. Figure 14 shows that the RMSE initially decreases as the features are added to the model, showing an increase in prediction performance. After reaching the lowest RMSE at 17 features, adding more features to the model increases the RMSE, showing a decrease in performance.

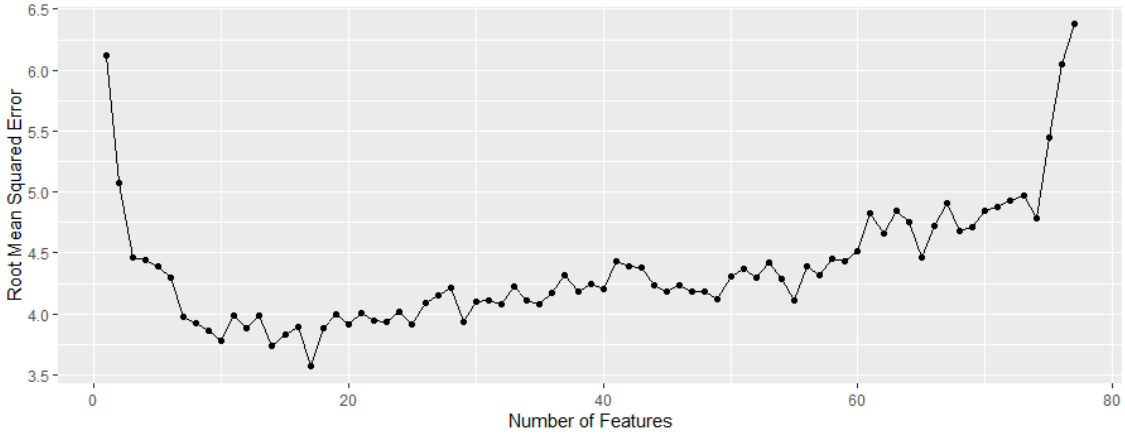



Figure 14: Model performance during feature selection.

The graph shows the average root mean squared error of the 10-fold cross validation of the best model at each step of feature selection. The RMSE declines, showing the models are increasing in accuracy until the lowest RMSE of 3.57 at 17 features. The RMSE then increases as more features are added to the model.

To investigate this in more detail, the highest ranked features can be investigated. Since 17 features were identified as the point at which the model stops improving in performance, the top ranked 17 features were subset as features of interest. Table 12 shows a summary of the top 17 features including the phenotype condition and the effects within the cell that the assay is

measuring under this condition. Some phenotype conditions are repeated where more than one of the variations (e.g. concentration) ranked within the top 17. This summary shows that phenotypes describing a wide range of stress, drug or nutrient responses are contributing to the lifespan prediction in the neural networks.

Table 12: Summary of the top 17 phenotype conditions ranked by the feature selection and the average root mean squared error (RMSE) of the 10-fold cross validation of the model at each step.

| Rank | Phenotype Condition | Effect | RMSE |
|---|---|---|---|
| **1** | Oligomycin | ATP synthase inhibition | 6.12 |
| **2** | Exhausted Media (2 days) | Low nutrients | 5.07 |
| **3** | Antimycin | Mitochondrial respiration inhibition | 4.46 |
| **4** | Serine | Supplemented amino acids | 4.45 |
| **5** | Exhausted Media (2 days) | Low nutrients | 4.40 |
| **6** | MMS | DNA damage | 4.10 |
| **7** | Proline | Supplemented amino acids | 3.98 |
| **8** | H2O2 | Oxidative stress | 3.92 |
| **9** | Caffeine and Rapamycin | TOR inhibition | 3.86 |
| **10** | Calcofluor | Cell wall integrity | 3.78 |
| **11** | Calcofluor and NaCl | Cell wall integrity and osmotic stress | 3.99 |
| **12** | Proline and Lysine | Supplemented amino acids | 3.89 |
| **13** | NaCl and MMS | Osmotic stress and DNA damage | 3.98 |
| **14** | EMM | Essential nutrients | 3.74 |
| **15** | LiCl | Translation inhibition | 3.84 |
| **16** | H2O2 | Oxidative stress | 3.89 |
| **17** | Isoleucine | Supplemented amino acids | 3.57 |

## 2.4 Discussion

### 2.4.1 Correlations

Initially, it was important to establish whether there were any evident relationships in the data. This would help to inform the choice of models used as well as the interpretation of any results. To begin that process, the mean lifespan training data was plotted against the mean phenotype data for each phenotype across the 56 strains available. The plots (figure 7) show that no single phenotype has any significant correlation with the lifespan data, this supports the development of a machine learning model to look for patterns in the data that are not otherwise evident. The lack of correlation across all the data also shows that none of the phenotypes should be considered more predictive than any others at this point and there is no need to introduce pre-emptive weighting of features into the models. The correlations do appear to have some non-linear structure which further supports the idea of a more complex relationship within the data.

To take this one step further, clustering can be used to create a heatmap to visualise any more hidden structures within the data. Using the heatmap seen in figure 8 we can look for any blocks of structure showing relationships between the mean lifespan dataset and the mean phenotype datasets across the 56 strains available. The clustering produced horizontal structure in the heatmap showing that some phenotype datasets are similar to each other and can cluster together which is to be expected. However, the heatmap does not show any horizontal structure. This outcome reveals that there is no evident structure in the data even when considering the relationship between lifespan and multiple

phenotypes and further supports the application of more complex machine learning models to search for structure and patterns within this dataset.

### 2.4.2 Linear regression

For a first attempt at creating a model to predict lifespan from simple phenotypes, linear regression was chosen. This is a relatively simple kind of modelling, closely related to the mathematics of pairwise correlations, making it the natural next step (IBM, 2023). The model would be able to determine if a combination of the linear relationships between the phenotypes and the lifespan data could be predictive. While the pairwise correlations suggested this will not be successful, it remained important to investigate the data in this way for linear relationships before moving on to more complex models.

Linear regression models are built on continuous data, making it an ideal choice for predicting the continuous lifespan data from the continuous, uncategorised phenotype datasets. However, alongside these, models were also built on training datasets using continuous lifespan data and categorised phenotype data. While this data is not traditionally as well suited to linear regression, it may provide more accurate biological insight since we usually describe phenotypes as sensitive or resistant.

Using *train_test_split* at 80/20 allowed for 80% of the data to be used as training data while 20% of it was withheld from the model as unseen test data. After training, this test data can be used to assess the model's ability to predict data it has never seen before and allow us to spot overfitting. After training, the model is asked to predict the lifespan of both the training dataset it has seen and the unseen test data and how closely it can predict these to the actual recorded lifespans is visualised in scatterplots to look for the correlation between the

actual lifespans and fitted (predicted) lifespans. The comparison between the seen and unseen data prediction accuracy allows us to check for overfitting, where a model has trained to predict the seen data very well but cannot predict unseen data. This occurs when a training process has resulted in the model learning only the specifics of the training dataset as opposed to an overarching connection in the data.

LASSO regression was chosen for the linear regression modelling in this work since it includes a penalty that shrinks coefficients which do not contribute to the prediction. It allows all the phenotypes to be inputted into the model without skewing the results as non-predictive phenotypes can have their coefficients shrunk. This was of particular advantage with our specific training data because we were unsure whether each phenotype could contribute to lifespan prediction.

Initially, LASSO models were built only on the training dataset LR5, a dataset with individual inputs of continuous lifespan data and categorised phenotype data. This dataset was altered to allow for three variants, one where each lifespan was paired to each phenotype. Building models with these three variants offered the opportunity to spot if there was any advantage or disadvantage to any of the pairings. Since all three models were virtually identical (table 5 and fig 9), we are able to conclude that there is no artificial signal from any of the pairings. This allows for further modelling to be free of the concern that the pairing of a lifespan repeat to a phenotype repeat could be affecting the outcome.

The six training datasets (defined in table 2 in *2.2.2 Linear regression*) were then used to create LASSO regression models. These models (table 6) were assessed by the use of root mean squared error and r-squared score metrics.

Root mean squared error is used in this way as a score of the prediction accuracy of the model as it describes how far predicted values lie from the true recorded values. R-squared is used as a measure of the strength of the linear relationship found in the data. By using root mean squared error, we were able to identify models built on LR1 and LR4 (based in mean lifespan and phenotype data) as much more accurate models. All the models had low or negative R-squared scores, showing weak to no linear relationship was found for the model to train on.

While the models built on LR1 and LR2 were able to predict lifespan from phenotypes better than random, both models performed better on predicting seen training data than on the unseen test data (fig 10), suggesting the models were overfitting slightly. As well as this, the models did not perform particularly well even on the seen training data which, in combination with the low r-squared scores, suggests that there is not a simple linear relationship within the data to build the model on.

As suggested by the lack of any visible correlations (fig 7), the linear model is not able to capture enough patterns within the data to predict lifespan from phenotypes in any meaningful way. However, the presence of some structure in the actual vs. fitted graphs for these models does suggest that there is a connection to find in this data using more sophisticated modelling techniques.

### 2.4.3 Random forest

As we could argue that the categorised data for both the lifespan and the phenotype could be more biologically intuitive, a model designed to predict on categorised data was the next choice to investigate the patterns in this data.

In the same way as the linear regression datasets, the random forest training datasets (defined in table 3 in *2.2.3 Random forest*) were subset 80/20 into seen training and unseen test data to facilitate analysis of the model's performance. The performance of the models was assessed using out of the bag error from training and an accuracy score for the predictions on unseen test data. The accuracy of the models was further visualised in confusion matrices, a table of the number of correctly categorised instances vs the incorrectly categorised ones. Overfitting can again be identified by the difference between performance on seen training data and unseen test data.

Out of the bag (OOB) error is a helpful metric when building random forest models since it computes during the training of the model without the need to reserve unseen testing data. It describes how much error the model makes when categorising yet unseen data within the training dataset. The accuracy score is a similar metric but this time on reserved unseen test data, showing the percentage of correctly categorised data in this group. When assessing our models, we were looking for the lowest OOB scores and the highest accuracy metrics.

Before building and assessing the models, a 10-fold cross validation grid search was used to establish the best values for *max_features* and *n_estimators*, the number of phenotypes to be used for a decision at each split and the number of trees in the model. Using 10-fold cross validation to do this allows for a highly accurate parameter search which makes the decision based on models built on 10 overlapping subsets of the data. This helps to remove any bias in the parameter selection and choose those most representative of the data.

The parameters used in the grid search (*max_features*: 7, 17, 27, 37, 47, 57, 67, and 77 and *n_estimators*: 30, 40, 50, 100, 200, 300, 400 and 500) were designed to cover a range so that the chosen parameter was not the highest or lowest option indicating that there was a better choice outside of the searched range. In no cases were the chosen parameters the highest options in the search, however 7 features were chosen twice, and 30 estimators were chosen once. While these are the lowest options, fewer than 7 features or fewer than 30 trees would be considered abnormally small for building a model of this kind, so the parameters were limited at this point.

Using the OOB and accuracy scores, while all the models predicted better than random (33%), it stood out that the models trained on RF1 and RF4 have the lowest error and the highest accuracy (table 8). This is similar to the performance seen in the linear regression models, where the datasets using mean lifespans and mean phenotypes performed better, suggesting that using the mean of these data is the best way to create meaningful predictions.

The models built on RF1 and RF4 both had an accuracy of over 50% (table 8), the confusion matrices were used for more in depth analysis of this. The model built on RF1 predicted 100% of the seen training data correctly (figure 11A) but correctly categorised only 6/11 of the unseen test data (figure 11B). In one case it misidentified, a long-lived strain as short and in another a short-lived strain as long which are complete opposite classes. The model built on RF4 misclassified 2 of the seen training data (figure 11C) but went on to correctly classify 7/11 of the unseen test data (figure 11D). Its misclassifications were classifying a long-lived strain as average in 3/4 cases and misclassifying one average strain as short.

Since the RF4 model had a better performance on unseen data and the RF1 model had a better performance on seen data we can suspect that the RF1 model may have been slightly overfitted, leading to higher error on unseen data. Not only did the RF4 model make less classification errors than the RF1 model but the errors it made were also only one class away as opposed to misclassifying long to short and short to long as the RF1 model did.

While this model works fairly successfully, and helps to confirm that the prediction is possible, it is not categorising accurately enough to conclude that this prediction is fully successful. The random forest models are capable of finding more complex patterns in the data than linear regression models, but it is likely that an even more sophisticated model is needed to fully capture this relationship.

### 2.4.4 A model to predict lifespan from simple phenotypes

Neural networks are one of the most sophisticated machine learning models, meaning they present the opportunity to establish patterns within datasets not identified by linear regression or random forest. Always structured as layers of nodes, the architecture can be configured in many different ways to address the problem, but there are some conventions. Based on these conventions, an architecture of 77 input nodes, a hidden layer of 40 nodes and an output layer of 1 node was created.

After some initial testing of different activation function, rectified linear unit (ReLU) was chosen for the model as this performed best. It is likely that the ReLU activation function performed better due to its common advantages over linear and sigmoid activation functions. ReLU is more likely to be able to identify complex patterns in data than the linear activation function which, given the low

success of linear regression models on this data, was a necessity for this model. Although sigmoid functions are generally capable of identifying these more complex patterns, they have limited sensitivity and saturation which ReLU does not suffer from.

Models were built using both Adam and SGD optimisers as both optimisers are appropriate choices for the dataset. SGD is the most commonly used optimiser in neural networks, but Adam may have advantages over it. Adam is used less commonly because it is more computationally intensive and so can slow down larger models. Since the dataset for this research is very small this is not a limitation we needed to consider in this work.

As with previous modelling, the neural network training datasets (defined in table 4 in *2.2.4 Neural networks*) were subset 80/20 into seen training and unseen test data to facilitate analysis of the model's performance. Performance of the models was assessed using root mean squared which is used to establish the prediction accuracy of the model on the unseen test data as it describes how far predicted values lie from the true recorded values.

Initial networks focussed on trialling a range of different learning rates to find the optimal learning rate for this data. Since the learning rate is usually a small number between 0 and 1 after a few pilot trials 0.0001, 0.0005 and 0.001 were chosen as a range to test. Using the mean squared error loss of the initial networks, it was clear that networks trained on the training dataset NN1 were the most accurate using both Adam and SGD optimisers and a learning rate of 0.0005 (table 9 and 10). This mirrors the results of the linear regression and random forest modelling as NN1 is once again a dataset made up of the mean lifespan and phenotypes rather than individual inputs.

These networks were then fine-tuned for learning rate optimisation. Networks were rebuilt using both optimisers but this time with learning rates of 0.0003, 0.0005 and 0.0007. From the root mean squared error of predictions on unseen test data, the model using the Adam optimiser with a learning rate of 0.0005 was the most accurate. When we look in more detail at these models in figures 12 and 13, we can see that the Adam optimiser likely performed better as the SGD model got stuck in local minima during training (figure 13A, C and E) affecting the training quality.

This most accurate network had smooth loss and error curves which converge towards 0 but do not then increase showing that the model trained smoothly and did not overfit (figure 12C). The scatterplot of the predicted lifespans vs the actual true value lifespans (figure 12D) shows again that the model did not perform overfit and perform better on seen training data than unseen test data, as well as visualising the high correlation between predicted and actual values for both seen training data and unseen test data. The network, with a root mean squared error of 2.83, is able to predict lifespan with a mean error of less than 3 days. If we consider the range of lifespans in the training data set (0-28 days), and that we usually only discuss lifespan as the qualitative labels of long, average and short, prediction within 3 days can be considered extremely successful.

Firstly, this result serves as a proof of concept that lifespan can be predicted from simple phenotypes. This has far-reaching implications within ageing by demonstrating that information within these stress response phenotypes can explain almost the entire lifespan phenotype. The ability to make this prediction supports current processes and factors though tot contribute to ageing, including oxidative stress and DNA damage, while suggesting that these

theories are needed in combination to explain the ageing phenotype. Secondly, the model serves as a genuine blueprint for a new kind of lifespan prediction. Since these kinds of stress response phenotypes can be predictive of yeast lifespan to within 3 days, it is possible that this concept could translate to higher eukaryotes, including humans, which have conserved stress pathways.

The concept that environmental stress responses and pathways can predict lifespan could inform the discovery of novel ageing and ageing-associated disease biomarkers as well as advice about lifespan and healthspan extending lifestyle choices. Because of these potential implications, it is an important next step to define which stress phenotypes and associated pathways are most predictive of ageing and therefore the most promising candidates for future study.

### 2.4.5 Identification of the most predictive phenotypes

Feature selection is not routinely performed on neural networks and usually the networks are treated as a ‘black box’ where the model intricacies are never fully understood. As the use of neural networks has become more widespread, there has been some development of feature selection processes to allow for this further step. Since the neural network was the much more successful model for this data, it was important to attempt to use this model for feature selection to allow for the most accurate feature importance rankings.

*A novel method for feature selection in in neural networks*

The feature selection script was based on the concept of the common feature selection method recursive feature elimination (RFE), used in other model types, including linear regression and random forest. In RFE, models are built and a metric for the importance of each feature is recorded during training, this

allows for the least important feature to be removed. This process is continued until a desired number of features or model accuracy metric has been reached. In neural networks there is no metric for the importance of a feature as the structure of the model does not allow for it and so RFE is impossible.

To create a viable alternative the feature selection script works in the opposite direction to RFE but with the same premise. At each step the most predictive feature is added to the model rather than eliminated, removing the need for a metric of feature importance by instead using a metric for model accuracy, root mean squared error (RMSE). At each step, the choice of the most predictive feature is established using 10-fold cross validation and the average RMSE of these networks. Using cross validation at every step increases the reliability of the feature selection by decreasing the impact of any outliers.

The output of the feature selection, a file of each feature ranked for importance along with the average RMSE from its 10-fold cross validation models, was then used to identify the most predictive phenotypes of lifespan. Initially, the rank was plotted against the RMSE (figure 14), which showed that the average model RMSE decreased as features were added to the lowest RMSE at feature 17 and then increased as features were added. This suggests that the first 17 features are improving the model and very predictive of ageing but that the subsequent features are at least redundant to lifespan prediction and possibly actively detrimental to it.

*Defining the most predictive phenotypes*

Therefore, the top ranked 17 features were subset as the predictive phenotypes. These phenotypes (table 12) are involved in a wide range of cellular processes, suggesting that the model is taking information about

several stress pathways into consideration when predicting lifespan. Phenotypes related to mitochondrial function, amino-acid availability, DNA damage, oxidative stress, target of rapamycin (TOR) inhibition, cell wall integrity, osmotic stress and translational inhibition were all identified as predictive phenotypes. Particularly the top 7 ranked phenotypes were associated with a steep decline in RMSE (figure 14) and can be considered the most predictive phenotypes. These related to ATP synthase inhibition, mitochondrial respiration inhibition, low nutrient availability, DNA damage and serine and proline supplementation.

*Mitochondrial stress is a most predictive phenotype for cellular lifespan*

The ranking of phenotypes related to ATP synthase inhibition and mitochondrial respiration inhibition suggests a strong connection between mitochondrial stress and lifespan. This is supported in the literature since mitochondrial stress has long been linked to ageing and ageing-associated diseases (Hill and Van Remmen, 2014, Lima et al., 2022, Dai et al., 2014). Mitochondrial hormesis is believed to have a biphasic response. Mild mitochondrial stress is thought to induce beneficial adaptations through metabolic reprogramming and epigenetic remodelling leading to mitochondrial import, proteostasis, oxidative phosphorylation, mitophagy and antioxidant defences. On the other hand, severe mitochondrial stress can lead to long term damage to the cell in through the reduction of ATP and the release of reactive oxygen species and mitochondrial damage associated molecular patterns leading to oxidative damage, inflammation, energetic crisis, cell damage and eventual death (Burtscher et al., 2023). In this way, we can consider that potentially the model may be identifying cells which have an increased resistance to severe

mitochondrial stress or an increased reaction to mild mitochondrial stress which lead to lifespan extending cellular processes.

*Nutrient availability is a most predictive phenotype for cellular lifespan*

Low nutrient availability also has long standing ties to ageing in the literature (Chen and Runge, 2009, Leitão et al., 2022, Pifferi et al., 2018). Caloric restriction has been shown to have extend ageing in many eukaryotes from extending chronological lifespan in yeast to extending lifespan by 1-5 years in humans (Flanagan et al., 2020, Leonov et al., 2017, Xiang and He, 2011). These results are often attributed to nutrient sensing pathways within the cell such as TOR or AMP kinase (AMPK) which modulate cellular processes in response to nutrient availability, leading to extended lifespan (Davinelli et al., 2012). We can therefore conclude that the model made associations between the ability to grow in low nutrient conditions on exhausted media and the lifespan of the strain. Given the extensive nature of the literature surrounding low nutrient availability extending lifespan it is most likely that this association was between resistance to low nutrient availability and longer lifespan.

*DNA damage is a most predictive phenotype for cellular lifespan*

DNA damage has both strong links to cellular ageing and the ageing-related disease cancer within the literature (Lee and Ong, 2021, Pal et al., 2018, Burhans and Weinberger, 2012). This process is another key research topic for ageing research and so its identification as a highly predictive phenotype for lifespan is supported by the literature. DNA damage can be caused by a range of processes including UV exposure, mitochondrial stress, and oxidative stress. It is thought to trigger cell senescence and even apoptosis by affecting many other cellular functions including the cell cycle and the tumour suppressor gene p53 which is a mechanism responsible for some types of cancer (Yousefzadeh

et al., 2021). It is likely that the model identified a relationship between increased resistance to DNA damage or increased capacity for DNA repair and longer lifespans.

*Amino-acid supplementation is a most predictive phenotype for cellular lifespan*

Serine and proline supplementation are the least well documented of the highly predictive phenotypes, however their involvement in lifespan and ageing is still supported by the literature (Rallis et al., 2021, Mirzaei et al., 2014). In budding yeast, serine supplementation has been shown to sensitise cells to oxidative stress and increased the chronological lifespan (Mirisola et al., 2014), and proline has been identified as a nitrogen source, stress protectant and energy source (Nishimura et al., 2021).

*A combination of simple phenotypes is needed for cellular lifespan prediction*

The key conclusion of this feature identification does not lie with any one most predictive phenotype, and it's support in the literature. The most interesting finding is that the model needed input from all of these phenotypes to be highly predictive of lifespan. For instance, prediction with only phenotypes relating to ATP synthase inhibition and low nutrient availability had a RMSE of more than 5 days. Addition of phenotypes relating to mitochondrial respiration inhibition, serine supplementation still only reduces the RMSE to more than 4.4 days. To bring the RMSE below 4 days, phenotypes relating to DNA damage and proline supplementation are needed. The need for input from different cellular processes at once to predict lifespan is clear, and this strongly suggests that the complex phenotype of lifespan is determined by a complex combination of these cellular processes.

Subsequent addition of phenotypes related to oxidative stress, TOR inhibition, cell wall integrity, osmotic stress and translational inhibition did go on to further reduce the RMSE and so improve the predictions of lifespan but in a much smaller range. This could be due to these processes being weaker determinants of lifespan or it could be due to redundancy. For example, TOR inhibition could be redundant to the higher ranked feature of nutrient availability as lack of nutrients can inhibit the TOR pathway, or oxidative stress could be redundant to the phenotypes relating to mitochondrial stress since mitochondrial stress can cause oxidative stress, or to DNA damage since this can be caused by oxidative stress. It is a limitation of the feature selection process that we cannot establish whether subsequent features have less effect on the RMSE due to biological importance or redundancy, but we can consider that all 17 features contributing to the model with the lowest RMSE are measurable determinants of cellular lifespan.

*Combinatory phenotypes for osmotic stress with cell wall integrity and DNA damage are predictive of cellular lifespan*

Osmotic stress has been shown to be linked to ageing with a decline in osmotic stress as cells age (Dues et al., 2016, Chen et al., 2020), and with hyperosmolarity shown to extend lifespan in *S. cerevisiae* and *C. elegans* (Chandler-Brown et al., 2015). Similarly, cell wall integrity has been shown to decrease and undergo remodelling with ageing in yeasts (Silva Vanessa et al., 2022, Molon et al., 2018). While not in the top 7 most predictive phenotypes, osmotic stress phenotypes appear twice in the top 17 predictive phenotypes. Both times, osmotic stress from NaCl appears in combination with another stress, first with cell wall integrity from calcofluor and then with DNA damage from MMS.

Since the feature selection filters out redundancy, and both calcofluor and MMS were already in the model, the addition of osmotic stress from NaCl must offer a new facet to the data. NaCl also appears as a single stress within the phenotypes and did not rank within the 17 phenotypes which improve model performance so we can infer that it is the combination of osmotic stress with cell wall integrity and DNA damage which is predictive. We can be confident that it is these combinations specifically which are predictive and not only the addition of another stress since NaCl, MMS and calcofluor all appear with other combinatory stresses which did not rank. This adds further weight to the conclusion that the model is based heavily on the combination of phenotypes and how potentially even how these phenotypes interact with each other.

*Future directions*

This model and feature selection has highlighted the complex and combinatory nature of the biological mechanisms which underpin ageing, further research could build on this by experimentally validating the predictions. By investigating the most predictive phenotypes and focussing on uncovering the mechanisms by which they are connected we could come closer to creating a more unified mechanism of ageing research.

# 3 Exploring the Role of Hsr1 in Cellular Ageing and Ageing-Associated Processes in Yeast

## 3.1 Introduction

### 3.1.1 Ageing networks

One of the obvious candidates for a mechanism by which different ageing pathways might interact is transcription factors. Cellular processes are all encoded within the DNA, but which ones go on to be expressed is determined by which genes are being transcribed. Transcription factors are one way the cell can control what genes are being expressed and when.

Transcription factors are proteins which bind to DNA and control the rate of transcription of their target genes. Since they are proteins themselves, and therefore transcribed, transcription factors can form feedback loops where one acts up or downstream of another. Eventually these interactions can go on to form large networks of transcriptional control which dictate how a cell responds to certain environmental stresses. This sort of network is a likely candidate to connect the many pathways thought to be involved in cellular ageing.

Over the last decade there has been mounting evidence for the role transcription factors play in cellular ageing. Given that organisms appear to age at different points in chronological age and that this process is reversible in some organisms (Matsumoto et al., 2019) ageing could be considered to be a gene-expression state. Transcription factors have been shown to affect many age-related cellular processes such as DNA repair, metabolism, and cell cycle control (Jin et al., 2020). This places transcription factor action in control of the initiation of many ageing processes, a theory backed up by the large number of transcription factor genes reported to have a lifespan effect when either deleted or overexpressed within a cell (Vachon et al., 2013, Rallis et al., 2014).

Consequently, downstream pathway inhibition has been found to alter lifespan such as the lifespan extending drugs rapamycin and Torin1, which inhibits the target of rapamycin (TOR) pathway (Fontana et al., 2010, Leontieva et al., 2015). TOR is a metabolic pathway which controls the response environmental and physiological conditions, including nutrient availability (Laplante and Sabatini, 2009). With this in mind, control of cellular response to nutrient availability within a cell is transcriptional in terms of both transcription factors and downstream pathways as different transcriptional profiles are found when cells are exposed to different conditions.

Transcription factors can also bridge the gaps between pathways previously considered to be independent of each other. In the case of transcription factors, different transcriptional pathways can often regulate each other in feedback loops. This mechanism is likely integral to a process such as ageing where many different pathways seem to play different and often opposing roles.

### 3.1.2 Gaf1 and Php transcription factors

One recently characterised ageing-associated transcription factor in the lab is the GATA factor Gaf1 (Rodríguez-López et al., 2020). Deletion of *gaf1* was shown to lead to shortened chronological lifespan and diminished Torin1-mediated longevity, suggesting Gaf1 involvement with the target of rapamycin (TOR) pathway. Gaf1 was found to be inhibited by TOR complex 1 (TORC1) and, in TORC1 inhibition by the inhibitor Torin1, Gaf1 was shown to regulate metabolism genes, tRNA genes and translation-related genes. The TOR pathway is a conserved pathway in eukaryotes which is a key regulator for cell growth and metabolism in response to nutrient availability (González and Hall, 2017, Gonzalez and Rallis, 2017, Valvezan and Manning, 2019). It has implications in stress response, protein synthesis, autophagy, and anabolic and catabolic processes (Matsuo et al., 2007, Poüs and Codogno, 2011, Weisman and Choder, 2001).

Within the lab, further exploration of ageing-associated transcription factors focussed on the Php factors: Php2, Php3, Php4 and Php5. All of these transcription factors bind to *gaf1* (Rodrígez-López, 2023, UCL, personal communication) and therefore help to regulate the downstream ageing-associated pathways of Gaf1. Php transcription factors have also been shown to be involved iron ion availability (Php2, Php3, Php4 and Php5) (Mercier et al., 2006). Nitrogen starvation has long been associated with extended cellular lifespan (Santos et al., 2016) and iron accumulation is associated with ageing and age-related diseases including neurodegenerative diseases (Mangan, 2021). In this way, Php transcription factors are implicated in ageing-associated cellular pathways both through regulating Gaf1 and other mechanisms.

Expanding our understanding of transcription factors can help to create a picture of the network of transcriptional control of ageing which bridges gaps between ageing-associated pathways within the cell. Elucidating this transcriptional control can lead to discovering new connections between theories of ageing and brings us closer to a unified mechanistic explanation.

### 3.1.3 Literature review of ageing transcription factor genes of interest

Six transcription factors thought to be involved in ageing or ageing-associated pathways, such as the TOR pathway, were selected for this investigation and discussed in detail below. Candidates were chosen due to ageing-related overexpression phenotypes, recorded long lifespan, resistance to ageing-related drugs, such as caffeine and rapamycin or Torin1, or interactions with other transcription factors studied in the lab: Gaf1 or Phps. Of the six chosen factors, four are partially investigated genes (*phx1, hsr1, moc3, rsv2*) and two are unknown and unnamed *S. pombe* genes whose transcription factor action are inferred from sequence models.

*Phx1 (SPAC32A11.03c)*

The gene *phx1* (2829bp) codes for the protein Phx1, a large protein with a conserved homeodomain. It appears to be highly homologous to Pho2, a known transcription factor in *S. cerevisiae* (Cheng et al., 2000 ), and has been shown to bind to DNA with the ability for transcriptional activation (Kim et al., 2012). GFP-tagged Phx1 showed that the protein is produced in the late exponential phase, then accumulated during stationary phase. Phx1 is located primarily in the nucleus. Northern blotting analysis of *phx1* mRNA transcripts showed a similar expression pattern to that of Phx1 protein suggesting that levels of Phx1

protein in the cell are directly driven by *phx1* expression (Kim et al., 2012). The production of Phx1 at stationary phase is believed to be at least partly due to nutrient starvation (Kim et al., 2012).

The *S. pombe* deletion strain *phx1*Δ shows increased viability in stationary phase (Kim et al., 2012) and increased fitness during Torin1 inhibition (Lie et al., 2018), however, it was not highlighted in a screen for caffeine and rapamycin resistance (Rallis et al., 2014). Kim et al. (2012) found that *phx1*Δ has reduced viability compared to wild type cells under normal and nutrient starved conditions. They also found it to have reduced stress tolerance for oxidative and heat stresses and that in wild type cells, *phx1* transcription increased in response to these stresses. An overexpression study of *S. pombe* transcription factors showed that ectopic expression of *phx1* leads to moderately impaired fitness and moderately elongated cells with aberrant septal deposition (Vachon et al., 2013).

Because Phx1 levels in the cell appear to be based primarily on *phx1* transcription, it is likely that *phx1* is under strict control by other transcription factors. The gene *phx1* is bound by Php4, an ageing related transcription factor of interest in the lab (Rodrígez-López, 2023, UCL, personal communication).

*Hsr1 (SPAC3H1.11)*

The gene *hsr1* (1749bp) codes for the transcription factor Hsr1 which functions in response to oxidative stress in the cell (Chen et al., 2008). *S. pombe hsr1*Δ is shown to have increased viability in stationary phase (Ohtsuka et al., 2011) but not highlighted on screens for resistance to caffeine and rapamycin (Rallis et al., 2014) or resistance or sensitivity to Torin1 (Lie et al., 2018). An overexpression study of *S. pombe* transcription factors showed that ectopic

expression of Hsr1 leads to moderately impaired fitness and moderately elongated cells with no cell cycle phenotype (Vachon et al., 2013).

Hsr1 requires both Pap1p and Prr1p for induction in response to oxidative stress with hydrogen peroxide (Chen et al., 2008), and *hsr1* is bound by Gaf1 with and without Torin1 treatment (Rodríguez-López et al., 2020), Php2, Php4 and Php5 (Rodríguez-López, 2023, UCL, personal communication) which are established ageing-related transcription factors. Hsr1 is of particular interest in this work due to the wide binding network to other transcription factors.

*Moc3 (SPAC821.07c)*

The gene *moc3* (1527bp) codes for the protein Moc3 containing a zinc finger binding motif which is localised in the nucleus involved in sexual development, ascus formation, and stress response in *S. pombe* (Goldar et al., 2005). Moc3 is a predicted transcription factor inferred from sequence model and the deletion strain *moc3*Δ leads to increased viability in stationary phase (Rallis et al., 2014). Additionally, *S. pombe moc3*Δ shows lower mating efficiency and forms aberrant asci. This deletion strain has been shown to be sensitive to $CaCl_2$ and DNA damaging agents, such as MMS and UV (Goldar et al., 2005).

Moc3, along with Moc1, Moc2, and Moc4, are thought to be positive regulators of *ste11* as overexpression of Moc3 was shown to lead to efficient induction of *ste11* (Paul et al., 2009). Ste11 is a key transcription factor responsible for positively regulating genes required for the initiation of mitosis. *S. pombe* ***ste11***Δ mutants are completely defective in mating and sporulation whereas the overexpression of *ste11*+ leads to sexual reproduction even in stress conditions (Kim et al., 2012) which may be why the strain *moc3*Δ shows lower mating efficiency. Furthermore, the ageing-related transcription factor Gaf1 has been

shown to downregulate *ste11* (Kim et al., 2012) and therefore might be part of a wider *ste11* regulation network with Moc3.

*Rsv2 (SPBC1105.14)*

The gene *rsv2* (1914bp) codes for the protein Rsv2 which induces stress-related genes during spore formation induced in middle/late meiosis (Mata et al., 2007) and is involved the in amino acid starvation response (Duncan et al., 2018). The deletion strain *rsv2*Δ is resistant to caffeine and rapamycin treatment (Rallis et al., 2014) and has increased viability in stationary phase (Ohtsuka et al., 2011); however, overexpression of *rsv2* also leads to increased chronological lifespan (Ohtsuka et al., 2011). The transcription factor Fil1 shows marginal affinity to *rsv2* (Duncan et al., 2018), suggesting further involvement with broader network via Php3 (Rodrígez-Lopez, 2023, UCL, personal communication).

*Novel transcription factors*

Two previously uncharacterised genes predicted to have transcription factor action were also chosen for this study: SPAC2H10.01 (1443bp) and SPAC11D3.17 (1758bp). Though neither of them are known to cause long-lived mutants in deletion strains, they are both bound by Gaf1(Rodríguez-López et al., 2020), Php2 and Php4. SPAC2H10.01 is bound by Php5 (Rodrígez-López, 2023, UCL, personal communication). This implicates them in a wider ageing-related transcription factor network.

# 3.2 Methodology

## 3.2.1 Generation of deletion and GFP tagged strains

DNA was amplified by PCR from the pFA6-NAT-MX6 plasmid for deletion constructs and the pFA6a-GFP(S65T)-natMX6 plasmid for the GFP-tagged constructs, using primers specific to each gene found in tables 13 and 14.

Table 13: Sequences of amplification primers for gene deletion

| Gene | Forward Sequence | Reverse Sequence |
|---|---|---|
| ***phx1*** | TTTCTTGGCATACTTTTTGAAGCAAATTTTTATTTCTCATAAGGATTTTATTTTCATTTCATTTTATTTCTAAGAACAATCGGATCCCCGGGTTAATTAA | TTCTCGTTATCAAAAGAAAACGAAAATAAGCAAGCTTCAAGCGAGTTTCAATTGTACCGTTAATCATCATATCACAAAATGAATTCGAGCTCGTTTAAAC |
| ***hsr1*** | TCTTTTCTTTAGTTGATTTTTTATTTTTGAAAAGTATTCGCTTACTTTCTTTTTATAATAATTCTTTTATCTTACTTTGTCGGATCCCCGGGTTAATTAA | AAATTATAATTGAAAACATTCTTATAAAACAGTTCAATGTAAAAAAAACCCGAATTTAGGCAGTTAATTTATAAAAATGCGAATTCGAGCTCGTTTAAAC |
| ***moc3*** | GGCTTTCTTTACTTTTGATTTGTTTAGTTCCTATTATCTTGTTATTCTTTTTTTTTCTATCTATTGTTTCCCTTGCAGTTCGGATCCCCGGGTTAATTAA | ATGTTTGATGCGTGATTTTGTCTAGCTATTATACAGTTCTATATCAATATTTATTGAAAAGCATCAAAATGATTTTAAAAGAATTCGAGCTCGTTTAAAC |
| ***rsv2*** | TGACAAGGGTGGTTTTTTCAATCAGACGCTTGACGTCATTCTTTATATATACATTGCCTCGCGCATCTTTGCTGGGTAGTCGGATCCCCGGGGTTAATTAA | AAGCGTAAGAGGTGAAAAATGACAATAGATAATGCAATAACCTTTTTTTAATATTTGTGGCAGAAGAGACTCCTTTCAATGAATTCGAGCTCGTTTAAAC |
| **SPAC2H10.01** | TACTTCACTAATTGCATTTACTTTTTTTTCCAACTCTAATTCCCCTTTCTTTCTGTAGGGATTTCCTTTAAACTGTTAACCGGATCCCCGGGTTAATTAA | TTATTCTCAAAAAACAACGAAGTTAGAGAAATAAACGTTTCCCGGTTACAACCAACAAAAAAGCCATGTCAAAATTTGACGAATTCGAGCTCGTTTAAAC |
| **SPAC11D3.17** | GTTAACACTTTTTAAAATTTTATACTTATTTATGCGTTGGATTTTCTTTGGGAAAGTTTTGTATTGATTTCTCCACAAATCGGATCCCCGGGTTAATTAA | AATAACTTGTTGTAGAAAGAAGAGAATTAGTAAAAACAAAACAGAGAAATAATAAATCTAAAAAAAAATAATATATAAAAGAATTCGAGCTCGTTTAAAC |

Table 14: Sequences of amplification primers for gene tagging

| Gene | Forward Sequence | Reverse Sequence |
|---|---|---|
| ***phx1*** | CATTTGAGGATGTTTACTCTCCTTCTGCTG GTATAGATTTTCAGAAACTTCGTGGTCAAC AATTTTCTCCGGACATGCAGCGGATCCCC GGGTTAATTAA | TTCTCGTTATCAAAAGAAAACGAAAATAAGC AAGCTTCAAGCGAGTTTCAATTGTACCGTTA ATCATCATATCACAAAATGAATTCGAGCTCG TTTAAAC |
| ***hsr1*** | CTAAACTACCAGTACAAACACCTAACCAAA AAATGCCTTTAATGAATCCGATGCATCAAT ATCAACCTTATCCTAGTTCTCGGATCCCCG GGTTAATTAA | AAATTATAATTGAAAACATTCTTATAAAACAG TTCAATGTAAAAAAAACCCGAATTTAGGCAG TTAATTTATAAAAATGCGAATTCGAGCTCGT TTAAAC |
| ***moc3*** | GGGCACAAATCGAATCTCATTTCGGAAGAT TGATGCTAGAACATTTTATGGATTGCAATG TCTTAAATCGACCAGTACTTCGGATCCCCG GGTTAATTAA | ATGTTTGATGCGTGATTTTGTCTAGCTATTA TACAGTTCTATATCAATATTTATTGAAAAGCA TCAAAATGATTTTAAAAGAATTCGAGCTCGT TTAAAC |
| ***rsv2*** | GATGTGAGATTTGTGGCGATCAACGCCATT TCAGTAGACATGATGCCTTGGTTAGGCATC TCCGTGTGAAACACGGTAGACGGATCCCC GGGTTAATTAA | AAGCGTAAGAGGTGAAAAATGACAATAGAT AATGCAATAACCTTTTTTTAATATTTGTGGCA GAAGAGACTCCTTTCAATGAATTCGAGCTC GTTTAAAC |
| **SPAC2H10.01** | TCTCTTCAAATACATCTCTGGATGATATGTT TTTCTTTATTCGTGATTTCGATGAGGATCAT CCAATTCAAACTGCATATCGGATCCCCGG GTTAATTAA | TTATTCTCAAAAAACAACGAAGTTAGAGAAA TAAACGTTTCCCGGTTACAACCAACAAAAAA GCCATGTCAAAATTTGACGAATTCGAGCTC GTTTAAAC |
| **SPAC11D3.17** | CTGCATTTGTTGCAAGCGCTTTGGATGTTG AAGGTGGATGGGGTATTGGTCCACTTCTTA CCAAAGCGTTCCGTCCAACACGGATCCCC GGGTTAATTAA | AATAACTTGTTGTAGAAAGAAGAGAATTAGT AAAAACAAAACAGAGAAATAATAAATCTAAA AAAAAATAATATATAAAAGAATTCGAGCTCG TTTAAAC |

PCR product was visualised by gel electrophoresis using ethidium bromide to ensure correct sizes of PCR product. This DNA was then used for the transformation of live yeast.

*Transformation procedure*

A 20ml YES culture of wild type yeast was grown overnight to OD0.2-0.5, pelleted and washed with sterile water. The cells were then washed in LiAcTE

(0.1M Lithium Acetate, 10mM Tris pH7.5, 1mM EDTA) before resuspension in 100µl LiAcTE. Then 10µl of DNA was added along with 2µl of 10mg/ml herring sperm carrier DNA and the cells were incubated at room temperature for 10 minutes. 260µl of LiAcTE-40%PEG (0.1M Lithium Acetate, 10mM Tris pH7.5, 1mM EDTA, 40% Polyethylene Glycol 4000) was added to the cells and they were incubated at 32°C for 60 minutes. 43µl of pre-warmed DMSO was added and the cells were heat shocked at 42°C for 5 minutes. The cells were pelleted, washed with sterile water and then plated in triplicate on YES plates. Plates were grown for 2 days at 32°C before replica plating on to YES NAT selective plates which were incubated at 32°C for up to two weeks until colonies appeared.

*Checking for correctly transformed colonies*

Once the colonies had grown, colony PCR was used to identify correctly transformed colonies. For this, checking primers were used in combination with internal primers which can be found in tables 15 and 16. All colonies were checked with outside-inside primer combinations to ensure correct orientation of DNA.

Table 15: Sequences of checking primers for gene deletion

| **Gene** | **Left-flanking Sequence** | **Right-flanking Sequence** |
|---|---|---|
| ***phx1*** | TTTGCGTCTCCTCAAGTACTCA | TTTTTCGCATCAAAGTTCTTCC |
| ***hsr1*** | GCTACGTTGTTTGCAGTCAAAA | GCCTAAAGATAGCAAAGCAGGA |
| ***moc3*** | AGACGACCATTGATTTTCACCT | CGTAATTCGTAAATTTCGGCTC |
| ***rsv2*** | TTGGCTCAATCAATGTAAAACG | GGATACGATGAAATGAAGAGGC |
| **SPAC2H10.01** | TCGTTTTATTTCCTTTCCGCTA | AAATGAACAAAAAGGGGGAAAT |
| **SPAC11D3.17** | TAACATTTTGCAATTGAGCCAC | CATCCGGAAGCGTATTTATTTT |
| **Nat cassette internal checking primer:** | | CGAGTACGAGATGACCACGA |

Table 16: Sequences of checking primers for gene tagging

| Gene | Left-flanking Sequence | Right-flanking Sequence |
|---|---|---|
| ***phx1*** | GAGCTTCGGGCTACTTATCTCA | TTTTTCGCATCAAAGTTCTTCC |
| ***hsr1*** | TTGTCTCAACAAATTGTCCCAG | GCCTAAAGATAGCAAAGCAGGA |
| ***moc3*** | GGACATTCTCTCTCAACCAACC | CGTAATTCGTAAATTTCGGCTC |
| ***rsv2*** | GGTACCTAATCCAACCAACCAA | GGATACGATGAAATGAAGAGGC |
| **SPAC2H10.01** | TTTGGAGATTGACCAAGGAAGT | AAATGAACAAAAAGGGGGAAAT |
| **SPAC11D3.17** | AAAATGTGATTTGGGGTTTACG | CATCCGGAAGCGTATTTATTTT |
| **Nat cassette internal checking primer:** | | CGAGTACGAGATGACCACGA |

## 3.2.2 Western blot analysis

Cells were grown in YES liquid media to OD 0.2-0.5. 25ml of cells were harvested, pelleted, washed with 25ml of sterile water and then transferred to microcentrifuge tubes for snap freezing with liquid nitrogen. The culture was then treated with $H_2O_2$ (0.5mM) and the process was repeated at 15 minutes, 30 minutes and 60 minutes after treatment for each strain. The samples used for protein extraction in this experiment are summarised in table 17.

Table 17: Table to summarise the samples used for protein extraction and western blot in this experiment including wild type negative control sample

| Strain | Timepoints | | | |
|---|---|---|---|---|
| **Hsr1:GFP** | Time 0/ Untreated | 15 minutes | 30 minutes | 60 minutes |
| **Wild type/JB22** | Time0/ Untreated | | | |

*Protein extraction*

The cell pellets were thawed on ice and resuspended on 100µl of lysis buffer. Glass beads were added, and the cells were lysed using a FastPrep mechanical lysis machine for 20 second intervals with cooling on ice for 5

minutes between. The cells were lysed until 80% of the cells were seen to be broken under the microscope (~4 cycles of lysis and cooling). The lysed cells were collected from the beads by centrifugation and the beads were eluted with another 100µl of lysis buffer to ensure complete collection of lysed cells. The elute was centrifuged at top speed for 10 minutes and the supernatant collected and saved to remove cell debris and insoluble proteins. The supernatant was then quantified by BCA assay and corrected to samples containing 50µg of protein for western blotting. NuPAGE LDS sample buffer and 0.1m DDT were added to the samples and they were boiled for 5 minutes at 80ºC.

*Western blotting and visualisation*

The samples were loaded into a Bio-Rad Mini-PROTEAN TGX Precast Gel 4%-20% and run at 200V for 50 minutes. Blots were transferred using a semi-dry transfer system for 10 minutes and the nitrocellulose was dyed with Ponceau S stain to check for even bands. The stain was washed off with distilled water and blocked for 2 hours in 5ml of PBS-T 5% milk. The primary GFP monoclonal antibody (Invitrogen, GF28F) was added at 1:1000 in 5ml of PBS-T 5% milk and incubated overnight at 4°C. The blot was then washed three times for 5 minutes with PBS-T before adding ABCAM Goat anti-mouse IgG at 1:5000 in 5ml of PBS-T 5% milk and incubated at room temperature for 2 hours. The blot washed three times for 15 minutes in PBS-T, three times for 15 minutes in PBS, then visualised using ECL.

To visualise the Cdc2 loading control protein the blot was then washed overnight in PSB-T and blocked again for 1 hour in PBS-T 5% milk. Primary pSTAIR antibody was added at 1:5000 PBS-T 5% milk and incubated at room temperature for two hours. The blot was washed three times with PBS-T for five minutes and then incubated for 1 hour at room temperature with Abcam Goat

anti-mouse IgG at 1:10,000 in 5ml of PBS-T 5% milk. The blot was washed three times for 15 minutes in PBS-T, three times for 15 minutes in PBS, then visualised using ECL.

### 3.2.3 Caffeine and rapamycin stress spot tests

Cells were grown to OD 0.2-0.5 and then corrected to OD 1.5 and serially diluted as shown in table 18. These cells were then spotted onto stress plates. Plates were prepared a day in advance using YES with 3% glucose and the appropriate stressors.

Table 18: Table of dilution factors for serial dilution of samples for stress spot plates.

| **Column** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| **Dilution factor** | None/ OD 1.5 | 1 in 4.3 | 1 in 18.5 | 1 in 79.5 | 1 in 341.2 | 1 in 1470.1 |

### 3.2.4 Chromatin immunoprecipitation sequencing (ChIP-Seq)

The experiment was designed as a $H_2O_2$ treatment time course using the *hsr1*-GFP strain with two repeats, using anti-GFP antibody for the experimental samples and anti-HA antibody for control immunoprecipitations (IPs). Samples are summarised in table 19.

Table 19: Summary of samples for ChIP-Seq, including repeat, antibody used for sample, and time after treatment with $H_2O_2$ the sample was taken.

| **Repeat** | **Antibody** | **Time (mins)** |
|---|---|---|
| 1 | Anti-GFP | 0 |
| 1 | Anti-GFP | 30 |
| 1 | Anti-GFP | 60 |
| 2 | Anti-GFP | 0 |
| 2 | Anti-GFP | 30 |

| | | |
|---|---|---|
| 2 | Anti-GFP | 60 |
| 1 | Anti-HA | 0 |
| 1 | Anti-HA | 30 |
| 1 | Anti-HA | 60 |
| 2 | Anti-HA | 0 |
| 2 | Anti-HA | 30 |
| 2 | Anti-HA | 60 |

*Extraction of chromatin*

For the control and experimental samples for each repeat 600ml of cells were grown to OD 1.0. 200ml were collected as a time 0 control. The remaining 400ml were subjected to H2O2 (0.5mM) treatment with 200ml collected after 30 minutes and the last 200ml collected after 60 minutes. Cells were fixed immediately upon collection by adding 5.4ml of 37% formaldehyde (1% final) for 30min at room temperature with gentle shaking. 10ml of 2.5M glycine were then added and the sample was incubated for 10min at room temperature. The cells were then pelleted by centrifugation at 3000rpm for 3 minutes, then washed in 40ml of ice cold 1xPBS. The pellet was then resuspended in 4ml of ice cold 1xPBS and the sample was split into 2 tubes, pelleted again and the supernatant was discarded. Pellets were then snap frozen and stored at -70°C.

The pellets were thawed on ice for 5min and resuspended in 800µl of ice-cold lysis buffer (50mM Hepes pH7.6, 1mM EDTA pH8, 150mM NaCl, 1% Triton X-100 and 0.1% Na-Doc) and 2x Roche EDTA-free Protease Inhibitors 1mM PMSF. 600µl of acid washed glass beads were added and the cells were broken using a FastPrep with 9-12x 20sec at 5.5 with 5 minute incubations on ice in between. 1µl aliquots of cells were visualised under the microscope to ensure breaking efficiency of >90%. A hole was poked into the bottom of the

tubes using a flamed, sterile needle and they were placed into clean tubes and centrifuged for 1min at 1000rpm at 4°C to collect the lysate. The glass beads were then washed with 400µl of lysis buffer (with protease inhibitors) and the centrifugation repeated so the flow through pooled with the collected lysate. The lysate was then centrifuged for 10 min at 20000rcf and the supernatant was discarded. The pellets were washed with 800µl of cold lysis buffer and then resuspended in 750µl of cold lysis buffer and the split samples were pooled into one tube again. PMSF was added to 1mM and the sample was split into 5x 300µl aliquots for sonication.

300µl of ceramic beads were added to the aliquots and they were sonicated using a Bioruptor for 35min with 30sec on/ 30sec off on high. The sonicated material was then spun at 20000rcf for 10min at 4°C and the supernatant was collected. This supernatant is the chromatin extract (CE). The CE was stored at -70°C.

5µl of the CE was used in a Bradford assay to check protein concentration and 50µl of the CE was used to check sonication efficiency using a Bioanalyzer before continuing. 50µl of CE was stored at -20°C for ‘input’ library construction.

*Immunoprecipitation*

100µl of Protein G-coated magnetic beads and 1ml of block solution (0.5% BSA (w/v) in LB) were added to 1.5ml microfuge tubes, one for each sample. Beads were collected using a magnetic stand and the supernatant removed. The beads were washed in 1ml of block solution twice more. Beads were then resuspended in block solution and 5µg of antibody (anti-GFP or anti-HA respectively) was added in a final volume of 250µl. The beads were then incubated overnight on a rotating platform at 4°C. 5mg of CE (volume calculated

using Bradford results) was then added to 50µl of the antibody/magnetic bead mix and these samples were gently mixed by rotating at 4°C overnight.

The beads were then collected using a magnetic stand and the supernatant was removed by aspiration. The beads were then washed in 0.8ml of each of the following buffers in semi-cold conditions.

- 2x in Lysis Buffer (50mM Hepes pH7.6, 1mM EDTA pH8, 150mM NaCl, 1% Triton X-100 and 0.1% Na-Doc)
- 2x in Lysis 500 (50mM Hepes pH7.6, 1mM EDTA pH8, 500mM NaCl, 1% Triton X-100)
- 2x in LiCl/NP-40 (10mM Tris-HCL pH8, 1mM EDTA pH8, 250mM LiCl, 1% NP-40, 1% Na-Doc)
- 1x in TE (10mM Tris-HCl pH8, 1mM EDTA)

Beads were then resuspended in 200µl of elution buffer (50mM Tris-HCl pH8, 10mM EDTA, 1% SDS) and incubated at 65°C for 6-18 hours to elute and perform reverse crosslinking. 150µl of TES was added to the reserved 50µl of 'input' CE samples and these were incubated alongside the other samples for reverse crosslinking and were subsequently treated the same as the samples.

200µl of TE and 5µl of DNase-free RNase (0.5mg/ml) was added to the samples and they were incubated at 37°C for 60min. 7µl of proteinase K (20mg/ml) was then added and they were incubated at 55°C for 2 hours. DNA was then purified using the MinElute Qiagen kit to the manufacturers protocol and these samples were stored at -20. The library was prepared using the NEB Next Ultra II library prep kit to the manufacturers protocol and sent for sequencing.

*Bioanalyzer analysis of ChIP-Seq samples*

150µl of TES buffer (10mM Tris-HCl pH8, 1mM EDTA, 1% SDS) was added to the 50µl samples and they were incubated overnight at 65°C. 200µl of TE buffer (10mM Tris-HCl pH8, 1mM EDTA) and 5µl of DNase-free RNase (0.5mg/ml)

were added to the samples and they were then incubated at 37°C for 30min. Then 7µl of proteinase K (20mg/ml) was then added to the samples and they were incubated at 55°C for 2 hours. The samples were then cleaned up using the MinElute Qiagen kit to the manufacturers protocol and then 1µl of sample was analysed on a Bioanalyzer.

*Mapping the reads*

Initial analysis of the ChIP-Seq reads was performed using the EU galaxy cluster (Community, 2022) as follows. First, read quality was assessed using *FastQC* (Andrews, 2010) and the reads were subsequently trimmed using *FASTQ Trimmer (Blankenberg et al., 2010)*. The reads were then aligned to the *S. pombe* genome (Wood et al., 2002) using *bowtie2* (Langmead and Salzberg, 2012, Langmead et al., 2009) and filtered to remove non-uniquely mapped reads using *Filter SAM or BAM* (Li et al., 2009). *bamCoverage* (Ramírez et al., 2016) was used to create bigWig files and these were viewed on the *PomBase* genome browser (Harris et al., 2021) to view the peaks before the peak call.

*Peak call and annotation*

*MACS2 callpeak* (Feng et al., 2012, Zhang et al., 2008) was used to call the peaks for each time point for each repeat using the samples treated with the anti-HA as the control files for the samples treated with anti-GFP. The peak files for the repeats were then joined, using *Join*, to create files containing only the peaks which occurred in both repeats, known as the high-confidence peaks. Peak annotation was performed in R using *ChIPseeker* (Wang et al., 2022, Yu et al., 2015) and the *S. pombe* annotated genome (Wood et al., 2002) on the high-confidence peaks, exported from galaxy.

*Binding motifs and gene ontology enrichment*

*MEME* (Bailey et al., 2015) was used on the high confidence peaks for time 30 to search for binding motifs and *TomTom* (Tanaka et al., 2011) was used to search for similar motifs in *S. cerevisiae* and humans. The high confidence peaks at all time points were then used for gene ontology enrichment using AnGeLi (Bitton et al., 2015).

### 3.2.5 Synthetic genetic array (SGA)

The synthetic genetic array (SGA) experiment was designed to compare the growth of double mutants of *hsr1*Δ and the strains in the deletion library, on both YES media and YES with caffeine (10mM) and rapamycin (100ng/l). For each of these conditions there were three repeats and a control using the *ade6*Δ strain rather than the *hsr1*Δ, a total of 12 SGA experiments.

*Preparation of the library, query, and control plates*

A prototrophic version of the Bioneer deletion library (*h*- with kanamycin antibiotic resistance markers) was woken up by pinning into 384 plate format, using the RoToR HDA robot (Singer Instruments) on to YES PlusPlates (Singer Instruments) and 96 long pins (Singer Instruments). Three complete sets of the library were woken up to ensure enough material for the whole experiment. Plates were incubated at 32°C for 4 days.

The query strain (*hsr1*Δ) and control strain (*ade6*Δ) (both *h*+ with nourseothricin antibiotic resistance markers) were woken up from glycerol stocks by streaking on to YES agar. These plates were incubated at 32°C for 2 days and then 6x 200ml YES cultures were inoculated for each strain. These cultures were incubated at 32°C with shaking overnight. The cultures were then pinned in 384 format using the RoToR HDA robot (Singer Instruments) on to YES PlusPlates

(Singer Instruments) and 384 long pins (Singer Instruments). Each culture was pinned on to 9 plates to ensure enough material for the whole experiment and plates were incubated at 32°C for 2 days.

*Mating of query and control strains to the library*

Matings were carried out by pinning using the RoToR HDA robot (Singer Instruments) on to ME PlusPlates (Singer Instruments) and 384 short pins (Singer Instruments). First the library was pinned to the plates with source mixing to ensure even material transfer, then this process was repeated with the respective query or control plate. After pinning, the RoToR HDA robot (Singer Instruments) and 384 long pins (Singer Instruments) were used to add a drop of water and pierce the agar at each spot, to increase mating efficacy. Plates were incubated at 25°C for 3 days for mating and then moved to 42°C for the subsequent 4 days to kill the adult cells and keep only the spores.

*Selection of double mutants*

Spores were woken up by pinning using the RoToR HDA robot (Singer Instruments) on to YES PlusPlates (Singer Instruments) with 384 short pins (Singer Instruments). Plates were incubated at 32°C for 2 days. The germinated spores were then pinned using the RoToR HDA robot (Singer Instruments), with 384 short pins (Singer Instruments), on to YES with nourseothricin and kanamycin or YES with nourseothricin, kanamycin, caffeine (10mM) and rapamycin (100ng/l). These plates select for only the double mutants with both antibiotic markers and show their growth on the YES control plates and under the treatment of caffeine (10mM) and rapamycin (100ng/l).

*Calculating genetic interactions*

Plates were imaged using a conventional scanner and a custom Unix script within the lab to crop and colony size quantification was carried out using the R package *gitter* (Wagih and Parts, 2014). The interactions were then calculated using a custom R script within the lab. First, extremely small colonies which are likely indicative of a strain not having woken up were identified and excluded from further analysis. All plates were then normalised to the plate median to account for plate dependant differences in growth and to the row and column medians to account for plate positional dependant differences in growth. The colonies were then mapped to the library deletion mutants to label each colony as it's respective deleted gene. The genetic interaction as $\log_2$ fold change was then calculated as $\log_2$(mean of *hsr1Δ* repeats / mean of *ade6Δ* repeats), using the *ade6Δ* colonies as a control for the *hsr1Δ* colonies. The p-values for the interaction calculations were also recorded.

*Exclusion of linked loci*

To identify the linked loci for both *hsr1* and *ade6*, positional information for all genes was extracted from the *S. pombe* annotated genome (Wood et al., 2002). Any genes with a locus less than 500,000bp away from either *hsr1* or *ade6* was identified as a gene with linked loci. Interactions for these genes were subsequently removed from the data before any further analysis.

*Identifying positive and negative interaction hits*

Volcano plots were created using the R package *EnhancedVolcano* (Kevin Blighe, 2021) to visualise the interactions with a p-value <0.05 and a $\log_2$ fold change of >±1 and >±0.5 Any interactions with a p-value of >0.05 was excluded from the analysis and lists of positive and negative interactions were created

over both $\log_2$ fold change thresholds. These lists were analyses for overlap using Venn diagrams.

## 3.3 Results

### 3.3.1 Generation of deletion and GFP tagged strains

For use in further experiments, deletion and GFP tagged strains for each of the chosen transcription factors were generated by homologous recombination. Transformations were carried out to generate twelve strains: *phx1*Δ, *hsr1*Δ, *moc3*Δ, *rsv2*Δ, SPAC2H10.01Δ, SPAC11D3.17Δ, *phx1*-GFP, *hsr1*-GFP, *moc3*-GFP, *rsv2*-GFP, SPAC2H10.01-GFP, and SPAC11D3.17-GFP.

Deletion strains were successfully generated for five of the six chosen transcription factors, selected by PCR. As shown in figure 15A, the deletion strains for *phx1, hsr1, moc3* and *rsv2* were all identified as having positive right and left flank PCR. The deletion strain for SPAC2H10.01 had only the right flank. No successfully transformed deletion strains were generated for SPAC11D3.17. GFP tagged strains were successfully generated for all six chosen transcription factors, selected by PCR. In figure 15B left and right flanks can be seen for all strains.

Western blot analysis was used to confirm the presence of the GFP tagged protein product (Hsr1-GFP) in the *hsr1*-GFP strain. Figure 16 shows bands at 30 minutes and 60 minutes after treatment with $H_2O_2$ at the correct size for Hsr1-GFP (~91kDa). The Cdc2 control band (~34kDa) is visible at all timepoints with consistent intensity.

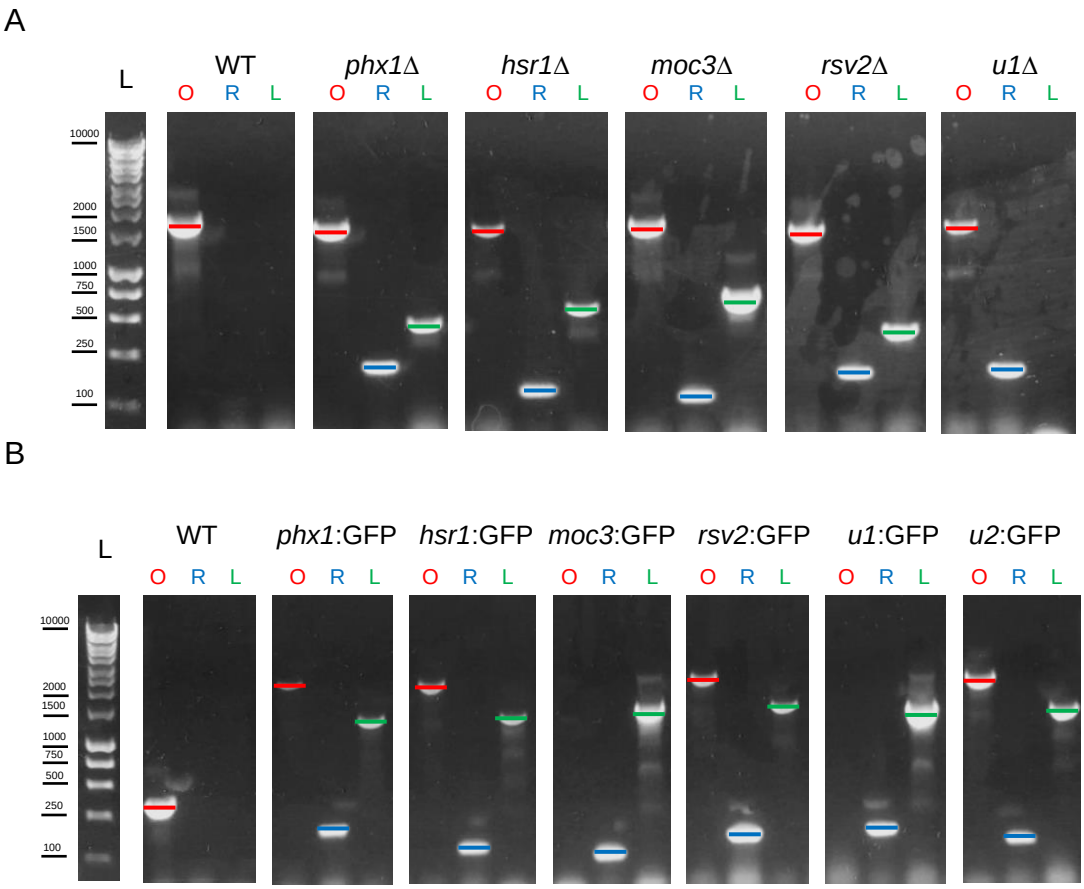



Figure 15: PCR analysis of mutants to ensure correct transformation.

Gels show (A) five deletion strains and (B) six tagged strains created. Wild type check was performed using SPAC2H10.01 primers in A and SPAC11D3.17 primers in B. For each strain the gel shows the PCR products of outside-outside primers, right-flank primers and left-flank primers from left to right respectively. Gel A shows that phx1, hsr1, moc3, and rsv2 deletion strains have both right and left flanks, while the SPAC2H10.01 deletion only shows a product for the right flank. Gel B shows left and right flanks for all six tagged strains, SPAC2H10.01 denoted as U1 and SPAC11D3.17 denoted at U2.

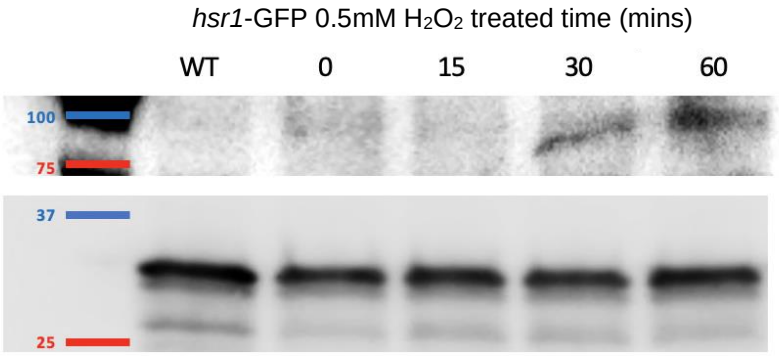



Figure 16: Western blot analysis of GFP tagged strains.

*hsr1*-GFP strain 0.5mM $H_2O_2$ treated time course. The band at ~91kDa not seen in the wild type control but seen strongly at 30 and 60 minutes after treatment in A shows the GFP tagged Hsr1 which accumulates in the cell in response to 0.5mM $H_2O_2$ treatment. All lanes show bands of even intensity for Cdc2 (~34kDa) which was visualised as a loading control.

### 3.3.2 Caffeine and rapamycin stress spot tests

Stress spot tests were performed on YES plates and YES plates with caffeine (10mM) and rapamycin (100ng/ml) to test the deletion strains for sensitivity or resistance to caffeine and rapamycin treatment, a key ageing-associated stress. As seen in figure 17, the *rsv2*Δ and *hsr1*Δ strains were resistant to caffeine and rapamycin treatment, showing no growth phenotype on YES but increased growth to the JB22 wild type strain on YES plates with caffeine (10mM) and rapamycin (100ng/ml).

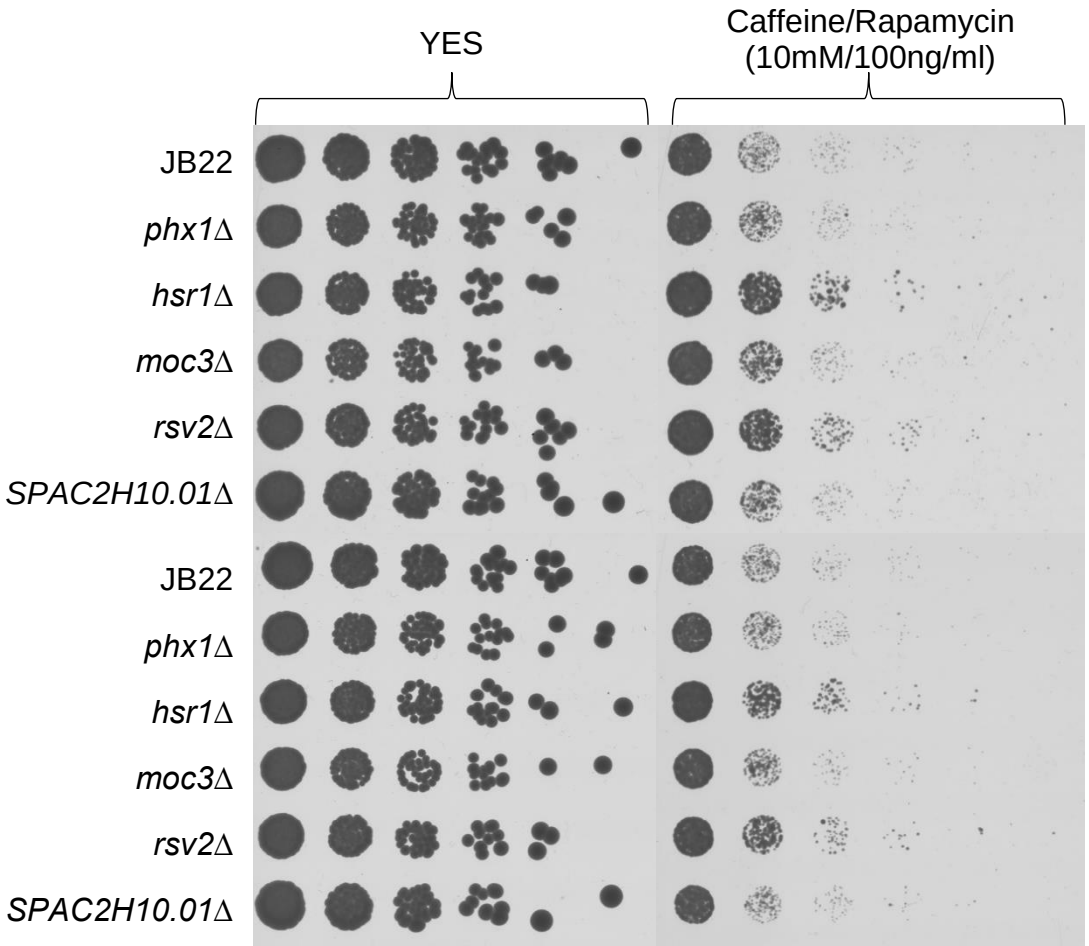



Figure 17: Stress spot test results.

YES agar with and without caffeine (10mM) and rapamycin (10ng/ml) showing *hsr1*Δ and *rsv2*Δ are growing better on the caffeine/rapamycin plate than the JB22 wild type control suggesting that these strains are resistant to caffeine/rapamycin treatment. *phx1*Δ, *moc3*Δ and *SPAC2H10.01*Δ showed no change in growth from the wild type control.

### 3.3.3 Chromatin immunoprecipitation sequencing (ChIP-Seq)

Following on from the novel fining that the *hsr1*Δ strain is resistant to caffeine and rapamycin treatment, Hsr1 was selected for further study in the form of target identification through ChIP-Seq. The ChIP-Seq experiment was constructed over a time course of $H_2O_2$ treatment. Since *hsr1* is known to be responsive to oxidative stress (Chen et al., 2008), and the *hsr1*-GFP strain showed increased levels of Hsr1-GFP protein during the western blot (figure 16), ChIP-Seq was then used to investigate Hsr1 binding during $H_2O_2$ treatment.

To create the time course, samples for the immunoprecipitations (IPs) were taken at time 0 (untreated), time 30 (30 minutes post treatment) and time 60 (60 minutes post treatment). The experimental IPs consisted of 2 repeats of the hsr1-GFP strain using an anti-GFP antibody. As a control, IPs were created from two more repeats of the hsr1-GFP strain but using an anti-HA antibody which cannot pull the GFP protein specifically. This is used to remove non-specific background binding during the analysis process.

The sequencing results from the ChIP-Seq experiment were assessed for read quality and trimmed to 75bp to remove the low read quality found at the ends of the sequences. These trimmed reads were mapped to the genome using bowtie2 and then filtered to remove any non-uniquely mapped reads. Peaks were then called using MACS2 where the HA control IPs were used as a background control for the GFP IPs for each repeat. The peaks for each repeat were then merged to leave only those common to both repeats for each time point. This set was used as the high confidence peaks.

The high confidence peaks at each time point were then analysed to look for overlap and create gene lists of biological interest. As seen in figure 18, there were 12 targets common to all time points, 4 exclusive to time 30, 5 common to times 30 and 60, and 3 common to times 60 and 0.

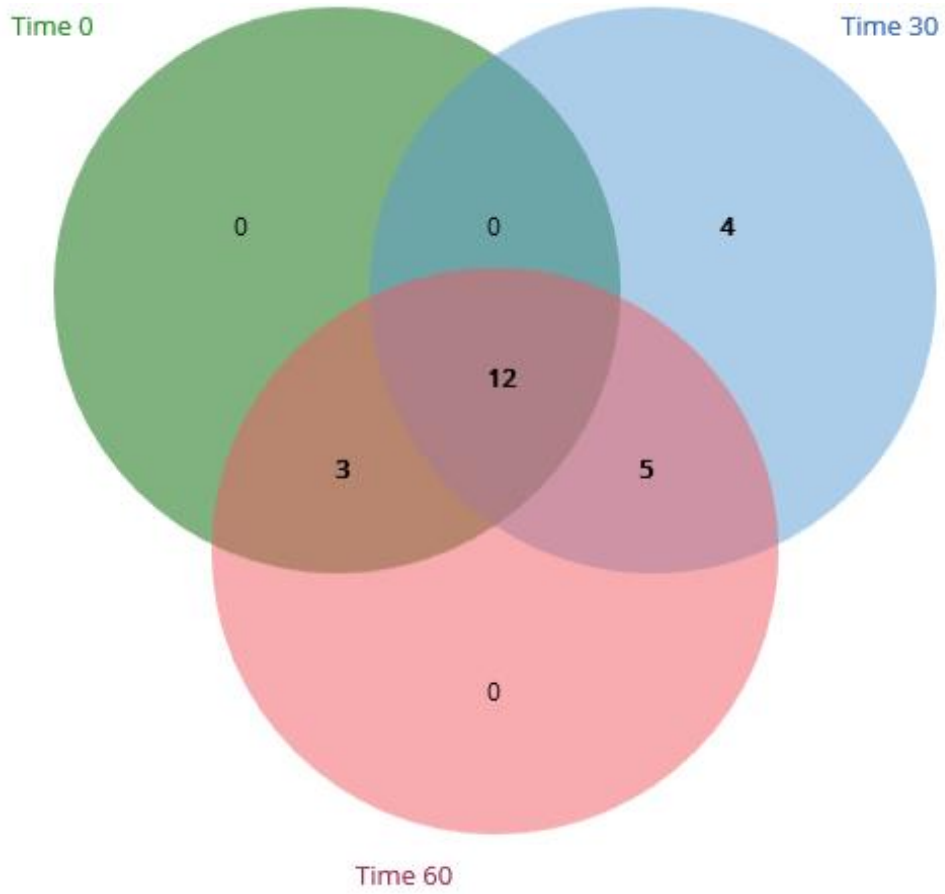



Figure 18: Venn diagram of high confidence Hsr1 binding sites for each time point.

Diagram shows the overlap between the high confidence peaks found in the untreated time 0 sample, the sample taken after 30 minutes of $H_2O_2$ treatment, and the sample taken 60 minutes after $H_2O_2$ treatment.

The first gene list of interest is the 9 targets which appear at 30 minutes post treatment but not in the time 0 control. 5 of which also appear at 60 minutes post treatment. The details of these targets can be found in table 19, they consist of 4 known and 4 unknown genes.

Table 20: Summary of targets found at 30 minutes post treatment but not in the time 0 control, including gene names, region of binding, and *PomBase* gene product annotation. The targets also found at 60 minutes post treatment are highlighted in grey.

| **Gene** | **Region** | ***PomBase* Product** |
|---|---|---|
| SPAP11E10.01 | Promoter | Ornithine cyclodeaminase-like protein |
| SPAC17A2.10c | Promoter | Unknown |
| SPATRNAALA.04 | Distal intergenic | tRNA Alanine |
| *pfl2* / SPAPB15E9.01c | Promoter | Cell surface glycoprotein, flocculin |
| *tdh1* / SPBC32F12.11 | Promoter | Glyceraldehyde-3-phosphate dehydrogenase |
| *ssn6* / SPBC23E6.09 | Intron | Transcriptional corepressor |
| SPCC320.03 | Promoter | DNA-binding transcription factor |
| *gcd1* / SPCC794.01c | Promoter | Glucose dehydrogenase |
| SPCTRNAASP.07 | Distal intergenic | tRNA Asparagine |

The mapped reads were also visualised on the *PomBase* genome browser to look at the structure and positioning of the peaks returned by the peak call using MACS2. Figures 19 and 20 show the visualised peaks for *tdh1* and *pfl2* on the genome browser. They shows clearly that the peaks lie immediately upstream of the genes, in their promoter regions, as identified in the peak call.

Figure 20 also shows smaller peaks upstream of the *pfl2*, potentially in the upstream promoter region of SPAPB15E9.02c. These peaks don't appear in the HA control IPs, or in the time 0 GFP IPs, but they did not pass the threshold during the peak call with MACS2 to be identified as a target.

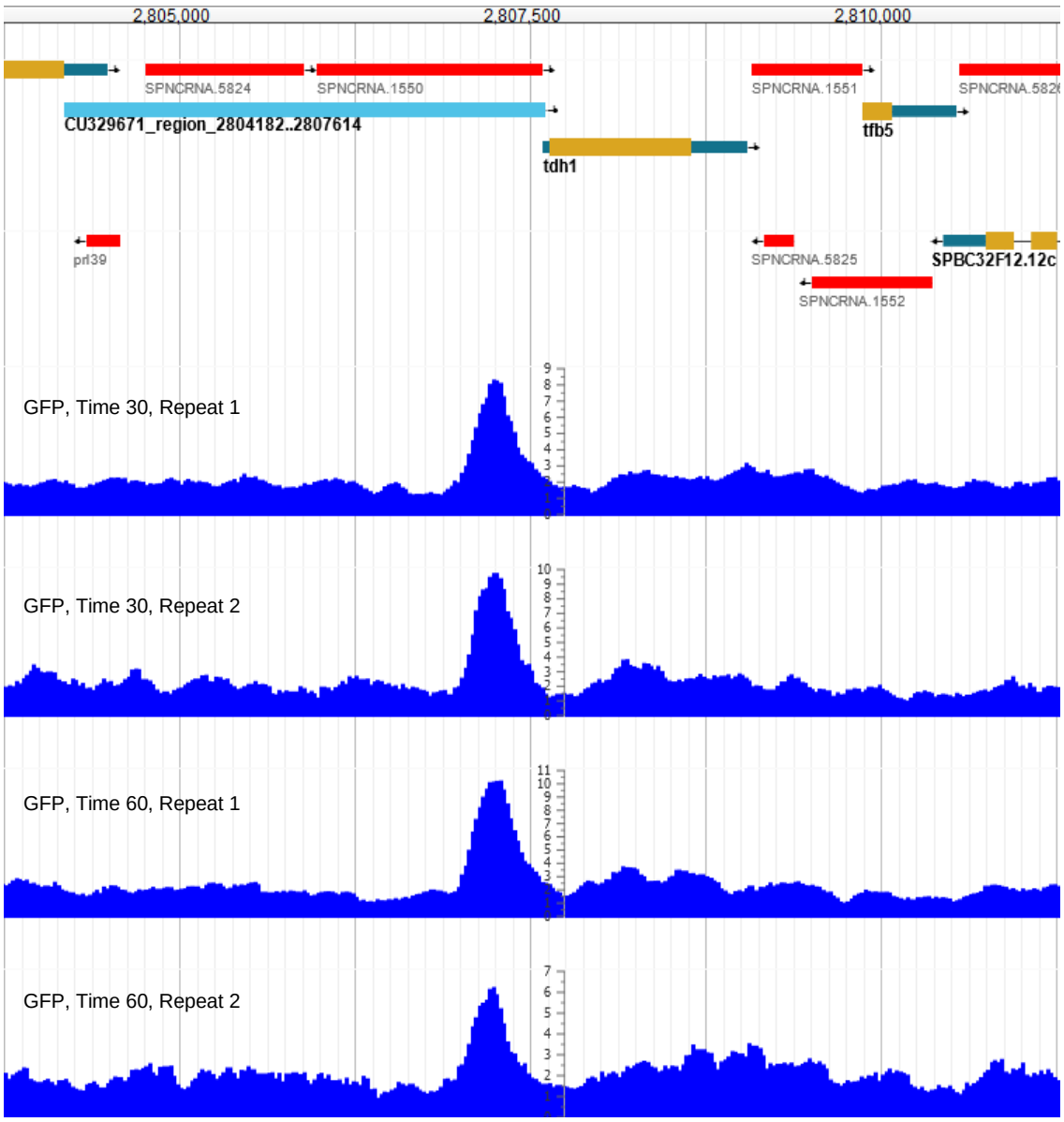



Figure 19: Visualisation of peaks in the promoter region of *tdh1*.

Using the mapped reads in the *PomBase* genome browser, defined peaks are seen in both repeats at time 30 and time 60 at the promoter region of gene *tdh1*.

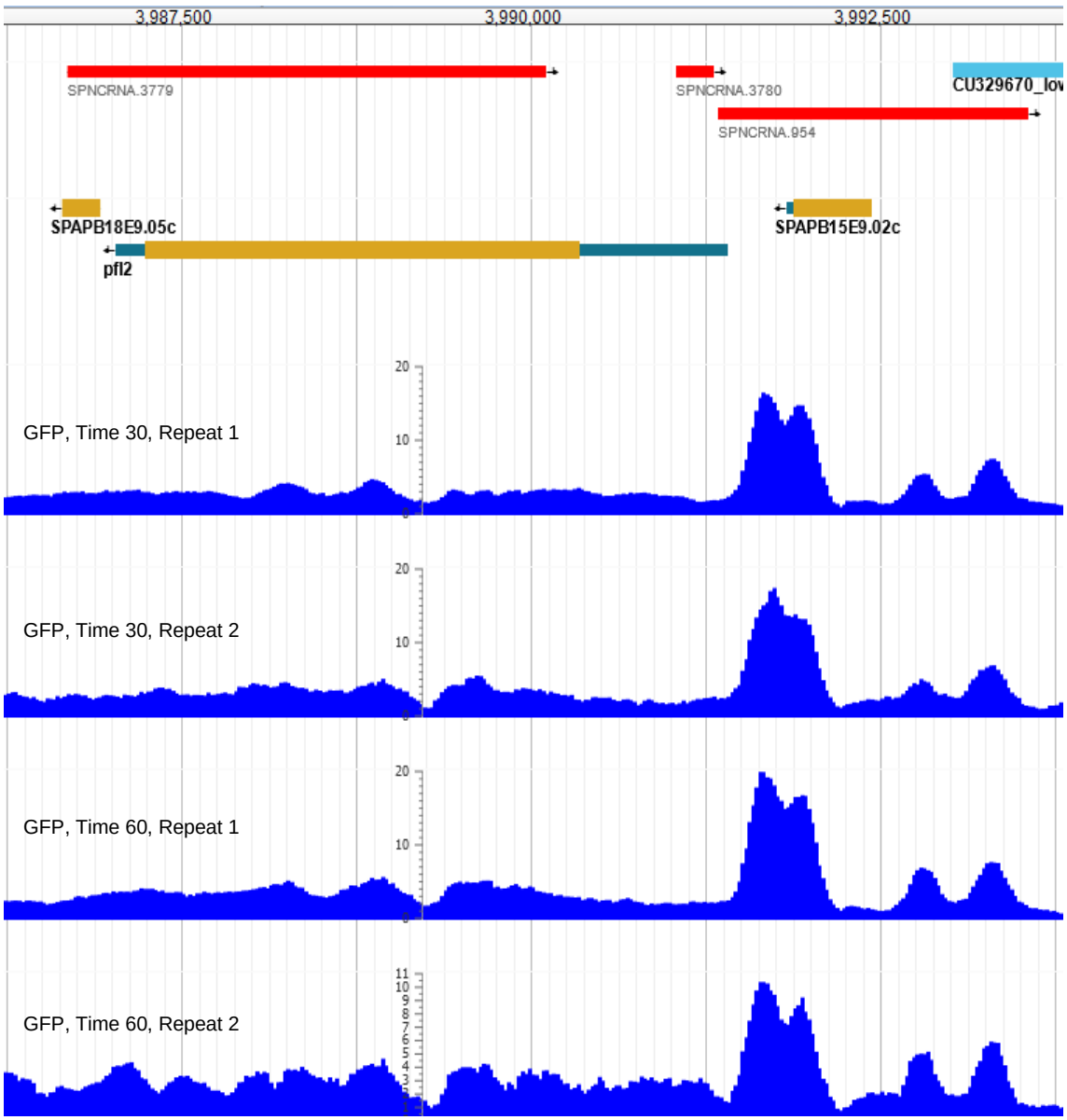



Figure 20: Visualisation of peaks in the promoter region of *pfl2*.

Using the mapped reads in the *PomBase* genome browser, defined peaks are seen in both repeats at time 30 and time 60 at the promoter region of gene *pfl2*. Smaller peaks are seen upstream of SPAPB15E9.02c, potentially in the promoter region but these peaks did not meet the threshold for the peak call with MACS2.

Using the mapped reads in the genome browser, it was visually confirmed that there was no peak at adh8/SPBC1773.06c and SPAC23D3.05c, two genes identified as having a lowered response to $H_2O_2$ treatment in *hsr1*Δ mutants. This suggests that their response is not caused by the direct binding of Hsr1.

The second list of interest is the 3 targets which appear at 60 minutes post treatment and in the time 0 control but not at 30 minutes post treatment. The details of these targets can be found in table 20, they consist of known and unknown genes.

Table 21: Summary of targets found at 60 minutes post treatment and in the time 0 control, but not at 30 minutes post treatment, including gene names, region of binding, and *PomBase* gene product annotation.

| Gene | Region | *PomBase* Product |
|---|---|---|
| SPRRNA.43 | Promoter | 18S ribosomal RNA |
| SPRRNA.44 | Promoter | 18S ribosomal RNA |
| SPRRNA.46 | Promoter | 18S ribosomal RNA |

The high confidence target genes at all three time points were used for gene ontology enrichment analysis. The Hsr1 targets were found to be enriched for flocculation, aggregation of unicellular organisms, cell aggregation and cytoplasmic translation (table 22). Flocculation was found to be almost 100 times more common in the Hsr1 high confidence targets than in the whole genome with the related processes of aggregation of unicellular organisms and cell aggregation similarly amplified. Cytoplasmic translation was found to be 6 times more frequent in the targets than in the whole genome and genes with this process accounted for almost half of all the high confidence targets.

Table 22: Gene ontology (GO) enrichment analysis results for a list of the high confidence targets of Hsr1 including percentage list frequency and percentage background frequency for each biological process found to be enriched in the list.

| GO Biological Process | List Frequency | Background Frequency |
|---|---|---|
| Flocculation | 16.67% | 0.17% |
| Aggregation of unicellular organisms | 16.67% | 0.20% |
| Cell aggregation | 16.67% | 0.20% |
| Cytoplasmic translation | 41.67% | 6.58% |

The list of Hsr1 high confidence binding site targets identified at timepoint 30 was used to search for binding motifs. One motif appeared in the majority of the targets and was identified as a likely binding motif for Hsr1. The motif and the details of its occurrences in the first list of interest can be found in figure 21.

A


B

| Gene | Motif Locations | Sequence |
|---|---|---|
| SPAP11E10.01 promoter | | GGGGGTAG |
| SPAC17A2.10c promoter | | GGAGGTAG |
| SPATRNAALA.04 distal intergenic | | TGTGGAAA |
| *pfl2* / SPAPB15E9.01c promoter | | TGAGGAAA |
| *tdh1* / SPBC32F12.11 promoter | | TGAGGTAG |
| *ssn6* / SPBC23E6.09 intron | | GGGTGAAG |
| SPCC320.03 promoter | | TGGGGTAG |
| *gcd1* / SPCC794.01c promoter | | GGGGGAAG |

Figure 21: Hsr1 binding motif details.

The binding motif identified for Hsr1 (A), and the binding motifs found within the high confidence targets found at t30 but not t0 (B).

This binding motif was then used to search for similar motifs in *S. cerevisiae* and humans. The search found similar binding motifs for the *S. cerevisiae* transcription factors Met32 and Rme1 (figure 22), involved in sulfur metabolism and meiosis (Carrillo et al., 2012, Toone et al., 1995), and the human transcription factors RUNX1, RUNX2 and RUNX3 (figure 23), involved in cell proliferation, differentiation and cell lineage specification (Mevel et al., 2019).

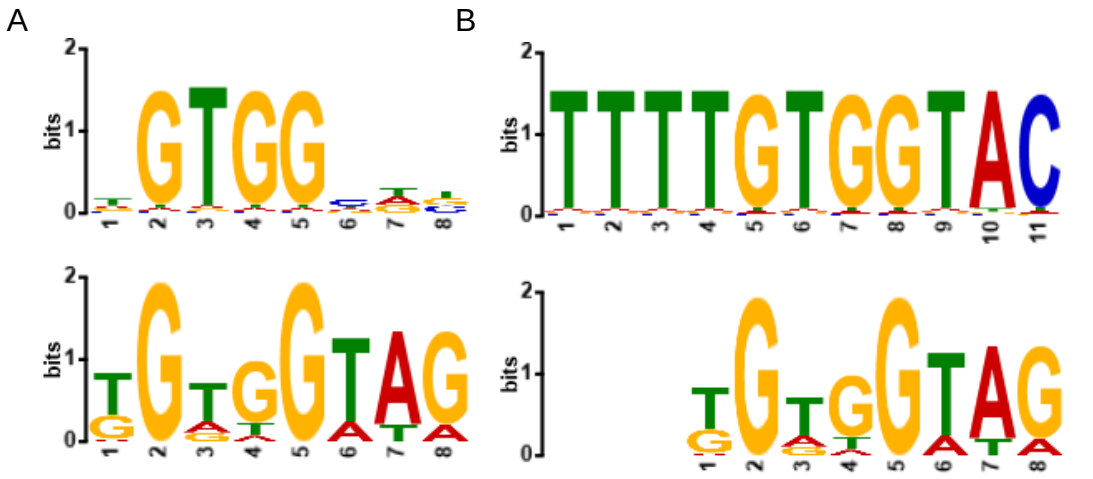



Figure 22: *S. cerevisiae* transcription factors with similar binding motifs to Hsr1: Met32 (A) and Rme1 (B)

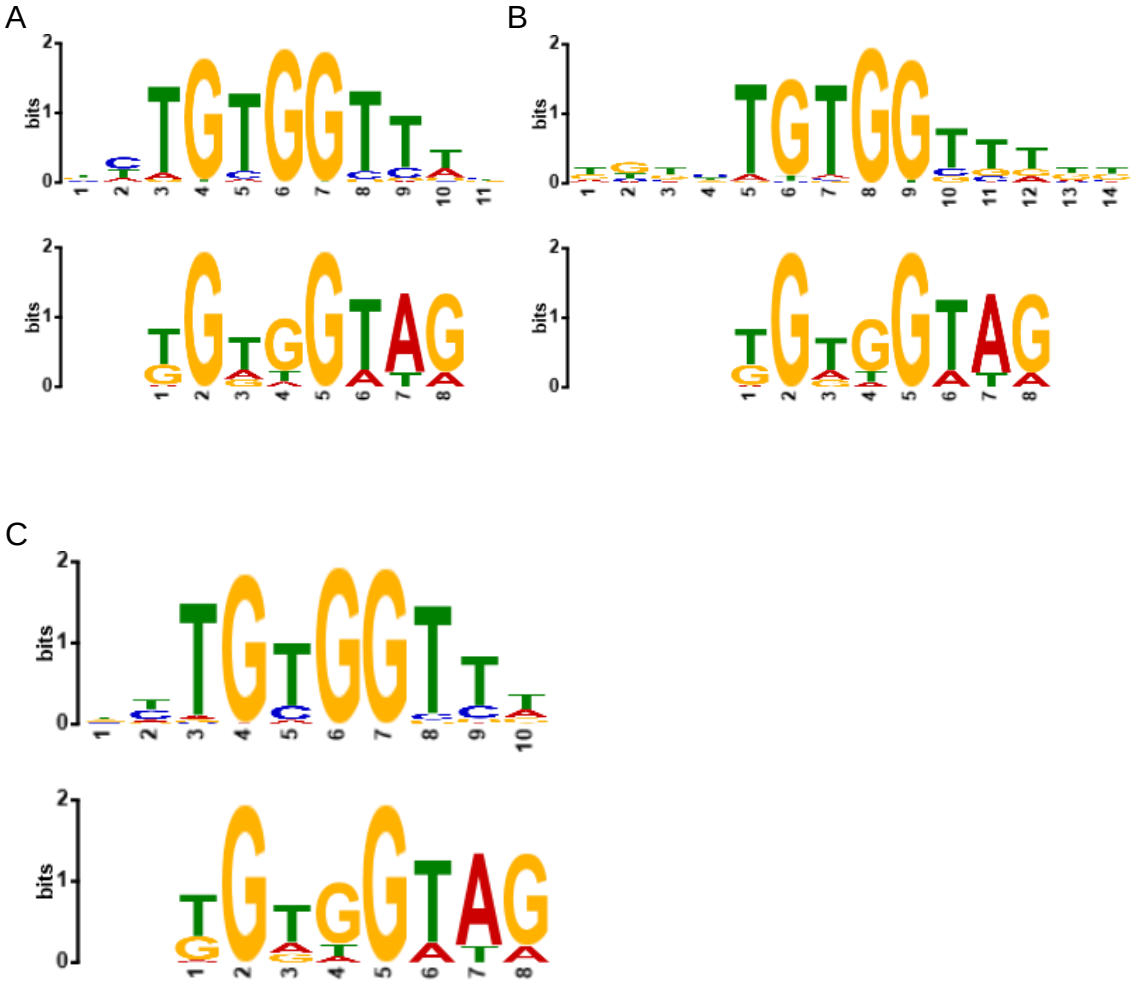



Figure 23: Human transcription factors with similar binding motifs to Hsr1: RUNX1 (A), RUNX2 (B) and RUNX3 (C).

### 3.3.4 Synthetic genetic array (SGA)

After the identification of Hsr1 binding targets, the involvement of Hsr1 in other pathways within the cell can be further elucidated with the investigation of genetic interactions with *hsr1*. A synthetic genetic array (SGA) creates an array of double mutants with deletions from the *hsr1*Δ strain and other non-lethal deletions from the Bioneer deletion library (Baryshnikova et al., 2010).

The SGA was carried out under the conditions of both YES and with caffeine and rapamycin stress to look for positive and negative genetic interactions with *hsr1*. Caffeine and rapamycin inhibit the ageing-associated target of rapamycin (TOR) pathway and can extend lifespan in fission yeast (Hillson et al., 2018), so genetic interactions exclusive to the caffeine and rapamycin condition can give insight into cellular processes linked to TOR inhibition and lifespan extension.

The scanned colonies for the SGA are first processed to extract the colony size using the number of pixels in the scanned image. These colony sizes are then analysed to remove the small colonies which denote spots where there was no growth, only the transferred material. These colonies are identified as a peak which is detached from the main normal distribution peak of colony sizes for each control sample. Figure 25 shows the segregated peak which is removed from each control sample. The genetic interactions were then calculated as log2 fold change between the average colony size of the three repeats of the hsr1Δ double mutants and the ade6Δ control double mutants. During the calculation, the p-value was also recorded.

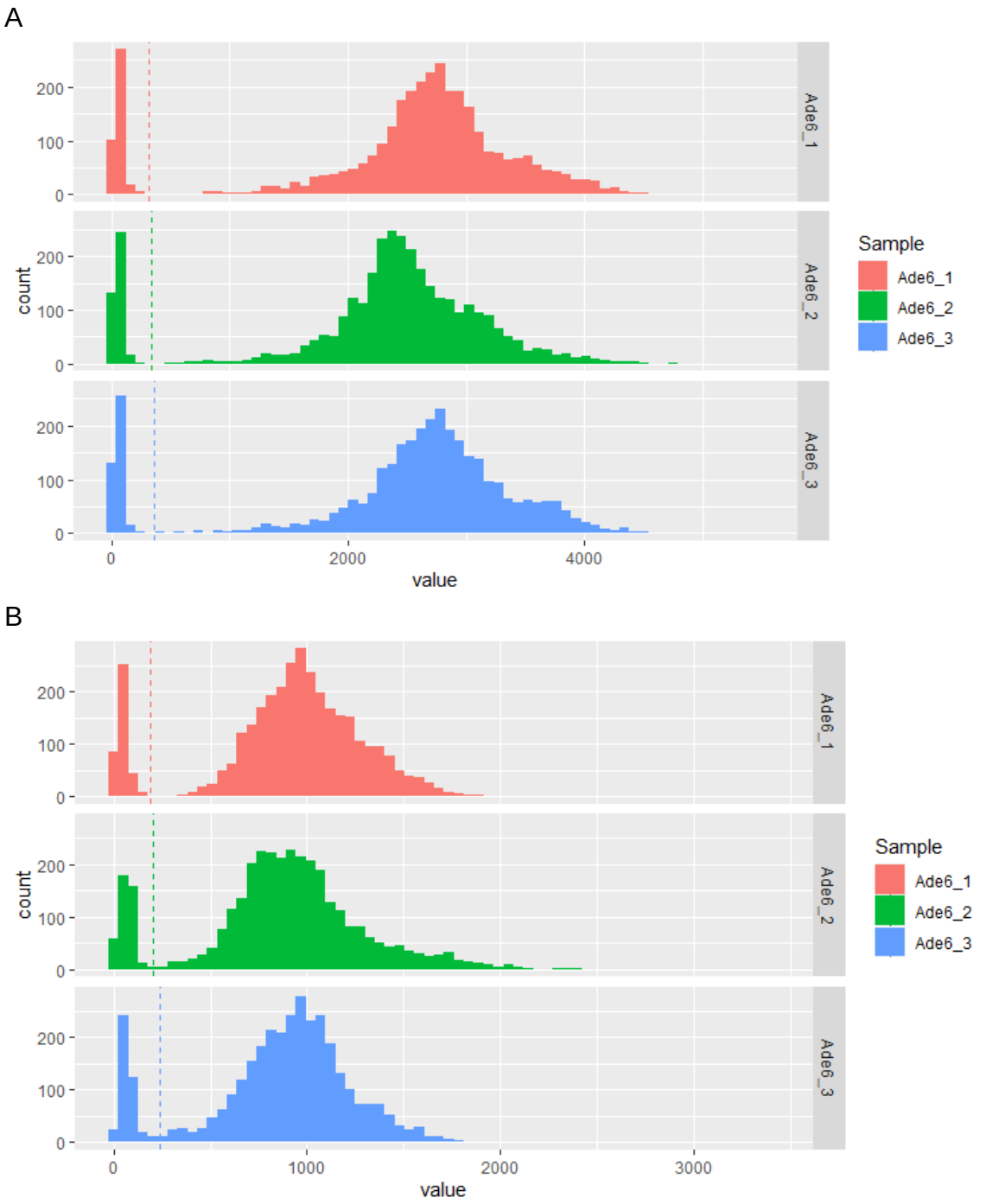



Figure 24: Small colony exclusions.

For the SGAs on YES(A) and caffeine and rapamycin (B), the distribution of colony sizes is shown as a histogram for each repeat of the control sample. The detached peak of small colonies is separated by a dotted line where the exclusion takes place to remove the small colonies from the rest of the data.

The linked loci for *hsr1* and *ade6* were identified as the genes with loci ±500,000bp from the loci of *hsr1* and *ade6*. Double mutants with linked loci can produce artificially strong negative and positive interactions which skew the

identification of true positive and negative interactions across the rest of the genome. Figure 24 shows the linked genes identified on each chromosome, highlighted in red for both the YES and caffeine and rapamycin SGAs. Decreased colony fitness can be seen around the locus of *hsr1* on chromosome I and increased colony fitness can be seen around the locus of *ade6* for both SGAs. Linked loci genes were then removed prior to the identification of positive and negative interaction hits.

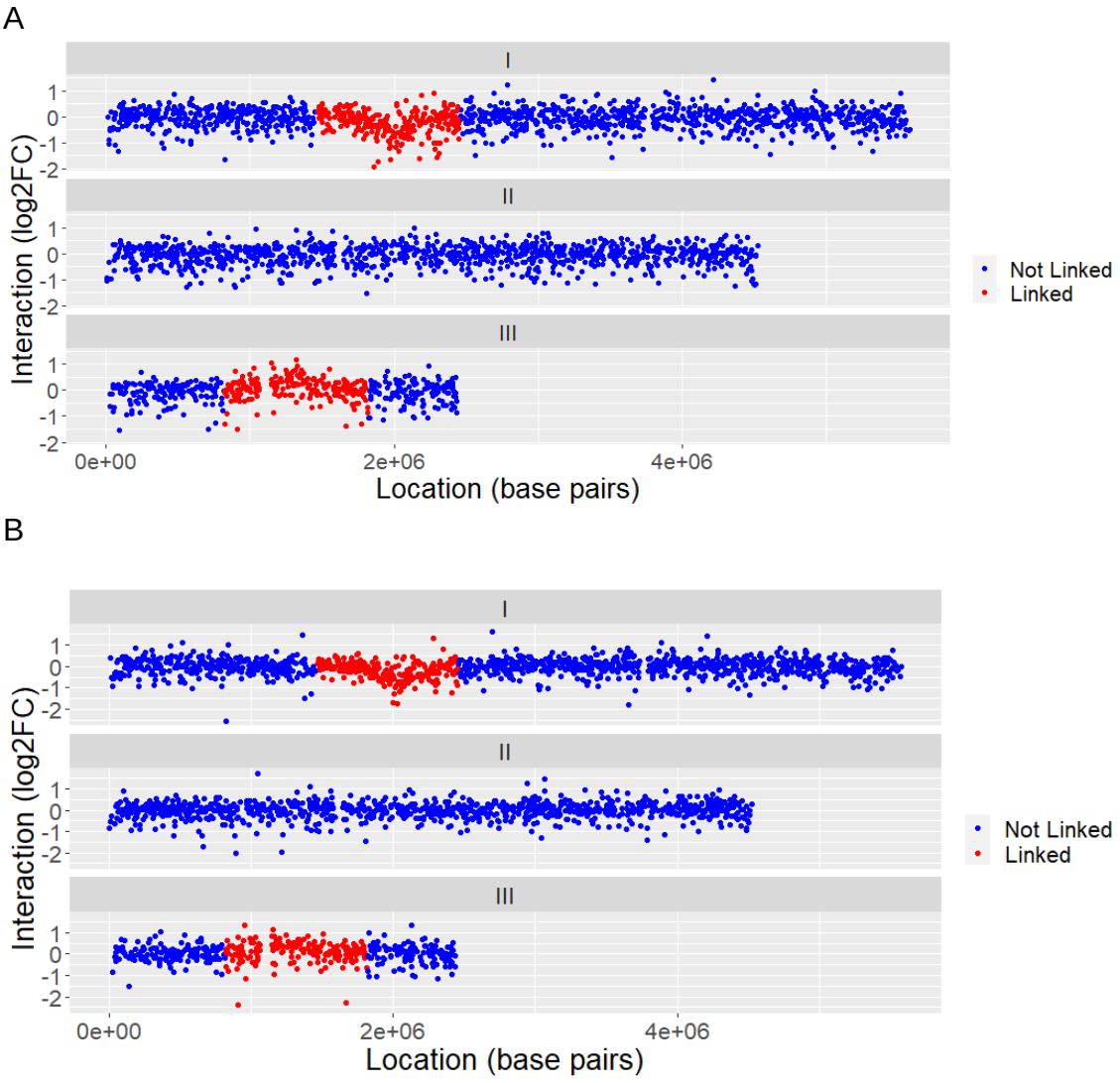



Figure 25: Linked loci to *hsr1* and *ade6*.

$Log_2$ fold change interaction scores are shown plotted against gene locus for the SGAs under YES (A) and caffeine and rapamycin (B) conditions. Highlighted in red are the loci for the genes ±500,000bp from the loci of *hsr1* and *ade6.* These genes are identified as linked loci for removal from the SGA results before identification of positive and negative interaction hits.

To visualise positive and negative interaction hits, two volcano plots were produced. Both plots had a p-value cut-off of $p<0.05$ so that only interactions with a confidence of 95% or more were considered as hits. Two $\log_2$ fold change cut-offs were visualised: ±1 (2 times as big or small) and ±0.5 (1.5 times as big or small). The volcano plots in figure 26 show the interaction hits for both cut-offs for the SGA on YES. With the more conservative $\log_2$ fold change cut-off of ±1, only two positive interaction hits were identified. By moving the $\log_2$ fold change cut-off to ±0.5, the number of positive interaction hits increased significantly, giving a less conservative but more inclusive group of interactions.

The volcano plots in figure 27 show the interaction hits for both cut-offs for the SGA on YES plus caffeine and rapamycin. With the more conservative $\log_2$ fold change cut-off of ±1, only one positive interaction hit was identified. By moving the $\log_2$ fold change cut-off to ±0.5, the number of positive interaction hits increased significantly, giving a less conservative but more inclusive group of interactions.

Hit lists of interactions at both $\log_2$ fold change cut-offs (±1 and ±0.5) were generated, both with a p-value cut off of $p<0.05$. These lists were stored for further analysis. Initially, the lists were searched for gene ontology enrichment against both the background of the genome and the background of all the interactions for that SGA, but this returned no enrichment for any of the lists.

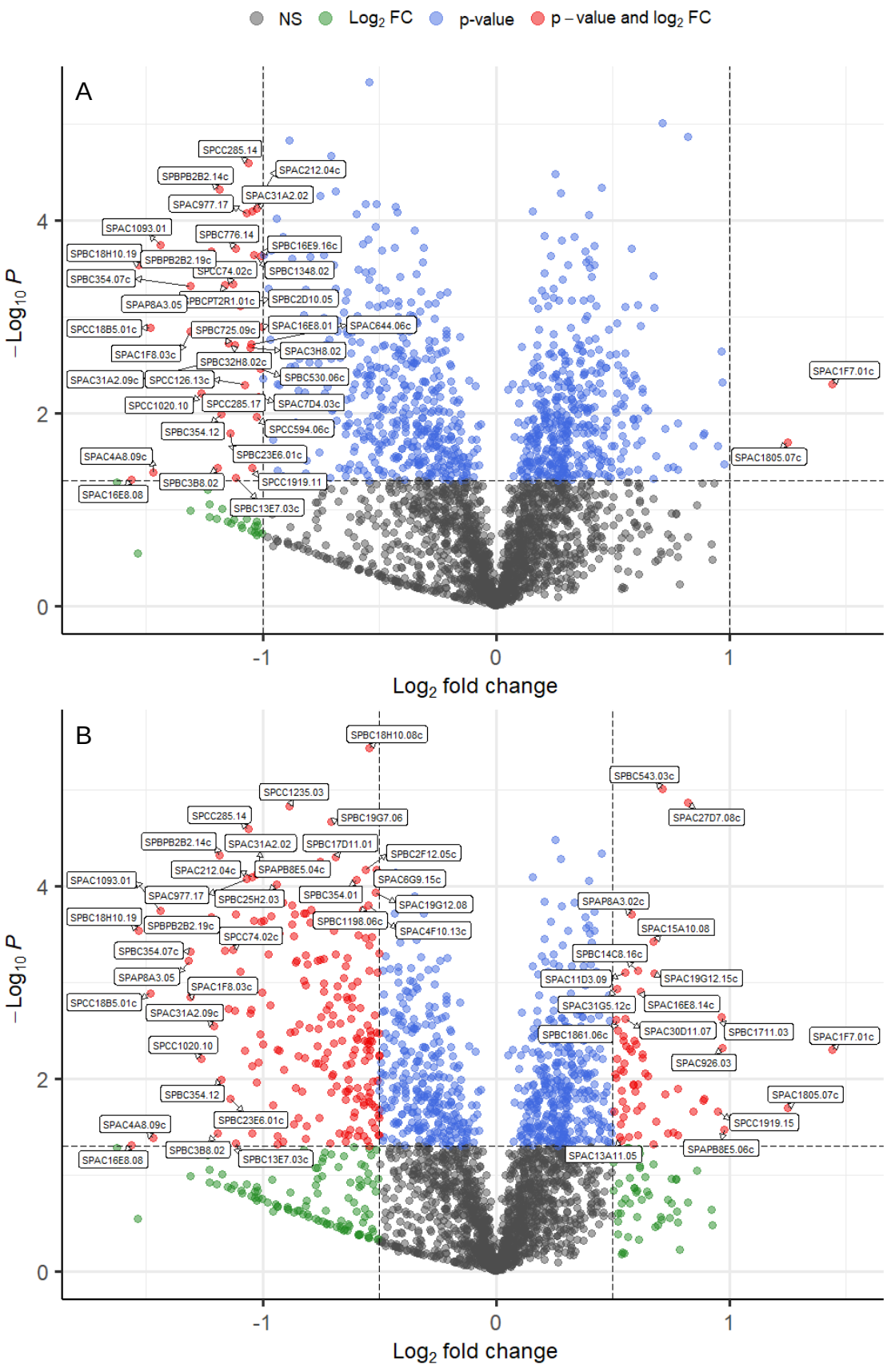



Figure 26: Volcano plots showing interaction hits for the YES SGA

Showing interactions for $\log_2$ fold change cut-offs of ±1(A) and ±0.5(B) with a p-value cut-off of $p<0.05$.

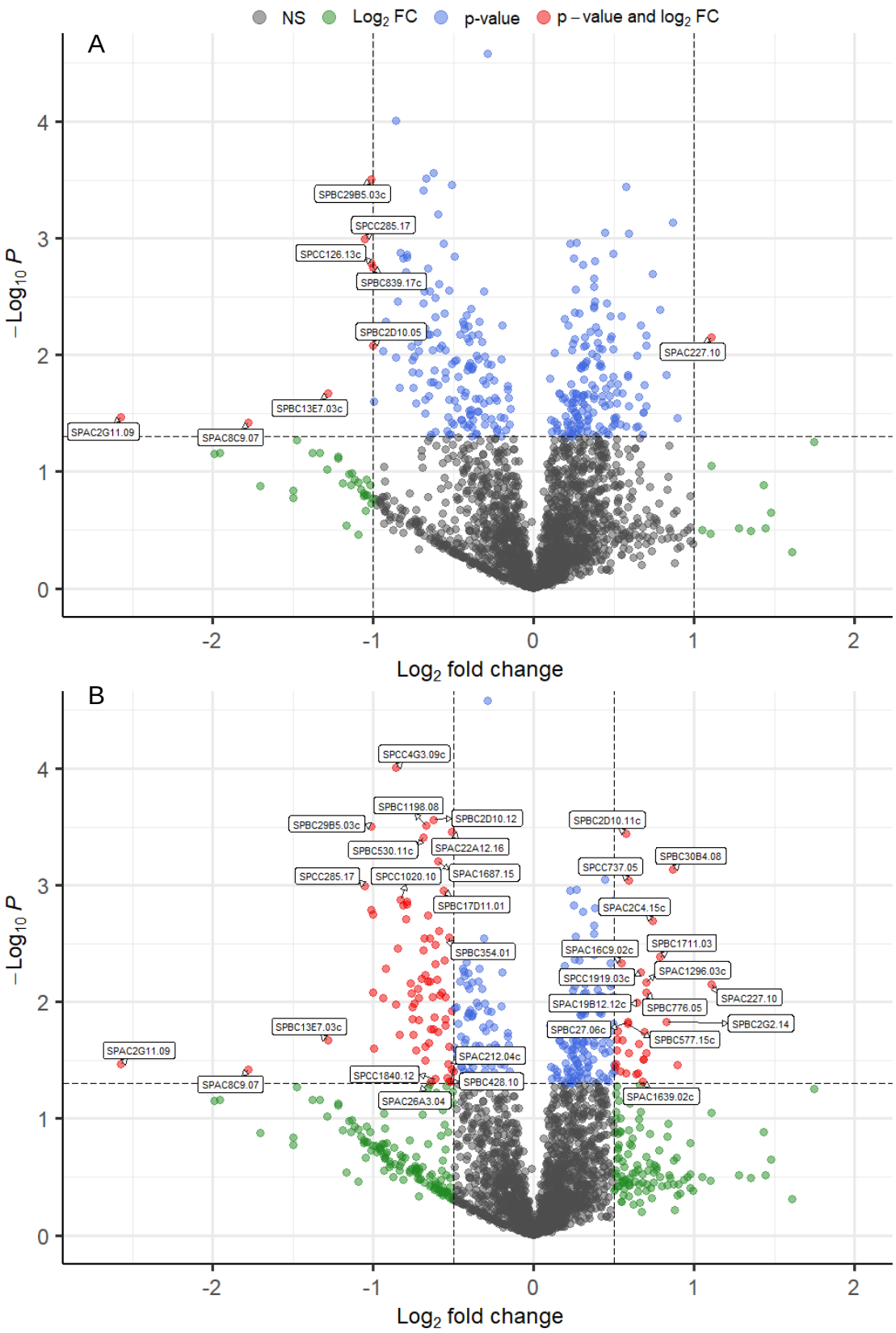



Figure 27: Volcano plots showing interaction hits for the caffeine and rapamycin SGA

Showing interactions for $\log_2$ fold change cut-offs of ±1(A) and ±0.5(B) with a p-value cut-off of $p<0.05$.

The hit lists were then analysed for overlap using Venn diagrams (figure 28). For both $log_2$ fold change cut-offs, ±1 and ±0.5, there was no overlap between the positive and negative interactions, even between the YES and caffeine and rapamycin conditions. There were overlaps between the positive interactions for both conditions at the $log_2$ fold change cut-off of ±0.5 and the negative interactions for both conditions at the $log_2$ fold change cut-offs of both ±1 and ±0.5.

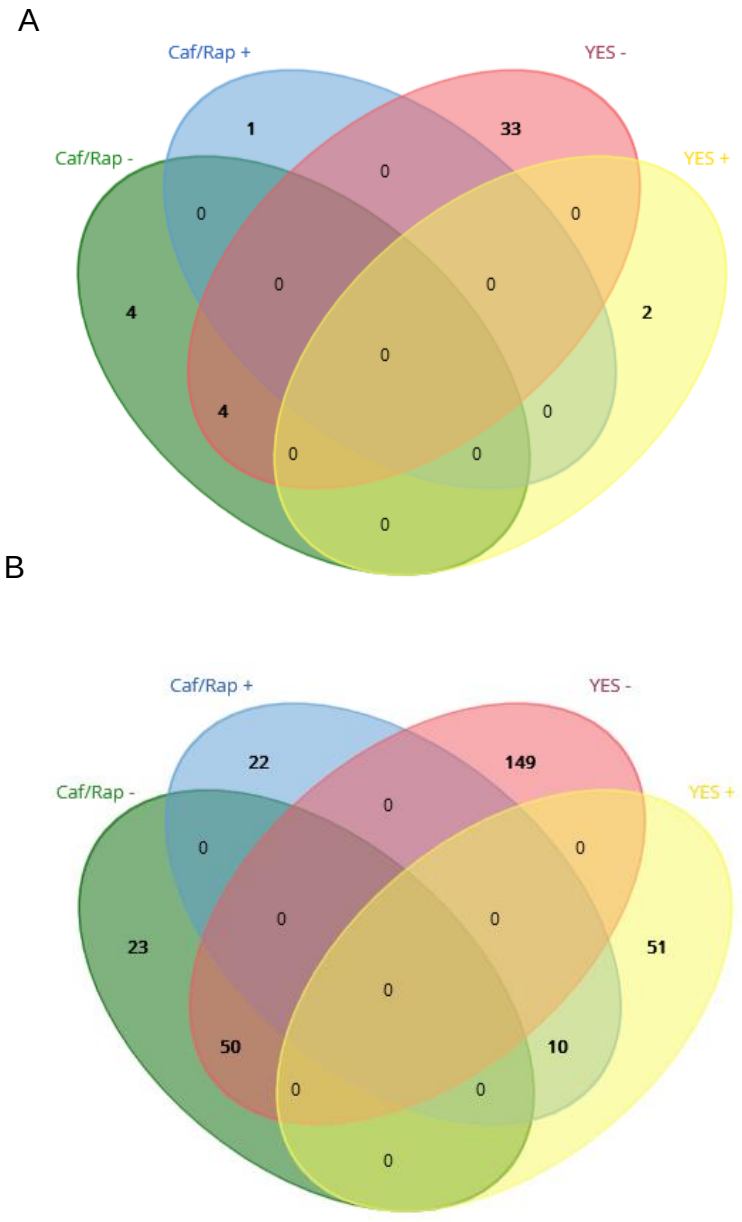



Figure 28: Overlap in hit lists using a $log_2$ fold change cut-off of ±1 (A) and $log_2$ fold change cut-off of ±0.5 (B).

Gene lists were produced for positive and negative interactions which appear under the ageing-associated condition of caffeine and rapamycin but not in YES (tables 23 and 24). Since there are far more negative interactions than positive interactions, the gene list for negative interactions was created using the more conservative $\log_2$ fold change cut-off of ±1, producing a list of 4 genes (figure 28A). For positive interactions, using the conservative cut-off returned only one gene (figure 28A), so the more inclusive $\log_2$ fold change cut-off of ±0.5 was used, creating a list of 22 genes (figure 28B).

Table 23: Details of genes with negative interactions with *hsr1* in caffeine and rapamycin but not YES, with a p-value < 0.05, and a log2 >1

| **Gene** | ***PomBase* Product** |
|---|---|
| *rpl26*/SPBC29B5.03c | 60S ribosomal protein |
| *fyv7*/SPAC8C9.07 | rRNA processing protein |
| SPAC2G11.09 | Calcium ion transmembrane transporter |
| *fkh1*/SPBC839.17c | FKBP-type peptidyl-prolyl cis-trans isomerase |

Table 24: Details of genes with positive interactions with *hsr1* in caffeine and rapamycin but not YES, with a p-value < 0.05, and a log2 >0.5

| **Gene** | ***PomBase* Product** |
|---|---|
| *rpp203*/SPAC1071.08 | 60S acidic ribosomal protein |
| *pfd2*/SPAC227.10 | Prefoldin subunit |
| SPAC22A12.14c | BSD domain protein |
| *amk2*/SPCC1919.03c | Serine/threonine protein kinase AMPK (beta) regulatory subunit |
| *ubx2*/SPAC2C4.15c | UBX domain protein |
| *pub3*/SPBC16E9.11c | HECT-type ubiquitin-protein ligase E3 |
| *aar2*/SPAC3H5.04 | U5 snRNP-associated protein |
| *sxa2*/SPAC1296.03c | Serine carboxypeptidase |
| SPBC887.17 | Plasma membrane guanine and adenine transmembrane transporter |
| SPCC330.03c | NADPH-hemoprotein reductase |
| SPCC737.05 | Peroxin Pex28/29 |

| | |
|---|---|
| *yip11*/SPAC19B12.12c | SMN complex subunit |
| *mta1*/SPAC16C9.02c | S-methyl-5-thioadenosine phosphorylase |
| *sim3*/SPBC577.15c | CENP-A chaperone, NASP family |
| *icp55*/SPAC12B10.05 | Mitochondrial intermediate cleavage peptidase |
| *rec24*/SPAC1952.15c | Meiotic recombination protein |
| *ale1*/SPBC16A3.10 | Membrane bound O-acyltransferase, MBOAT |
| SPBC428.12c | Peptidyl-prolyl cis-trans isomerase E |
| *gpc1*/SPBC776.05 | Glycerophosphocholine acyltransferase (GPCAT) |
| *bun62*/SPAC12B10.03 | WD repeat protein Wdr20 |
| *mug73*/SPCC31H12.02c | Multispanning 7TM plasma membrane rhodopsin family protein, implicated in signalling |
| *trk2*/SPAC1639.02c | Plasma membrane potassium ion transmembrane transporter |

Interactions were also searched for all of the genes known to be associated with Hsr1 from previous experiments. Many did not overlap with the SGA data either due to linked loci, small colony exclusions or high p-value of interactions, but the data available for known associates is summarized in table 25. It shows that most associated genes had a negative interaction and that *php5* and *gcd1* had significant negative interactions under the YES condition.

Table 25: SGA results for genes with known *hsr1* involvement

| | **YES** | | **Caffeine and Rapamycin** | |
|---|---|---|---|---|
| **Gene** | **$Log_2$ FC** | **P-value** | **$Log_2$ FC** | **P-value** |
| tdh1/ SPBC32F12.11 | 0.24 | 0.04 | | |
| SPCC320.03 | -0.40 | 0.04 | -0.25 | 0.04 |
| gcd1/SPCC794.01c | -0.86 | 0.00 | | |
| php2/SPBC725.11c | -0.16 | 0.03 | | |
| php5/SPBC3B8.02 | -1.19 | 0.04 | | |
| mbx2/ SPBC317.01 | -0.34 | 0.02 | | |

## 3.4 Discussion

### 3.4.1 Generation of deletion and GFP tagged strains

Deletion strains for five of the six chosen transcription factors were transformed and selected by PCR. As shown in figure 15A the deletion strains for *phx1, hsr1, moc3* and *rsv2* were all identified as having correctly oriented deletion constructs due to positive right and left flank PCR. The deletion strain for SPAC2H10.01 had only the right flank however this was repeated and confirmed. The presence of only one flank band is still enough to suggest this deletion construct is correctly orientated. GFP tagged strains were transformed and selected for all six chosen transcription factors. In figure 15B left and right flanks can be seen for all strains showing that the GFP tag construct is correctly orientated in all strains.

Western blot analysis was used to confirm the presence of the GFP tagged protein in the *hsr1*-GFP strain. Figure 16 shows a band 30 minutes and 60 minutes after treatment with $H_2O_2$ at the correct size for Hsr1-GFP suggesting that Hsr1 was produced in the cell in response to $H_2O_2$ treatment. The Cdc2 control band is visible at all timepoints with consistent intensity. The consistent intensity of the Cdc2 band shows that the samples all contained a consistent quantity of protein and so the increase in Hsr1 has biological significance.

### 3.4.2 Caffeine and rapamycin stress spot tests

Stress spot tests were carried out to check for an ageing-related growth phenotype. Growth phenotypes in the presence of caffeine and rapamycin suggest an involvement with the ageing-associated target of rapamycin (TOR) pathway as caffeine and rapamycin treatment inhibits TOR activity in fission yeast.

The *rsv2*Δ strain is seen to have a resistant phenotype to the caffeine and rapamycin treatment as it has more growth than the JB22 control despite showing identical growth on the untreated YES control plate. This is the expected result as the Bioneer deletion of *rsv2* had previously been shown to be resistant to caffeine and rapamycin (Rallis et al., 2014).

The spot test also showed that the *hsr1*Δ strain was resistant to caffeine and rapamycin. In this test the resistance was equal to that of *rsv2*Δ but *hsr1*Δ had not been previously identified as resistant to caffeine and rapamycin, making this a novel finding in this work.

### 3.4.3 Targets of the transcription factor Hsr1

Initially, it is important to ensure that the ChIP-Seq data is of high quality so that subsequent conclusions can be considered reliable and valid. Quality control of the reads showed that all reads were of good quality up to 75bp so the reads were trimmed to this length to ensure the quality of the subsequent analyses. After mapping to the genome using bowtie2, the mapped reads were filtered to remove a small number of non-uniquely mapped reads to ensure the reliability of the peak call. The peak call with MACS2 was designed to remove any background of non-specific binding by using the samples created using the anti-HA as a background for those created with ant-GFP and after the peak call, the repeats for each time point were joined to keep only targets which appeared in both repeats. In this way ‘high confidence’ peaks were created which have background removed and two repeats. Use of these high confidence samples is a conservative method for analysing ChIP-Seq data which risks loss of peaks which have biological significance, but it allows for high-confidence conclusions to be drawn about the binding of Hsr1.

The annotated high confidence peaks showed that Hsr1 bound to *tdh1* at 30- and 60-minutes post treatment with $H_2O_2$, and to *gcd1* and *pfl2* at 30-minutes post treatment, but not to these genes at the untreated time 0 control. This shows that binding to these genes was in response to the oxidative stress caused by $H_2O_2$, which has been shown to activate Hsr1 (Chen et al., 2008).

*Hsr1 regulates transcription of glycolytic enzyme Tdh1 in response to oxidative stress*

*tdh1*, a high-confidence target gene, was identified as having oxidative-stress dependent binding with binding at 30- and 60-minutes post treatment with $H_2O_2$, but not in the untreated time 0 control. *tdh1* is a protein coding gene which codes for Tdh1, a glyceraldehyde-3-phosphate dehydrogenase (GAPDH) enzyme which catalyses the sixth step of the glycolytic pathway and has also been shown to be involved in the cellular response to oxidative stress (Morigasaki et al., 2008, Morigasaki and Shiozaki, 2010, Morigasaki and Shiozaki, 2013). During the multistep phosphorelay in response to oxidative stress in *S. pombe*, sensor histidine kinases Mak2 and Mak3 activate the phosphotransferase Mpr1 which activates the response regulator Mcs4. Mcs4 then goes on to trigger a MAPK cascade to activate the MAPK Spc1. Tdh1 has been shown to form a complex with Mcs4 in response to oxidative stress, which facilitates this phosphorelay. In response to $H_2O_2$ treatment, Tdh1 also undergoes oxidation of one of its cysteine residues which promotes its interaction with Mcs4. This cysteine residue modification is conserved in other organisms and may be a mechanism for Tdh1 to avoid irreversible oxidative inactivation. Hsr1 binding to the promoter region of *tdh1* in this work was induced by oxidative stress from $H_2O_2$ treatment.

Potentially, transcriptional regulation of *tdh1* by Hsr1 in response to oxidative stress could regulate the availability of Tdh1 for the phosphorelay complex with Mcs4, as well as regulating the glycolytic pathway by transcriptomic control of Tdh1. During oxidative stress, cells have been shown to redirect glycolysis through to the pentose phosphate pathway to generate reducing NADPH which can limit oxidation by reducing reactive oxygen species within the cell (Mullarky and Cantley, 2015). This redirection would require regulation of glycolysis and also of the pentose phosphate pathway.

*Hsr1 regulates transcription of glucose dehydrogenase Gcd1 in response to oxidative stress*

*gcd1*, a high-confidence target gene, was identified as having oxidative-stress dependent binding with binding at 30-minutes post treatment with $H_2O_2$, but not at 60-minutes post treatment or in the untreated time 0 control. *gcd1* is a protein coding gene which codes for Gcd1, an $NAPD^+$-dependent glucose dehydrogenase which has been shown to function as an alternative route into the pentose phosphate pathway in *S. pombe* (Corkins et al., 2017). Gcd1 is suggested to function as a shunt into the pentose phosphate pathway which bypasses the rate limiting enzyme Glu-6-P dehydrogenase. In this way it would be possible to Gcd1 to function as a shunt into this pathway in response to oxidative stress in the cell, regulated by Hsr1 transcriptional control. We can propose that Hsr1 regulates the redirect from glycolysis to the pentose phosphate pathway in fission yeast by increasing transcription of Gcd1 which can increase entry to the pathway. This increased capacity for entry into the pentose phosphate pathway within the cell would allow for much greater NADPH production and offer protection against the reactive oxygen species from the oxidative stress (Mullarky and Cantley, 2015).

*Hsr1 regulates flocculation and cellular adhesion in response to oxidative stress through pfl2*

Flocculation, the process of cell aggregation and separation from the media, has been shown to be a protective phenotype for cells under stress (Smukalla et al., 2008). The flocculation allows the inner cells of the floc to be protected from the environmental stressors by the physical barrier of the external cells of the floc. In this way, flocculation can be seen as a response to environmental stressors to protect some cells in the population from the effects of the stress.

*pfl2*, a high-confidence target gene, was identified as having oxidative-stress dependent binding with binding at 30-minutes post treatment with $H_2O_2$, but not at 60-minutes post treatment or in the untreated time 0 control. *pfl2* is a protein encoding gene which codes for Pfl2, a cell surface glycoprotein involved in flocculation whose overexpression has been shown to produce the second highest degree of flocculation of all the pombe flocculins (Kwon et al., 2012). Since overexpression of *pfl2* causes an increase in flocculation we can assume that the transcription levels of *pfl2* will affect the degree of flocculation. Hsr1 binding to the *pfl2* promoter in response to oxidative stress regulates this transcription and therefore likely regulates the degree of flocculation of the cell. *pfl2* is also the direct transcriptional target of the flocculation transcription factors Mbx2 and Rfl1 (Kwon et al., 2012), meaning that Hsr1 likely functions as part of this network of flocculation transcription factors.

Gene ontology enrichment analysis for the high confidence Hsr1 targets also showed enrichment for flocculation, aggregation of unicellular organisms and cell aggregation at nearly 100 times that of the background frequency in the genome (table 22). This enrichment was caused by four of the Hsr1 targets having these associated biological processes. Aside from *pfl2*, Hsr1 also bound

to flocculation related genes SPAPB2C8.01, *pfl3* / SPBC947.04, SPBPJ4664.02, *pfl5* / SPBC1289.15 and *gsf2* / SPCC1742.01. This binding was not dependent on oxidative stress in the same way as *pfl2* binding, but it does point to a wider involvement of Hsr1 in regulating cellular flocculation. Three of the genes are likely directly involved with the *pfl2* pathway: three other named pombe flocculins, *pfl3* and *pfl5* and *gsf2*, also regulated by Mbx2 and Rfl1 (Kwon et al., 2012). From this, it can be inferred that Hsr1 is involved in flocculation regulation in both an oxidative stress dependent and independent way.

*Hsr1 regulates ribosomal subunit 18s in response to oxidative stress*

8 ribosomal RNA genes were found to be high confidence targets of Hsr1. Of these, binding at SPRRNA.43, SPRRNA.44, and SPRRNA.46 was found to be responsive to oxidative stress due to Hsr1 unbinding by 30 minutes of H2O2 treatment and rebinding by 60 minutes of $H_2O_2$ treatment (table 21). These genes make up the 18s ribosomal subunit (EMBL-EBI, 2021) meaning that Hsr1 regulates transcription of the 18s ribosomal subunit in response to oxidative stress. mRNA levels and protein levels are known to decrease during oxidative stress with mRNA returning to baseline ~1h after treatment but protein levels continuing to change due to a number of factors (Vogel et al., 2011). Reduced transcription upon oxidative stress has been attributed to ribosome stalling on tryptophan codons causing ribosome accumulation upstream (Rubio et al., 2021). With the oxidative stress time-dependant binding of Hsr1 to the genes which make up the 18s ribosomal subunit, we can suggest that Hsr1 is another mechanism by which translation is stalled during oxidative stress before returning to baseline. This mechanism appears to mirror the decrease in transcription which also returns to baseline within 60 minutes of treatment.

*Hsr1 functions as part of a wider network of ageing-associated transcription factors*

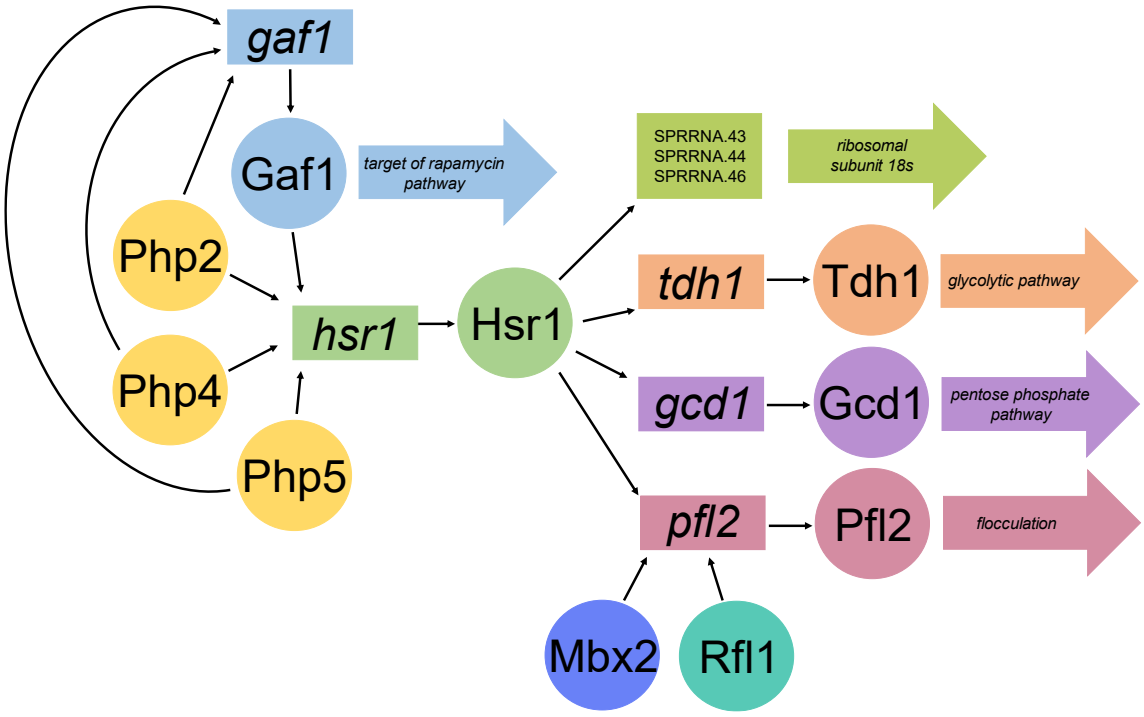



Figure 29: Hsr1 functions as part of a wider network.

Hsr1 is regulated by ageing-associated transcription factors including Gaf1, Php2, Php4 and Php5, and goes on to regulate ribosomal subunit 18s, the glycolytic pathway, the pentose phosphate pathway and flocculation.

Since *hsr1* is regulated itself by the ageing-associated transcription factors Gaf1, Php2, Php4 and Php5, it forms part of a wider network of the transcriptional control of ageing-associated pathways in the cell (figure 29). Hsr1, through regulation by these transcription factors, goes on itself to regulate components of the glycolytic and pentose phosphate pathways. Both of these pathways are strongly implicated in ageing-associated stress responses to nutrient availability and oxidative stress (Albers et al., 2007, Mullarky and Cantley, 2015). Along with the transcription factors Mbx2 and Rfl1, Hsr1 also regulates flocculation in fission yeast which can be a protective phenotype for cells undergoing stress leading to increased chances of survival and longevity (Di Gianvito et al., 2017). Hsr1 also regulates the genes which code for the

ribosomal subunit 18s and could be a mechanism by which translation is reduced upon oxidative stress.

This network of transcriptional control helps to provide mechanisms for the interplay between the pathways involved in ageing including the target of rapamycin pathway, the glycolytic pathway, and the pentose phosphate pathway. Building up a picture of the interplay between ageing related transcription factors can elucidate the ways in which ageing-associated cellular processes are connected and dependent on one another.

*Transcription factors involved in sulfur metabolism and negative regulation of meiosis in s. cerevisiae have similar binding motifs to Hsr1*

The suggested TGTGGT binding motif for Hsr1 (figure 21) has high similarity to the binding motifs of the *S. cerevisiae* transcription factors Met32 and Rme1 (figure 22) which can suggest conserved functionality. Met32 is responsible for activation of the sulfur metabolism pathway including sulfate assimilation and sulfonate metabolism (Carrillo et al., 2012). Rme1, like Hsr1, is a zinc finger transcription factor which has been shown to be a negative regulator of meiosis and a positive activator of G1 cyclin gene expression leading to entry initiation of the cell cycle in *s. cerevisiae* (Toone et al., 1995). Sulfur metabolism, meiosis and regulation of the cell cycle are all processes associating with ageing processes within the cell (Tyers et al., 1993, Gire and Dulic, 2015, Jeon et al., 2018, Boselli et al., 2009), suggesting that the transcription factors with similar binding sites to Hsr1 are also involved with ageing-associated cellular pathways.

*Transcription factors involved in cell proliferation, differentiation and cell lineage specification in humans have similar binding sites to Hsr1*

In humans, the RUNX family of transcription factors have a similar binding motif to Hsr1 (figure 23). RUNX transcription factors are highly conserved across metazoans and are involved in a range of cellular processes (Mevel et al., 2019). RUNX1 has been shown to be involved in cell differentiation in haematopoiesis, RUNX2 in skeletal development and RUNX3 in neurogenesis. Mutations in the RUNX transcription factor family have been shown to be involved in the proliferation and development of cancers, summarised in table 26.

Table 26: Summary of the cancers which have been shown to be contributed to by mutations in RUNX transcription factors.

| | |
|---|---|
| **RUNX1** | Epithelial tumours such as skin and oral cancers (Scheitz et al., 2012).<br>Tumorigenesis of hormone related organs including breast, ovarian, uterine, and prostate cancers (Riggio and Blyth, 2017). |
| **RUNX2** | Osteosarcoma development (Martin et al., 2011).<br>Breast and Prostate cancer bone metastasis (Chuang et al., 2017). |
| **RUNX3** | Solid-tissue tumorigenesis in the gastrointestinal system, pancreas, and lungs (Chuang et al., 2017, Lotem et al., 2017). |

Cancer is widely considered to be an ageing related disease since age is the biggest risk factor for the disease, with cell proliferation and differentiation a key component to its physiology (Berben et al., 2021). Therefore, the human transcription factors with similar binding motifs to Hsr1 are also involved in ageing-associated cellular processes.

*Future directions*

Future research into Hsr1 and its targets should focus on defining its function as a repressor or an activator for each of its targets. Notably, published microarray data for a *hsr1*Δ deletion strain (Chen et al., 2008) showed no overlap with respect to the differentially expressed genes in the mutant and the stress-dependant high-confidence targets identified by this ChIP-Seq experiment. This discrepancy may reflect differences in the particular oxidative-stress conditions which can lead to experimental variaton, so new transcriptional analysis would be necessary. Transcriptional analysis such as RNA-seq as well as phenotyping of double mutants could help to fully explain the impact of Hsr1 binding on its targets and therefore further define its role within cellular ageing and ageing-associated processes.

### 3.4.4 Genetic interactions of *hsr1*

In an SGA, negative genetic interactions usually highlight that the protein products are involved in parallel or compensatory pathways whereas positive interactions usually highlight that the protein products operate in the same linear pathway (Ryan et al., 2013). In this way, the SGA can be used to identify other components of Hsr1's linear pathway as well as components of pathways which act in a compensatory manner to it. The screen highlighted the genetic interactions between hsr1 and two of its known associated genes under the YES condition, along with other ageing-associated genes dependent on caffeine and rapamycin treatment.

*hsr1 has a negative genetic interaction with its known associates php5 and gcd1*

Under YES conditions, *hsr1* had negative genetic interactions with *php5* and *gcd1*. As discussed in 3.4.3, Php5 binds to *hsr1* and Hsr1 binds to *gcd1*. The

negative interaction of hsr1 with these two genes is indicative of the genes belonging to the same compensatory network as *hsr1*. These interactions are therefore consistent with the results from Php5 and Hsr1 ChIP-Seq experiments which report these bindings. This adds to the evidence of the connections between Hsr1 and Php5/Gcd1 as well as providing a positive control which helps to reassure that the SGA is producing reliable and valid findings.

*hsr1 has caffeine and rapamycin dependent negative interactions with ribosomal associated genes rpl26 and fyv7*

*hsr1* showed caffeine and rapamycin dependent negative interactions with the 60s ribosomal protein encoding gene *rpl26* (Leng et al., 2014) and the rRNA processing gene *fyv7* (Peng et al., 2003). The involvement of these genes suggests that Hsr1 is involved in compensatory ribosomal processes within the cell, adding to the ChIP-Seq result that Hsr1 regulates ribosomal subunit 18s in response to oxidative stress (section 3.4.3). This suggests that Hsr1 is also involved with ribosomal processes during TOR inhibition. *rpl26* also has a human ortholog RPL26, which is implicated in cancer as it regulates the tumour suppressor genes p53 and p73 (Gazda et al., 2012, Zhang et al., 2016). This adds to Hsr1's potential involvement in cancer processes through its similar binding site to the RUNX genes (section 3.4.3).

*hsr1 has a caffeine and rapamycin dependent negative interaction with the ageing related gene fkh1*

*hsr1* also showed a caffeine and rapamycin dependent negative interaction with *fkh1*, a conserved forkhead transcription factor required for stress response, cell cycle progression and longevity (Malo et al., 2016). A negative interaction suggests that Hsr1 and Fkh1 are acting in compensatory biological pathways, supporting Hsr1's involvement in stress- and ageing-associated pathways. *fkh1*

deletion strains show reduced chronological lifespan and increased stress sensitivity, likely reflection that Fkh1 is a target of rapamycin in fission yeast (Weisman et al., 2001, Malo et al., 2016). Since Fkh1 is a target of rapamycin, its deletion would confer a rapamycin resistant strain, with the subsequent deletion of *hsr1* leading to even greater resistance to rapamycin with the double mutant with *hsr1* showing a reduction in growth from the ade6 double mutant control by more than half. This is strongly suggestive that Hsr1 and Fkh1 operate in a compensatory manner during TOR inhibition. Operating in a compensatory TOR associated pathway could also explain the caffeine and rapamycin resistance of the *hsr1*Δ strain as seen in figure 17.

*hsr1 has a caffeine and rapamycin dependent positive interaction with the ageing related gene amk2*

*hsr1* showed a caffeine and rapamycin dependent positive interaction with *amk2*, the AMP-activated protein kinase (AMPK). AMPK is closely linked to ageing by controlling autophagy through TOR, being involved in cellular stress resistance and metabolic regulation (Salminen and Kaarniranta, 2012). A decline in AMPK activation in ageing is associated with decreased autophagy, increased oxidative stress, increased endoplasmic stress and increased apoptotic resistance, all key elements of ageing and ageing-associated disease (Salminen and Kaarniranta, 2012). The positive interaction with *hsr1* suggests that Hsr1 is involved in the same linear pathway as AMPK, this would cement Hsr1 as being part of an ageing-associated pathway and confirm its involvement with ageing and lifespan. Since AMPK interacts with the TOR pathway and the positive interaction was dependent on the condition of TOR inhibition it is likely that Hsr1 is involved in the AMPK/TOR cascade.

*Future directions*

Future studies could look to create SGAs for known associates of *hsr1* including *php5* which also showed a genetic interaction with *hsr1* in this work. These SGAs could be used to look for overlaps between the genetic interactions and create a clearer picture of the relationships between these pathways. To build on this further, phenotype tests of double mutants with interaction hits such as *fkh1* and *amk2* could be used to investigate the genetic interaction more thoroughly.

# Conclusions

The current theories for the mechanism of ageing range across several complex pathways and processes, such as mitochondrial stress, DNA damage, and free radicals. All these theories have supporting evidence which shows them to be, at least in part, an accurate description of the ageing process. However, the theories also conflict in ways that make a unified theory of ageing hard to define. This evidences that the mechanism which underpins the ageing process is likely a process which involves multiple pathways acting in a combinatory manner. As with any complex biological process, when we have only part of the story, defining a mechanism can seem like an impossible task. However, as research progresses, we are collectively able to fill in the blanks and develop a clear picture of each new step in the pathway. In the current landscape of ageing research, we are likely at this juncture from which we can only see parts of a mechanism larger than we can currently conceive.

There are two ways to approach beginning to fill in the gaps of this problem. Firstly, we can look at the mechanism from the bottom, up, and continue to define interactions of individual pathways one at a time. This is the traditional approach of molecular biology, which has brought the field much of the current theories of ageing. However, engaging with this bottom-up approach alone

often results in a granular understanding of complex mechanisms, resulting in situations such as with ageing where we understand parts in great detail but are missing connections. The second approach is to instead study the mechanism from the top, down. Here, we would engage approaches like machine learning and high throughput screens, which don't provide the same mechanistic detail but instead bring us overarching ideas and provide evidence for connections between pathways.

In scientific research, we are always hampered by the limitations of our techniques and our own preconceived understanding. We are only able to look at a small part of big answers at any one time and then try to build the connections between the evidence to create an understandable narrative. Taking on questions as integral to our own lives as 'what is ageing?' is an enormous responsibility, and we must make every effort to see the problem from all angles. To take the fullest advantage of the technologies available to us in the ageing field, we can design projects which tackle research questions both from the bottom up and the top down, giving us the best opportunity to catch a glimpse of the complete picture. With this research, I took the question of 'how are ageing-associated pathways connected?' and approached it from both a bottom-up and top-down direction. Combining machine learning built on high throughput lifespan and phenotype screens and traditional molecular biology techniques for defining transcriptional control of ageing pathways, I created a multi-pronged approach to the research. By doing so, I was able to begin to unpick some of the complex and combinatory mechanisms which connect the cellular ageing-associated pathways.

Making use of a phenotypically rich collection of wild type yeast strains from around the world, I developed a machine learning model to predict the complex

phenotype of lifespan from simple phenotypes to within 3 days. When we consider that we usually describe fission yeast lifespans in the relative terms of being ‘short’ or ‘long’, a quantitative error of 3 days within a range of 0-28 days is negligible, and the model can be described as successfully predicting lifespan from simple phenotypes. This is an important step in itself, as it evidences that these phenotypes contain all the information necessary for determining lifespan. From this, we can infer that several stress response pathways are involved in the ageing process, with the combinatory nature of this involvement demonstrated by the sophisticated form of machine learning model needed for the prediction.

Further to this question, I then leveraged the model to ask which of these phenotypes, and consequently which ageing-associated pathways, are most involved in ageing. To do this, I created a custom feature selection script, which was able to select the most predictive phenotypes for lifespan by building the model in a stepwise manner. This feature selection provided more evidence that a combination of data from different stress pathways was necessary for accurate lifespan prediction, highlighting the combinatory nature of these pathways’ contribution to cellular ageing and lifespan. The feature selection was also able to identify mitochondrial stress, nutrient availability, DNA damage and amino acid supplementation as the most predictive phenotypes of ageing, and, therefore, implicate them as the pathways most closely linked with cellular ageing.

This work then went on to investigate some of the finer mechanisms by which these pathways may be connected in the form of the ageing-associated transcription factors. Characterisation of the *hsr1*Δ deletion strain showed it to have a resistant phenotype to TOR inhibition by caffeine and rapamycin

treatment, and subsequently, Hsr1 was further investigated for its ageing-related connections. Hsr1 functions as part of a wider network of ageing-associated transcription factors, including Gaf1 and Phps, and target identification by ChIP-Seq revealed that it regulates ribosomal subunit 18s, flocculation, the glycolytic pathway, and the pentose phosphate pathway in response to oxidative stress. Genome-wide genetic interaction analysis of hsr1 then uncovered its interaction with multiple ageing-associated genes, including components of the TOR pathway.

In this way, this research was able to highlight the combinatory, complex nature of the ageing-associated pathways and their contribution to cellular ageing, beginning to provide some clues for how they are connected. The research showed from both perspectives that DNA damage and repair are integral to the ageing process by highlighting DNA damage from the top-down machine learning approach and the pentose phosphate pathway, which is involved in DNA synthesis, in the bottom-up ChIP-Seq approach. ChIP-Seq analysis of Hsr1 targets also implicated the connection of flocculation to ageing-associated pathways, connecting with the machine learning model, which identified amino acid and nutrient availability as strongly ageing-associated. While these associations are preliminary, the work has highlighted some areas of future interest which would allow for further research to build on these results. By continuing to approach ageing research questions from multiple angles, it is my firm belief that the field will go on to develop the nuanced and unified mechanistic model of cellular ageing we have been searching for.

# References

ALBERS, E., LARSSON, C., ANDLID, T., WALSH, M. C. & GUSTAFSSON, L. 2007. Effect of nutrient starvation on the cellular composition and metabolic capacity of Saccharomyces cerevisiae. *Appl Environ Microbiol,* 73**,** 4839-48.

ANDREWS, S. 2010. *FastQC: a quality control tool for high throughput sequence data* [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed].

BAILEY, T. L., JOHNSON, J., GRANT, C. E. & NOBLE, W. S. 2015. The MEME Suite. *Nucleic Acids Research,* 43**,** W39-W49.

BARYSHNIKOVA, A., COSTANZO, M., DIXON, S., VIZEACOUMAR, F. J., MYERS, C. L., ANDREWS, B. & BOONE, C. 2010. Chapter 7 - Synthetic Genetic Array (SGA) Analysis in Saccharomyces cerevisiae and Schizosaccharomyces pombe. *Methods in Enzymology.* Academic Press.

BERBEN, L., FLORIS, G., WILDIERS, H. & HATSE, S. 2021. Cancer and Aging: Two Tightly Interconnected Biological Processes. *Cancers (Basel),* 13.

BITTON, D. A., SCHUBERT, F., DEY, S., OKONIEWSKI, M., SMITH, G. C., KHADAYATE, S., PANCALDI, V., WOOD, V. & BÄHLER, J. 2015. AnGeLi: A Tool for the Analysis of Gene Lists from Fission Yeast. *Front Genet,* 6**,** 330.

BLANKENBERG, D., GORDON, A., VON KUSTER, G., CORAOR, N., TAYLOR, J., NEKRUTENKO, A. & TEAM, T. G. 2010. Manipulation of FASTQ data with Galaxy. *Bioinformatics,* 26**,** 1783-1785.

BOSELLI, M., ROCK, J., UNAL, E., LEVINE, S. S. & AMON, A. 2009. Effects of age on meiosis in budding yeast. *Dev Cell,* 16**,** 844-55.

BROWNLEE, J. 2019. *How to Choose a Feature Selection Method For Machine Learning* [Online]. Available: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/ [Accessed].

BROWNLEE, J. 2020a. *A Gentle Introduction to the Rectified Linear Unit (ReLU)* [Online]. Available: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/ [Accessed].

BROWNLEE, J. 2020b. *Recursive Feature Elimination (RFE) for Feature Selection in Python* [Online]. Available: https://machinelearningmastery.com/rfe-feature-selection-in-python/ [Accessed].

BROWNLEE, J. 2020c. *Understand the Impact of Learning Rate on Neural Network Performance* [Online]. Available: https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/ [Accessed].

BROWNLEE, J. 2021. How to Develop LASSO Regression Models in Python. Available: https://machinelearningmastery.com/lasso-regression-with-python/ [Accessed 07/02/2023].

BRYS, K., VANFLETEREN, J. R. & BRAECKMAN, B. P. 2007. Testing the rate-of-living/oxidative damage theory of aging in the nematode model Caenorhabditis elegans. *Exp Gerontol,* 42**,** 845-51.

BURHANS, W. C. & WEINBERGER, M. 2012. DNA damage and DNA replication stress in yeast models of aging. *Subcell Biochem,* 57**,** 187-206.

BURTSCHER, J., SOLTANY, A., VISAVADIYA, N. P., BURTSCHER, M., MILLET, G. P., KHORAMIPOUR, K. & KHAMOUI, A. V. 2023. Mitochondrial stress and mitokines in aging. *Aging Cell,* 22**,** e13770.

CARRILLO, E., BEN-ARI, G., WILDENHAIN, J., TYERS, M., GRAMMENTZ, D. & LEE, T. A. 2012. Characterizing the roles of Met31 and Met32 in coordinating Met4-activated transcription in the absence of Met30. *Mol Biol Cell,* 23**,** 1928-42.

CASSON, R. J. & FARMER, L. D. M. 2014. Understanding and checking the assumptions of linear regression: a primer for medical researchers. *Clinical & Experimental Ophthalmology,* 42**,** 590-596.

CHANDLER-BROWN, D., CHOI, H., PARK, S., OCAMPO, B. R., CHEN, S., LE, A., SUTPHIN, G. L., SHAMIEH, L. S., SMITH, E. D. & KAEBERLEIN, M. 2015. Sorbitol treatment extends lifespan and induces the osmotic stress response in Caenorhabditis elegans. *Frontiers in Genetics,* 6.

CHEN, B. R. & RUNGE, K. W. 2009. A new Schizosaccharomyces pombe chronological lifespan assay reveals that caloric restriction promotes efficient cell cycle exit and extends longevity. *Exp Gerontol,* 44**,** 493-502.

CHEN, D., WILKINSON, C. R. M., WATT, S., PENKETT, C. J., TOONE, W. M., JONES, N. & BÄHLER, J. 2008. Multiple Pathways Differentially Regulate Global Oxidative Stress Responses in Fission Yeast. *Molecular Biology of the Cell,* 19**,** 308-317.

CHEN, K., SHEN, W., ZHANG, Z., XIONG, F., OUYANG, Q. & LUO, C. 2020. Age-dependent decline in stress response capacity revealed by proteins dynamics analysis. *Sci Rep,* 10**,** 15211.

CHENG, L., ZHIYONG, Y., JUN, Y., ZANXIAN, X. & SHIZHOU, A. 2000 Regulation of the Yeast Transcriptional Factor PHO2 Activity by Phosphorylation. *The Journal of Biological Chemistry***,** 31972-31978.

CHOI, R. Y., COYNER, A. S., KALPATHY-CRAMER, J., CHIANG, M. F. & CAMPBELL, J. P. 2020. Introduction to Machine Learning, Neural Networks, and Deep Learning. *Transl Vis Sci Technol,* 9**,** 14.

CHUANG, L. S. H., ITO, K. & ITO, Y. 2017. Roles of RUNX in Solid Tumors. *In:* GRONER, Y., ITO, Y., LIU, P., NEIL, J. C., SPECK, N. A. & VAN WIJNEN, A. (eds.) *RUNX Proteins in Development and Cancer.* Singapore: Springer Singapore.

COMMUNITY, T. G. 2022. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research,* 50**,** W345-W351.

CORKINS, M. E., WILSON, S., COCURON, J. C., ALONSO, A. P. & BIRD, A. J. 2017. The gluconate shunt is an alternative route for directing glucose into the pentose phosphate pathway in fission yeast. *J Biol Chem,* 292**,** 13823-13832.

CORNELIUS, E. 1972. Increased incidence of lymphomas in thymectomized mice--evidence for an immunological theory of aging. *Experientia,* 28**,** 459.

DAI, D. F., CHIAO, Y. A., MARCINEK, D. J., SZETO, H. H. & RABINOVITCH, P. S. 2014. Mitochondrial oxidative stress in aging and healthspan. *Longev Healthspan,* 3**,** 6.

DAVIDOVIC, M., SEVO, G., SVORCAN, P., MILOSEVIC, D. P., DESPOTOVIC, N. & ERCEG, P. 2010. Old age as a privilege of the "selfish ones". *Aging Dis,* 1**,** 139-46.

DAVINELLI, S., WILLCOX, D. C. & SCAPAGNINI, G. 2012. Extending healthy ageing: nutrient sensitive pathway and centenarian population. *Immunity & Ageing,* 9**,** 9.

DI GIANVITO, P., TESNIÈRE, C., SUZZI, G., BLONDIN, B. & TOFALO, R. 2017. FLO5 gene controls flocculation phenotype and adhesive properties in a Saccharomyces cerevisiae sparkling wine strain. *Scientific Reports,* 7**,** 10786.

DOSHI, S. 2019. *Various Optimization Algorithms For Training Neural Network* [Online]. Available: https://towardsdatascience.com/optimizers-for-training-neural-network-59450d71caf6 [Accessed].

DUES, D. J., ANDREWS, E. K., SCHAAR, C. E., BERGSMA, A. L., SENCHUK, M. M. & VAN RAAMSDONK, J. M. 2016. Aging causes decreased resistance to multiple stresses and a failure to activate specific stress response pathways. *Aging (Albany NY),* 8**,** 777-95.

DUNCAN, C. D. S., RODRÍGUEZ-LÓPEZ, M., RUIS, P., BÄHLER, J. & MATA, J. 2018. General amino acid control in fission yeast is regulated by a nonconserved transcription factor, with functions analogous to Gcn4/Atf4. *Proceedings of the National Academy of Sciences of the United States of America,* 115**,** E1829-E1838.

EMBL-EBI. 2021. *Schizosaccharomyces pombe (fission yeast) 18S ribosomal RNA* [Online]. Available: https://rnacentral.org/rna/URS000055C986/4896?tab=2d [Accessed].

FENG, J., LIU, T., QIN, B., ZHANG, Y. & LIU, X. S. 2012. Identifying ChIP-seq enrichment using MACS. *Nature Protocols,* 7**,** 1728-1740.

FIGUEROA BARRAZA, J., LÓPEZ DROGUETT, E. & MARTINS, M. R. 2021. Towards Interpretable Deep Learning: A Feature Selection Framework for Prognostics and Health Management Using Deep Neural Networks. *Sensors (Basel),* 21.

FLANAGAN, E. W., MOST, J., MEY, J. T. & REDMAN, L. M. 2020. Calorie Restriction and Aging in Humans. *Annu Rev Nutr,* 40**,** 105-133.

FONTANA, L., PARTRIDGE, L. & LONGO, V. D. 2010. Extending healthy life span--from yeast to humans. *Science,* 328**,** 321-6.

GALKIN, F., MAMOSHINA, P., KOCHETOV, K., SIDORENKO, D. & ZHAVORONKOV, A. 2021. DeepMAge: A Methylation Aging Clock Developed with Deep Learning. *Aging Dis,* 12**,** 1252-1262.

GAZDA, H. T., PRETI, M., SHEEN, M. R., O'DONOHUE, M. F., VLACHOS, A., DAVIES, S. M., KATTAMIS, A., DOHERTY, L., LANDOWSKI, M., BUROS, C., GHAZVINIAN, R., SIEFF, C. A., NEWBURGER, P. E., NIEWIADOMSKA, E., MATYSIAK, M., GLADER, B., ATSIDAFTOS, E., LIPTON, J. M., GLEIZES, P. E. & BEGGS, A. H. 2012. Frameshift mutation in p53 regulator RPL26 is associated with multiple physical abnormalities and a specific pre-ribosomal RNA processing defect in diamond-blackfan anemia. *Hum Mutat,* 33**,** 1037-44.

GIRE, V. & DULIC, V. 2015. Senescence from G2 arrest, revisited. *Cell Cycle,* 14**,** 297-304.

GLADYSHEV, V. N. 2014. The free radical theory of aging is dead. Long live the damage theory! *Antioxid Redox Signal,* 20**,** 727-31.

GOLDAR, M. M., JEONG, H. T., TANAKA, K., MATSUDA, H. & KAWAMUKAI, M. 2005. Moc3, a novel Zn finger type protein involved in sexual development, ascus formation, and stress response of Schizosaccharomyces pombe. *Current Genetics,* 48**,** 345.

GONZÁLEZ, A. & HALL, M. N. 2017. Nutrient sensing and TOR signaling in yeast and mammals. *Embo j,* 36**,** 397-408.

GONZALEZ, S. & RALLIS, C. 2017. The TOR Signaling Pathway in Spatial and Temporal Control of Cell Size and Growth. *Front Cell Dev Biol,* 5**,** 61.

HARRIS, M. A., RUTHERFORD, K. M., HAYLES, J., LOCK, A., BÄHLER, J., OLIVER, S. G., MATA, J. & WOOD, V. 2021. Fission stories: using PomBase to understand Schizosaccharomyces pombe biology. *Genetics,* 220.

HAYFLICK, L. 2007. Biological aging is no longer an unsolved problem. *Ann N Y Acad Sci,* 1100**,** 1-13.

HILL, S. & VAN REMMEN, H. 2014. Mitochondrial stress signaling in longevity: A new role for mitochondrial function in aging. *Redox Biology,* 2**,** 936-944.

HORVATH, S. 2013. DNA methylation age of human tissues and cell types. *Genome Biol,* 14**,** R115.

HUGHES, K. A. & REYNOLDS, R. M. 2005. Evolutionary and mechanistic theories of aging. *Annu Rev Entomol,* 50**,** 421-45.

IBM. 2023. *What is linear regression?* [Online]. Available: https://www.ibm.com/uk-en/topics/linear-regression [Accessed 07/02/2023 2023].

JAUL, E. & BARRON, J. 2017. Age-Related Diseases and Clinical and Public Health Implications for the 85 Years Old and Over Population. *Front Public Health,* 5**,** 335.

JEFFARES, D. C., JOLLY, C., HOTI, M., SPEED, D., SHAW, L., RALLIS, C., BALLOUX, F., DESSIMOZ, C., BÄHLER, J. & SEDLAZECK, F. J. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications,* 8**,** 14061.

JEFFARES, D. C., RALLIS, C., RIEUX, A., SPEED, D., PŘEVOROVSKÝ, M., MOURIER, T., MARSELLACH, F. X., IQBAL, Z., LAU, W., CHENG, T. M., PRACANA, R., MÜLLEDER, M., LAWSON, J. L., CHESSEL, A., BALA, S., HELLENTHAL, G., O'FALLON, B., KEANE, T., SIMPSON, J. T., BISCHOF, L., TOMICZEK, B., BITTON, D. A., SIDERI, T., CODLIN, S., HELLBERG, J. E., VAN TRIGT, L., JEFFERY, L., LI, J. J., ATKINSON, S., THODBERG, M., FEBRER, M., MCLAY, K., DROU, N., BROWN, W., HAYLES, J., CARAZO SALAS, R. E., RALSER, M., MANIATIS, N., BALDING, D. J., BALLOUX, F., DURBIN, R. & BÄHLER, J. 2015. The genomic and phenotypic diversity of Schizosaccharomyces pombe. *Nat Genet,* 47**,** 235-41.

JEON, J. S., OH, J. J., KWAK, H. C., YUN, H. Y., KIM, H. C., KIM, Y. M., OH, S. J. & KIM, S. K. 2018. Age-Related Changes in Sulfur Amino Acid Metabolism in Male C57BL/6 Mice. *Biomol Ther (Seoul),* 26**,** 167-174.

JIN, K. 2010. Modern Biological Theories of Aging. *Aging Dis,* 1**,** 72-74.

JIN, Y., LIANG, Z. & LOU, H. 2020. The Emerging Roles of Fox Family Transcription Factors in Chromosome Replication, Organization, and Genome Stability. *Cells,* 9.

KALITA, A., HESLES, E. E., POWER, L. N., WANG, D., SINGH, P. K. & SMITH, J. S. 2021. Isonicotinamide extends yeast chronological lifespan through a mechanism that diminishes nucleotides. *bioRxiv***,** 2021.07.11.451986.

KAMRAD, S., RODRÍGUEZ-LÓPEZ, M., COTOBAL, C., CORREIA-MELO, C., RALSER, M. & BÄHLER, J. 2020. Pyphe, a python toolbox for assessing microbial growth and cell viability in high-throughput colony screens. *Elife,* 9.

KEVIN BLIGHE, S. R. A. M. L. 2021. *EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.* [Online]. R package version 1.12.0. Available: https://github.com/kevinblighe/EnhancedVolcano [Accessed].

KIM, J.-Y., KWON, E.-S. & ROE, J.-H. 2012. A homeobox protein Phx1 regulates long-term survival and meiotic sporulation in Schizosaccharomyces pombe. *BMC Microbiology,* 12**,** 86-86.

KWOLEK-MIREK, M. & ZADRAG-TECZA, R. 2014. Comparison of methods used for assessing the viability and vitality of yeast cells. *FEMS Yeast Research,* 14**,** 1068-1079.

KWON, E. J., LADEROUTE, A., CHATFIELD-REED, K., VACHON, L., KARAGIANNIS, J. & CHUA, G. 2012. Deciphering the transcriptional-regulatory network of flocculation in Schizosaccharomyces pombe. *PLoS Genet,* 8**,** e1003104.

LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods,* 9**,** 357-359.

LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology,* 10**,** R25.

LAPLANTE, M. & SABATINI, D. M. 2009. mTOR signaling at a glance. *J Cell Sci,* 122**,** 3589-94.

LEE, J. W. & ONG, E. B. B. 2021. Genomic Instability and Cellular Senescence: Lessons From the Budding Yeast. *Frontiers in Cell and Developmental Biology,* 8.

LEE, J. W., ONG, T. G., SAMIAN, M. R., TEH, A.-H., WATANABE, N., OSADA, H. & ONG, E. B. B. 2021. Screening of selected ageing-related proteins that extend chronological life span in yeast Saccharomyces cerevisiae. *Scientific Reports,* 11**,** 24148.

LEITÃO, C., MIGNANO, A., ESTRELA, M., FARDILHA, M., FIGUEIRAS, A., ROQUE, F. & HERDEIRO, M. T. 2022. The Effect of Nutrition on Aging-A Systematic Review Focusing on Aging-Related Biomarkers. *Nutrients,* 14.

LENG, X. M., DIAO, L. T., LI, B., BI, Y. Z., CHEN, C. J., ZHOU, H. & QU, L. H. 2014. The ribosomal protein rpl26 promoter is required for its 3' sense terminus ncRNA transcription in Schizosaccharomyces pombe, implicating a new transcriptional mechanism for ncRNAs. *Biochem Biophys Res Commun,* 444**,** 86-91.

LEONOV, A., FELDMAN, R., PIANO, A., ARLIA-CIOMMO, A., LUTCHMAN, V., AHMADI, M., ELSASER, S., FAKIM, H., HESHMATI-MOGHADDAM, M., HUSSAIN, A., ORFALI, S., RAJEN, H., ROOFIGARI-ESFAHANI, N., ROSANELLI, L. & TITORENKO, V. I. 2017. Caloric restriction extends yeast chronological lifespan via a mechanism linking cellular aging to cell cycle regulation, maintenance of a quiescent state, entry into a non-quiescent state and survival in the non-quiescent state. *Oncotarget,* 8**,** 69328-69350.

LEONTIEVA, O. V., DEMIDENKO, Z. N. & BLAGOSKLONNY, M. V. 2015. Dual mTORC1/C2 inhibitors suppress cellular geroconversion (a senescence program). *Oncotarget,* 6**,** 23238-48.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & SUBGROUP, G. P. D. P. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-2079.

LIE, S., BANKS, P., LAWLESS, C., LYDALL, D. & PETERSEN, J. 2018. The contribution of non-essential Schizosaccharomyces pombe genes to fitness in response to altered nutrient supply and target of rapamycin activity. *Open Biology,* 8**,** 180015.

LIMA, T., LI, T. Y., MOTTIS, A. & AUWERX, J. 2022. Pleiotropic effects of mitochondria in aging. *Nature Aging,* 2**,** 199-213.

LIN, S. J. & AUSTRIACO, N. 2014. Aging and cell death in the other yeasts, Schizosaccharomyces pombe and Candida albicans. *FEMS Yeast Res,* 14**,** 119-35.

LITHGOW, G. J., DRISCOLL, M. & PHILLIPS, P. 2017. A long journey to reproducible results. *Nature,* 548**,** 387-388.

LOTEM, J., LEVANON, D., NEGREANU, V., BAUER, O., HANTISTEANU, S., DICKEN, J. & GRONER, Y. 2017. Runx3 in Immunity, Inflammation and Cancer. *In:* GRONER, Y., ITO, Y., LIU, P., NEIL, J. C., SPECK, N. A. & VAN WIJNEN, A. (eds.) *RUNX Proteins in Development and Cancer.* Singapore: Springer Singapore.

LUÍZA DA COSTA, N., DIAS DE LIMA, M. & BARBOSA, R. 2021. Evaluation of feature selection methods based on artificial neural network weights. *Expert Systems with Applications,* 168**,** 114312.

LUJAN, C., TYLER, E. J., ECKER, S., WEBSTER, A. P., STEAD, E. R., MARTINEZ MIGUEL, V. E., MILLIGAN, D., GARBE, J. C., STAMPFER, M. R., BECK, S., LOWE, R., BISHOP, C. L. &

BJEDOV, I. 2020. A CellAgeClock for expedited discovery of anti-ageing compounds. *bioRxiv*, 803676.

MALO, M. E., POSTNIKOFF, S. D., ARNASON, T. G. & HARKNESS, T. A. 2016. Mitotic degradation of yeast Fkh1 by the Anaphase Promoting Complex is required for normal longevity, genomic stability and stress resistance. *Aging (Albany NY),* 8**,** 810-30.

MANGAN, D. 2021. Iron: an underrated factor in aging. *Aging (Albany NY),* 13**,** 23407-23415.

MARTIN, J. W., ZIELENSKA, M., STEIN, G. S., VAN WIJNEN, A. J. & SQUIRE, J. A. 2011. The Role of RUNX2 in Osteosarcoma Oncogenesis. *Sarcoma,* 2011**,** 282745.

MATA, J., WILBREY, A. & BÄHLER, J. 2007. Transcriptional regulatory network for sexual differentiation in fission yeast. *Genome Biology,* 8**,** R217-R217.

MATSUMOTO, Y., PIRAINO, S. & MIGLIETTA, M. P. 2019. Transcriptome Characterization of Reverse Development in Turritopsis dohrnii (Hydrozoa, Cnidaria). *G3 (Bethesda),* 9**,** 4127-4138.

MATSUO, T., OTSUBO, Y., URANO, J., TAMANOI, F. & YAMAMOTO, M. 2007. Loss of the TOR kinase Tor2 mimics nitrogen starvation and activates the sexual development pathway in fission yeast. *Mol Cell Biol,* 27**,** 3154-64.

MERCIER, A., PELLETIER, B. & LABBÉ, S. 2006. A transcription factor cascade involving Fep1 and the CCAAT-binding factor Php4 regulates gene expression in response to iron deficiency in the fission yeast Schizosaccharomyces pombe. *Eukaryot Cell,* 5**,** 1866-81.

MEVEL, R., DRAPER, J. E., LIE-A-LING, M., KOUSKOFF, V. & LACAUD, G. 2019. RUNX transcription factors: orchestrators of development. *Development,* 146**,** dev148296.

MIRISOLA, M. G., TAORMINA, G., FABRIZIO, P., WEI, M., HU, J. & LONGO, V. D. 2014. Serine- and threonine/valine-dependent activation of PDK and Tor orthologs converge on Sch9 to promote aging. *PLoS Genet,* 10**,** e1004113.

MIRZAEI, H., SUAREZ, J. A. & LONGO, V. D. 2014. Protein and amino acid restriction, aging and disease: from yeast to humans. *Trends Endocrinol Metab,* 25**,** 558-66.

MOLON, M., WOZNICKA, O. & ZEBROWSKI, J. 2018. Cell wall biosynthesis impairment affects the budding lifespan of the Saccharomyces cerevisiae yeast. *Biogerontology,* 19**,** 67-79.

MORIGASAKI, S., SHIMADA, K., IKNER, A., YANAGIDA, M. & SHIOZAKI, K. 2008. Glycolytic enzyme GAPDH promotes peroxide stress signaling through multistep phosphorelay to a MAPK cascade. *Mol Cell,* 30**,** 108-13.

MORIGASAKI, S. & SHIOZAKI, K. 2010. Chapter 15 - Two-Component Signaling to the Stress MAP Kinase Cascade in Fission Yeast. *Methods in Enzymology.* Academic Press.

MORIGASAKI, S. & SHIOZAKI, K. 2013. Phosphorelay-dependent and -independent regulation of MAPKKK by the Mcs4 response regulator in fission yeast. *Commun Integr Biol,* 6**,** e25020.

MULLARKY, E. & CANTLEY, L. 2015. Diverting Glycolysis to Combat Oxidative Stress. *In:* NAKAO K, M. N., UEMOTO S (ed.) *Innovative Medicine: Basic Research and Development.*

NG, P. & MAECHLER, M. 2007. A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling,* 7**,** 315-328.

NISHIMURA, A., YOSHIKAWA, Y., ICHIKAWA, K., TAKEMOTO, T., TANAHASHI, R. & TAKAGI, H. 2021. Longevity Regulation by Proline Oxidation in Yeast. *Microorganisms,* 9.

OHTSUKA, H., AZUMA, K., KUBOTA, S., MURAKAMI, H., GIGA-HAMA, Y., TOHDA, H. & AIBA, H. 2011. Chronological lifespan extension by Ecl1 family proteins depends on Prr1 response regulator in fission yeast. *Genes to Cells,* 17**,** 39-52.

OLSHANSKY, S. J. 2018. From Lifespan to Healthspan. *Jama,* 320**,** 1323-1324.

PAL, S., POSTNIKOFF, S. D., CHAVEZ, M. & TYLER, J. K. 2018. Impaired cohesion and homologous recombination during replicative aging in budding yeast. *Sci Adv,* 4**,** eaaq0236.

PARK, D. C. & YEO, S. G. 2013. Aging. *Korean J Audiol,* 17**,** 39-44.

PAUL, S. K., OOWATARI, Y. & KAWAMUKAI, M. 2009. A large complex mediated by Moc1, Moc2 and Cpc2 regulates sexual differentiation in fission yeast. *The FEBS Journal,* 276**,** 5076-5093.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. & DUCHESNAY, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* 12**,** 2825-2830.

PENG, W.-T., ROBINSON, M. D., MNAIMNEH, S., KROGAN, N. J., CAGNEY, G., MORRIS, Q., DAVIERWALA, A. P., GRIGULL, J., YANG, X., ZHANG, W., MITSAKAKIS, N., RYAN, O. W., DATTA, N., JOJIC, V., PAL, C., CANADIEN, V., RICHARDS, D., BEATTIE, B., WU, L. F., ALTSCHULER, S. J., ROWEIS, S., FREY, B. J., EMILI, A., GREENBLATT, J. F. & HUGHES, T. R. 2003. A Panoramic View of Yeast Noncoding RNA Processing. *Cell,* 113**,** 919-933.

PIFFERI, F., TERRIEN, J., MARCHAL, J., DAL-PAN, A., DJELTI, F., HARDY, I., CHAHORY, S., CORDONNIER, N., DESQUILBET, L., HURION, M., ZAHARIEV, A., CHERY, I., ZIZZARI, P., PERRET, M., EPELBAUM, J., BLANC, S., PICQ, J.-L., DHENAIN, M. & AUJARD, F. 2018. Caloric restriction increases lifespan but affects brain integrity in grey mouse lemur primates. *Communications Biology,* 1**,** 30.

POÜS, C. & CODOGNO, P. 2011. Lysosome positioning coordinates mTORC1 activity and autophagy. *Nat Cell Biol,* 13**,** 342-4.

RALLIS, C., CODLIN, S. & BÄHLER, J. 2013. TORC1 signaling inhibition by rapamycin and caffeine affect lifespan, global gene expression, and cell proliferation of fission yeast. *Aging Cell,* 12**,** 563-73.

RALLIS, C., LÓPEZ-MAURY, L., GEORGESCU, T., PANCALDI, V. & BÄHLER, J. 2014. Systematic screen for mutants resistant to TORC1 inhibition in fission yeast reveals genes involved in cellular ageing and growth. *Biology Open,* 3**,** 161-171.

RALLIS, C., MÜLLEDER, M., SMITH, G., AU, Y. Z., RALSER, M. & BÄHLER, J. 2021. Amino Acids Whose Intracellular Levels Change Most During Aging Alter Chronological Life Span of Fission Yeast. *The Journals of Gerontology: Series A,* 76**,** 205-210.

RAMÍREZ, F., RYAN, D. P., GRÜNING, B., BHARDWAJ, V., KILPERT, F., RICHTER, A. S., HEYNE, S., DÜNDAR, F. & MANKE, T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research,* 44**,** W160-W165.

RIGGIO, A. I. & BLYTH, K. 2017. The enigmatic role of RUNX1 in female-related cancers – current knowledge & future perspectives. *The FEBS Journal,* 284**,** 2345-2362.

RODRÍGUEZ-LÓPEZ, M., GONZALEZ, S., HILLSON, O., TUNNACLIFFE, E., CODLIN, S., TALLADA, V. A., BÄHLER, J. & RALLIS, C. 2020. The GATA Transcription Factor Gaf1 Represses tRNAs, Inhibits Growth, and Extends Chronological Lifespan Downstream of Fission Yeast TORC1. *Cell Reports,* 30**,** 3240-3249.e4.

ROLLO, C. D. 2010. Aging and the Mammalian regulatory triumvirate. *Aging Dis,* 1**,** 105-38.

ROMILA, C. A., TOWNSEND, S., MALECKI, M., KAMRAD, S., RODRÍGUEZ-LÓPEZ, M., HILLSON, O., COTOBAL, C., RALSER, M. & BÄHLER, J. 2021. Barcode sequencing and a high-throughput assay for chronological lifespan uncover ageing-associated genes in fission yeast. *Microb Cell,* 8**,** 146-160.

ROUX, A. E., QUISSAC, A., CHARTRAND, P., FERBEYRE, G. & ROKEACH, L. A. 2006. Regulation of chronological aging in Schizosaccharomyces pombe by the protein kinases Pka1 and Sck2. *Aging Cell,* 5**,** 345-357.

RUBIO, A., GHOSH, S., MÜLLEDER, M., RALSER, M. & MATA, J. 2021. Ribosome profiling reveals ribosome stalling on tryptophan codons and ribosome queuing upon oxidative stress in fission yeast. *Nucleic Acids Res,* 49**,** 383-399.

RUNGE, K. W. & ZHANG, H. 2018. Chapter 30 - The Budding and Fission Yeast Model Systems for Aging Biology: Rapid Advancement With New Technologies. *In:* RAM, J. L. & CONN, P. M. (eds.) *Conn's Handbook of Models for Human Aging (Second Edition).* Academic Press.

SALMINEN, A. & KAARNIRANTA, K. 2012. AMP-activated protein kinase (AMPK) controls the aging process via an integrated signaling network. *Ageing Research Reviews,* 11**,** 230-241.

SANTOS, J., LEITÃO-CORREIA, F., SOUSA, M. J. & LEÃO, C. 2016. Nitrogen and carbon source balance determines longevity, independently of fermentative or respiratory metabolism in the yeast Saccharomyces cerevisiae. *Oncotarget,* 7**,** 23033-42.

SCHEITZ, C. J. F., LEE, T. S., MCDERMITT, D. J. & TUMBAR, T. 2012. Defining a tissue stem cell-driven Runx1/Stat3 signalling axis in epithelial cancer. *The EMBO Journal,* 31**,** 4124-4139.

SCHMIDHUBER, J. 2015. Deep learning in neural networks: an overview. *Neural Netw,* 61**,** 85-117.

SILVA VANESSA, K. A., BHATTACHARYA, S., OLIVEIRA NATALIA, K., SAVITT ANNE, G., ZAMITH-MIRANDA, D., NOSANCHUK JOSHUA, D. & FRIES BETTINA, C. 2022. Replicative Aging Remodels the Cell Wall and Is Associated with Increased Intracellular Trafficking in Human Pathogenic Yeasts. *mBio,* 13**,** e00190-22.

SMUKALLA, S., CALDARA, M., POCHET, N., BEAUVAIS, A., GUADAGNINI, S., YAN, C., VINCES, M. D., JANSEN, A., PREVOST, M. C., LATGÉ, J. P., FINK, G. R., FOSTER, K. R. & VERSTREPEN, K. J. 2008. FLO1 is a variable green beard gene that drives biofilm-like cooperation in budding yeast. *Cell,* 135**,** 726-37.

SOARES, E. V. 2011. Flocculation in Saccharomyces cerevisiae: a review. *J Appl Microbiol,* 110**,** 1-18.

SONG, Y. Y. & LU, Y. 2015. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry,* 27**,** 130-5.

SPIVEY, E. C., JONES, S. K., JR., RYBARSKI, J. R., SAIFUDDIN, F. A. & FINKELSTEIN, I. J. 2017. An aging-independent replicative lifespan in a symmetrically dividing eukaryote. *Elife,* 6.

STRATFORD, M. 1989. Evidence for two mechanisms of flocculation in Saccharomyces cerevisiae. *Yeast,* 5 Spec No**,** S441-5.

TANAKA, E., BAILEY, T., GRANT, C. E., NOBLE, W. S. & KEICH, U. 2011. Improved similarity scores for comparing motifs. *Bioinformatics,* 27**,** 1603-1609.

TANAKA, N., AWAI, A., BHUIYAN, M. S., FUJITA, K., FUKUI, H. & TAKEGAWA, K. 1999. Cell surface galactosylation is essential for nonsexual flocculation in Schizosaccharomyces pombe. *J Bacteriol,* 181**,** 1356-9.

TOONE, W. M., JOHNSON, A. L., BANKS, G. R., TOYN, J. H., STUART, D., WITTENBERG, C. & JOHNSTON, L. H. 1995. Rme1, a negative regulator of meiosis, is also a positive activator of G1 cyclin gene expression. *Embo j,* 14**,** 5824-32.

TYERS, M., TOKIWA, G. & FUTCHER, B. 1993. Comparison of the Saccharomyces cerevisiae G1 cyclins: Cln3 may be an upstream activator of Cln1, Cln2 and other cyclins. *Embo j,* 12**,** 1955-68.

VACHON, L., WOOD, J., KWON, E.-J. G., LADEROUTE, A., CHATFIELD-REED, K., KARAGIANNIS, J. & CHUA, G. 2013. Functional Characterization of Fission Yeast Transcription Factors by Overexpression Analysis. *Genetics,* 194**,** 873-884.

VALVEZAN, A. J. & MANNING, B. D. 2019. Molecular logic of mTORC1 signalling as a metabolic rheostat. *Nature Metabolism,* 1**,** 321-333.

VAN HEEMST, D. 2010. Insulin, IGF-1 and longevity. *Aging Dis,* 1**,** 147-57.

VAN RAAMSDONK, J. M. & HEKIMI, S. 2009. Deletion of the mitochondrial superoxide dismutase sod-2 extends lifespan in Caenorhabditis elegans. *PLoS Genet,* 5**,** e1000361.

VOGEL, C., SILVA, G. M. & MARCOTTE, E. M. 2011. Protein expression regulation under oxidative stress. *Mol Cell Proteomics,* 10**,** M111.009217.

WAGIH, O. & PARTS, L. 2014. gitter: a robust and accurate method for quantification of colony sizes from plate images. *G3 (Bethesda),* 4**,** 547-52.

WANG, Q., LI, M., WU, T., ZHAN, L., LI, L., CHEN, M., XIE, W., XIE, Z., HU, E., XU, S. & YU, G. 2022. Exploring Epigenomic Datasets by ChIPseeker. *Current Protocols,* 2**,** e585.

WEISMAN, R. & CHODER, M. 2001. The fission yeast TOR homolog, tor1+, is required for the response to starvation and other stresses via a conserved serine. *J Biol Chem,* 276**,** 7027-32.

WEISMAN, R., FINKELSTEIN, S. & CHODER, M. 2001. Rapamycin blocks sexual development in fission yeast through inhibition of the cellular function of an FKBP12 homolog. *J Biol Chem,* 276**,** 24736-42.

WOOD, V., GWILLIAM, R., RAJANDREAM, M. A., LYNE, M., LYNE, R., STEWART, A., SGOUROS, J., PEAT, N., HAYLES, J., BAKER, S., BASHAM, D., BOWMAN, S., BROOKS, K., BROWN, D., BROWN, S., CHILLINGWORTH, T., CHURCHER, C., COLLINS, M., CONNOR, R., CRONIN, A., DAVIS, P., FELTWELL, T., FRASER, A., GENTLES, S., GOBLE, A., HAMLIN, N., HARRIS, D., HIDALGO, J., HODGSON, G., HOLROYD, S., HORNSBY, T., HOWARTH, S., HUCKLE, E. J., HUNT, S., JAGELS, K., JAMES, K., JONES, L., JONES, M., LEATHER, S., MCDONALD, S., MCLEAN, J., MOONEY, P., MOULE, S., MUNGALL, K., MURPHY, L., NIBLETT, D., ODELL, C., OLIVER, K., O'NEIL, S., PEARSON, D., QUAIL, M. A., RABBINOWITSCH, E., RUTHERFORD, K., RUTTER, S., SAUNDERS, D., SEEGER, K., SHARP, S., SKELTON, J., SIMMONDS, M., SQUARES, R., SQUARES, S., STEVENS, K., TAYLOR, K., TAYLOR, R. G., TIVEY, A., WALSH, S., WARREN, T., WHITEHEAD, S., WOODWARD, J., VOLCKAERT, G., AERT, R., ROBBEN, J., GRYMONPREZ, B., WELTJENS, I., VANSTREELS, E., RIEGER, M., SCHÄFER, M., MÜLLER-AUER, S., GABEL, C., FUCHS, M., FRITZC, C., HOLZER, E., MOESTL, D., HILBERT, H., BORZYM, K., LANGER, I., BECK, A., LEHRACH, H., REINHARDT, R., POHL, T. M., EGER, P., ZIMMERMANN, W., WEDLER, H., WAMBUTT, R., PURNELLE, B., GOFFEAU, A., CADIEU, E., DRÉANO, S., GLOUX, S., LELAURE, V., et al. 2002. The genome sequence of Schizosaccharomyces pombe. *Nature,* 415**,** 871-880.

XIANG, L. & HE, G. 2011. Caloric restriction and antiaging effects. *Ann Nutr Metab,* 58**,** 42-8.

YIU, T. 2019. Understanding Random Forest. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [Accessed 07/02/2023].

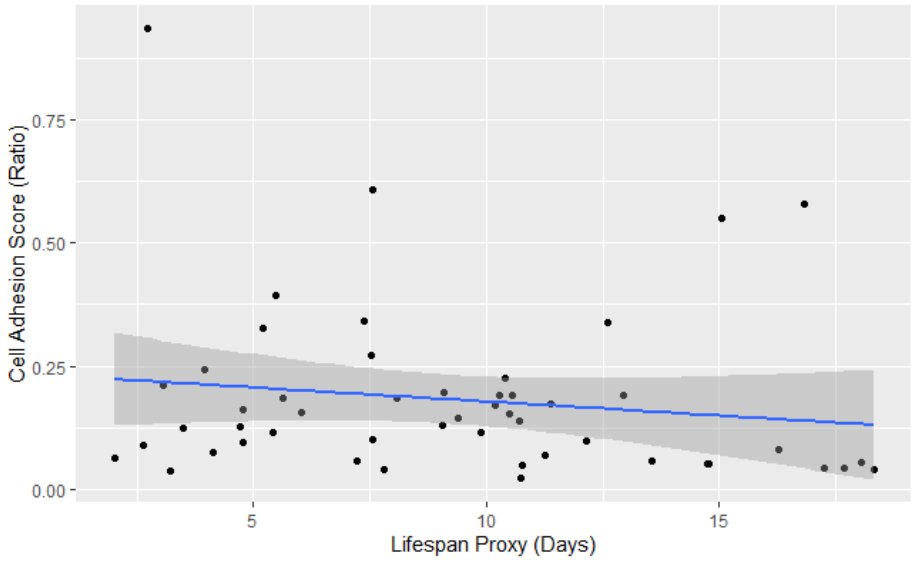
YOUSEFZADEH, M., HENPITA, C., VYAS, R., SOTO-PALMA, C., ROBBINS, P. & NIEDERNHOFER, L. 2021. DNA damage-how and why we age? *Elife,* 10.

YU, G., WANG, L.-G. & HE, Q.-Y. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics,* 31**,** 2382-2383.

ZHANG, M., ZHANG, J., YAN, W. & CHEN, X. 2016. p73 expression is regulated by ribosomal protein RPL26 through mRNA translation and protein stability. *Oncotarget,* 7**,** 78255-78268.

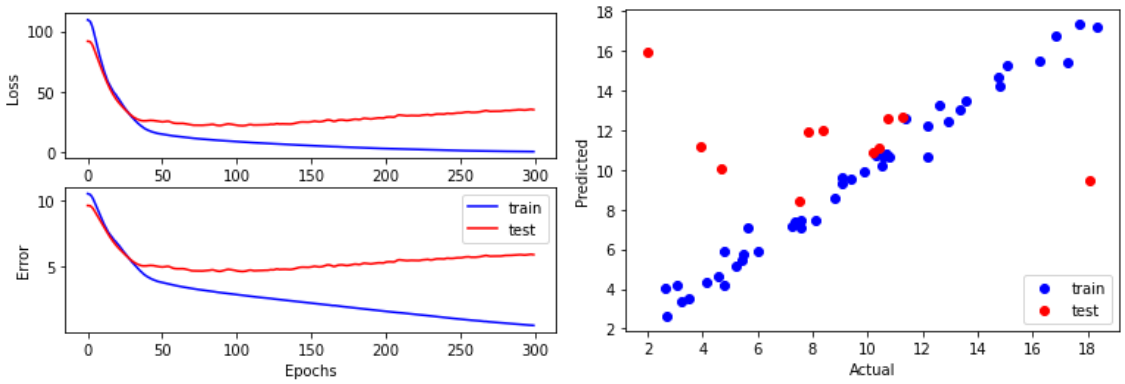
ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W. & LIU, X. S. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology,* 9**,** R137.

# Supplementary Figures




Supplementary Figure 1: Pairwise correlation of mean *DeadOrAlive* lifespan proxy for all repeats and cell adhesion score for the wild type strains.

A B




Supplementary Figure 2: Example of overfitting for a neural network built on dataset NN1 using the Adam optimiser with a learning rate of 0.0005. The loss and error curves for the test data begin to rise after the initial fall (A), and the predicted lifespans for the training data are strongly correlated to the actual values whereas the predicted lifespans for the test data are not correlated to the actual values (B). Both of these show overfitting.