INFERRING BIOLOGICAL NETWORKS FROM GENOME-WIDE TRANSCRIPTIONAL AND FITNESS DATA

By

WAZEER MOHAMMAD VARSALLY

A thesis submitted to

The University of Birmingham

for the degree of

Doctor of Philosophy

College of Life and Environmental Sciences
School of Biosciences
The University of Birmingham
July 2013

ABSTRACT

In the last 15 years, the increased use of high throughput biology techniques such as genome-wide gene expression profiling, fitness profiling and protein interactomics has led to the generation of an extraordinary amount of data. The abundance of such diverse data has proven to be an essential foundation for understanding the complexities of molecular mechanisms and underlying pathways within a biological system. One approach of extrapolating biological information from this wealth of data has been through the use of reverse engineering methods to infer biological networks.

This thesis demonstrates the capabilities and applications of such methodologies in identifying functionally enriched network modules in the yeast species *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. This study marks the first time a mutual information based network inference approach has been applied to a set of specific genome-wide expression and fitness compendia, as well as the integration of these multi-level compendia. This work highlights how network inference can infer potentially novel biological relationships by identifying gene modules that exhibit similar response across samples. This information can then be used to identify strong candidate interactions which can be tested experimentally.

In particular, this work has generated hypotheses in *S. pombe* that have led to a deeper understanding of the relationship between ribosomal proteins and energy metabolism, a recently discovered pathway termed riboneogenesis. To date, this link has only been reported in *S. cerevisiae.* Experimental validation of this hypothesis using ChIP-chip data has led to new theories on the role of energy metabolism enzymes in controlling ribosome biogenesis in *S. pombe*, including the novel finding that fructose-1, 6-bisphosphatase (FBP1) may have roles in both gluconeogenesis and riboneogenesis.

This thesis also demonstrates how the use of multi-level data allows for comprehensive insight into nuclear functions of the *S. pombe* nonsense-mediated mRNA decay protein, UPF1. This study provides a substantial amount of evidence demonstrating the role of UPF1 in DNA replication. The applicability of fitness data in identifying targets of metal and metalloid toxicity in *S. cerevisiae* has also been investigated.

Ultimately, this thesis reports the effectiveness of using a systems biology and network inference approach to identify, elucidate and understand complex biological pathways in *S. cerevisiae* and *S. pombe*.

ACKNOWLEDGEMENTS

First and foremost I want to dedicate my thesis to my parents, Reshad and Seeyreen Varsally. Without their unconditional love, support and sacrifice, I would not be where I am today. Their encouragement and motivation throughout the years, especially during the tougher times gave me the strength to continue pushing forward. I would also like to thank my brother Nadiim Varsally, who has always met me with a smile, no matter the situation.

Professionally, I would like to thank my supervisors Dr. Francesco Falciani and Dr. Saverio Brogna for their guidance and support throughout both my undergraduate and PhD study. Their expertise and insight in their respective fields have been invaluable in helping me progress as a scientist.

I would like to thank all my colleagues. Philipp Antczak, for his vast scientific insight, no matter what problem I had, he always had a solution. To Nil Turan-Jurdzinski for her listening skills and company, especially when the office was quiet. Anna Stincone for teaching me some Italian in our often very entertaining conversations. Jaanika Kronberg for her amazing cake baking skills, they always made my day brighter. Thanks to Harriet Davies whose energy and enthusiasm never failed to make me smile. To Sandip De for his exceptional experimental work in *S. pombe* which helped validate my results and I would also like to thank the rest of the group, Kim Clarke, Rita Gupta, Helani Munasinghe and Peter Davidsen.

I wish to thank all my friends who have been there for me over the years. Special thanks to Anisah for her often humorous but trustworthy advice. To Vanica and Chahat for all the good times we’ve shared and to Yash, for our incredibly long conversations about life.

LIST OF PUBLICATIONS

[1] **Varsally, Wazeer** and Saverio Brogna. UPF1 involvement in nuclear functions. *Biochemical Society Transactions* 40.4 (2012): 778.

[2] De, Sandip and **Varsally, Wazeer** and Falciani, Francesco and Brogna, Saverio. Ribosomal proteins' association with transcription sites peaks at tRNA genes in *Schizosaccharomyces pombe*. *RNA* 17.9 (2011): 1713-1726.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1 Introduction to systems biology approaches for omics data analysis

Systems biology has taken its place as a mainstream approach to research since the late 90s [1]. The overall aim of systems biology is to obtain a quantitative understanding of biological systems by analysing the relationships among their components, including information from and between genes, mRNA, proteins and metabolites [2]. Mathematical models can also be used to describe how each component interacts with each other and to predict their behaviour [3]. Its rise in popularity is due to biologists examining entire biological systems rather than focussing on individual mechanisms. A biological system may be defined as a set of relationships amongst genes, proteins and macromolecules that result in the life and viability of the system [4]. Thereby a system can be defined as a pathway, mechanism, single cell, tissue, organ or an entire complex organism [1]. A systems biology approach requires obtaining information on the various components mentioned above, in accordance with the principle that genes are not the sole keepers of information.

1.1.1 The analysis of genome-wide omics datasets

Systems biology does not only focus on the single ‘omics’ resources available but allows for the inclusion of multiple data sources in an integrative manner. The integration of multiple datasets can reveal cellular mechanisms that would have otherwise remained undetected if only a single omic data source was used. Acquiring such data requires high throughput experiments and therefore, powerful statistical computational analyses are needed [2].

In the next section, I provide an overview of the stages of a typical analytical pipeline and the current bioinformatics tools available for system biology approaches. In this case I focus on microarray technology but the methods and principles can be applied to all other omics technologies.

1.1.2 Microarray data processing and normalisation

To separate the true biological variation from the technical variation due to the laboratory equipment and human error, normalization procedures are necessary. Noisy data can be caused by laboratory instruments, and variation in experimental procedure such as labelling, scanning and image analysis procedures between labs. Several normalisation methods exist to correct for technical differences [5]. Application of normalisation techniques is dependent on the array type (single or two channel arrays, manufacture type). Two channel arrays require different normalisation techniques compared to single channel arrays. Variables such as dye bias can influence signal intensity as microarray designs vary, with some containing more than one feature for a given target. In these cases, specific normalisation techniques need to be applied. Typically, a log2 transformation is applied to the data to remove low signal bias and reduce variation among measures with high magnitude. Robust Multi-array Analysis (RMA) [6], GeneChip RMA (GCRMA) [7, 8], and Microarray Suite (MAS5) [9] are the most commonly used normalisation methods for single channel arrays produced by Affymetrix. RMA normalisation provides background correction for each probe, quantile normalisation and summarisation into expression measurements, with quantile normalisation assuming that gene abundances are distributed similarly across all samples.

GCRMA is an improved version of RMA which incorporates sequence specific probe affinities into the methodology. MAS5 builds regression models for subsets of the entire dataset, and then transforms each probe accordingly. This thesis utilises microarray data from Affymetrix arrays. The methods described can be implemented using the 'affy' [10] package in R [11]. The normalised data can be interrogated using a series of statistical methodologies, which are described below. The acquisition and analysis of fitness data and ChIP-chip data are reported in sections 1.2.1 and 1.2.2 respectively.

1.1.3 Data exploration

Exploratory analysis of the data is an important step in outlier detection and general understanding of the experimental setup. There are a number of techniques which can perform such analysis. Principal component analysis (PCA) [12] and independent component analysis (ICA) [13] for example, belong to a group of dimensional reduction techniques that can be used to summarise the variance across samples into principal components, thereby reducing the dimensionality of the data. The first component would contain the most amount of variance, with the variance decreasing as the number of components increase. The principal components, typically the first two, can be visualised on a 2D (or 3D) plot which represents the differences in samples. Other tools used for exploring large scale datasets include clustering and tree based algorithms that provide a visual representation of the data. Using such methods can reveal the similarity between genes or samples and output results in a visual form. Tools include the self organising tree algorithm (SOTA) [14] and hierarchical ordered partitioning and collapsing hybrid

(HOPACH) [15]. These particular methods output results as a gradient heatmap which represents signal intensity across samples and genes and are useful in visualisation of the correlation structures in the data.

1.1.4 Identification of differentially expressed genes

One of the key questions in analysing biological data, is the whether a set of features are differentially expressed between samples. Numerous statistical approaches have been developed and implemented to answer this question (Table 1.1).

Statistical Test	Applicability
t-test	Compares averages of classes (max 2) and produces a p-value for each gene. Can be one class and two class (paired and unpaired)
ANOVA	Extension of t-test. Compares means across all sample groups. Eliminates the need to perform multiple class comparisons for each pair of classes. Tests include one-way ANOVA and two-way ANOVA
n-way ANOVA	Analysis of multiple factors with the same classification groups.
SAM	Modification of the t-test that removes the stability problem. Can be used for comparing one class, two class (paired and unpaired), multiclass, time course and quantitative data (providing there is a continuous response variable)

Table 1. 1 Summary of most popular statistical tests

Deciding the statistical method to use is dependent on the experimental design. The traditional t-test (or the non parametric version; Wilcox-test) compares the averages of a maximum of two classes, and specifies a p-value for each gene. The p-value is the probability of a significant difference in gene expression between two experimental conditions being due to random chance, a statistically significant low p-value (typically $\leq$ 0.05) indicates strong evidence again this null hypothesis. A statistical threshold is chosen by the user to define genes that are significantly differentially expressed, therefore rejecting the null hypothesis. A drawback to using a t-test is the ‘stability problem’ in which genes with low variance can be statistically significant despite having a very low fold change, creating a bias towards highly reproducible genes. Analysis of variance (ANOVA) is a generalised extension of the t-test, and is not limited to only two classes. This technique compares the means for all sample groups, reducing the requirement for performing two class comparisons between each class pair permutation. The analysis of multiple factors is known as n-way ANOVA. Significance analysis of microarrays (SAM) [16] is a modification of the t-test which removes the stability problem, and can be applied to one class, two class (paired and unpaired), multiclass, quantitative and time course data. The t-test and ANOVA return p-values for each gene.

Due to the often massive size of microarray datasets, p-values require correction for multiple testing. Several correction methods have been developed [17] [18] [19], however the most widely used correction method is known as False Discovery Rate (FDR), published by Benjamini and Hochberg [20]. This method aims to capture the highest number of true positives whilst controlling the number of false positives. Correction for multiplicity testing is essential. With statistical tests, the user typically picks a cut-off threshold representing the probability of identifying false positives; however, this is only valid when a single hypothesis is being tested. Therefore, to correct for the problem of multiple comparisons, an FDR correction must be applied. Studies of multiple FDR thresholds are required to optimise sample separation and functional annotation whilst maintaining a reasonable number of genes, before finally settling on an FDR threshold.

1.1.5 Functional analysis

Despite microarray technologies having the ability to measure gene expression at the genome-wide level, interpreting results of such a large number of genes can prove challenging. Differential expression or clustering analysis alone does not provide a full picture of the molecular changes occurring across experimental conditions. Functional annotation tools test whether, in a given list of genes, genes belonging to a specific functional process are present at a higher frequency than by random chance. One of the most popular and well known databases is The Database for Annotation, Visualisation and Integrated Discovery (DAVID) [21] [22]. DAVID is a web based service that integrates multiple functional annotation databases, including gene ontology (GO) [23] and Kyoto Encyclopedia of Genes and Genome (KEGG) [24]. Therefore, DAVID provides a well validated functional clustering service, which identifies statistically significant enrichment utilising multiple databases and curated pathways [21] [22], and returns corrected FDR scores for any functional enrichment identified.

1.1.6 The current state of network inference

Network inference aims to use the vast amount of data generated by high-throughput technologies and identify interactions between entities such as genes, proteins and experimental perturbations. Network inference approaches provide a means to identify and visualise dependencies between these entities. Constructing networks aid in the understanding of cellular mechanisms, such as predicting regulatory pathways through the identification of network modules. A module is defined as a group of entities that are co-regulated, functionally similar or regulated by a common factor [25].

This type of hypothesis driven research has led to the development and implementation of several reverse engineering methodologies [26], however the choice of which method to apply is dependent on the type of data (steady state or time course measurements) and the number of genes and samples to be analysed. Network inference methodologies can be grouped into gene pairwise association (such as correlation or mutual information), probabilistic methods (such as Bayesian networks) and differential equations.

Probabilistic methods such as Bayesian networks require a large amount of data as the algorithm is based on the estimation of a probability density distribution. Advantages of using Bayesian networks are the ability to infer connection directionality and analyse time course data [27], an advantage not shared with pairwise scoring methodologies. However the dependency on published data makes Bayesian networks mainly suitable for model systems. The Bayesian networks methodology can only be applied to a limited number of genes [27] as the entire network needs to be rescored after each edge manipulation, this limitation is also shared by differential equations [28]. Therefore, Bayesian networks and differential equations are best applied when the focus is on a specific biological question or pathway, which involves the analysis of a limited number of genes. Several methodologies have been developed with the primary focus of inferring networks using a Bayesian approach [27] and differential equations [29].

A key advantage of using pairwise scoring methods over probabilistic methods and differential equations is the ability to use an entire dataset as input, which can contain hundreds or even thousands of genes and samples. Metrics such as correlation and mutual information (MI) can be computed between all gene pairs within the dataset. Correlation algorithms such as Spearman and Pearson are typically suited to datasets that have a smaller number of samples. MI based methods can capture positive, negative, linear and non-linear dependencies; however a large number of samples (> 50) is required. It should be noted that MI based pairwise methods are limited by their inability to infer edge directionality. Therefore, when dealing with static data, containing thousands of genes and samples, pairwise association based methods are the most suitable. Many algorithms have been developed including the MI based methods Algorithm for the Reconstruction of

Accurate Cellular Networks (ARACNE) [30] and Context of Likelihood of Relatedness (CLR) [31].

1.1.6.1 Reverse engineering approaches: ARACNE

In this study, the well validated reverse engineering method ARACNE was used to infer gene networks from large datasets. ARACNE defines an edge as a statistical dependency between gene expression profiles [30]. Potentially indirect connections can be eliminated using the Data Processing Inequality (DPI) principle [30]. This principle involves removing the edge with the smallest information from a triplet of gene connections. One of the biggest advantages of using ARACNE over other means of correlation is that it is able to identify non-linear correlations as well as positive and negative relationships, making it very effective in identifying connections which are biologically relevant [30, 32]. MI is scored between 0 and 1, and similarly to linear correlation methods, the higher the value, the greater the MI that is shared between two genes. However, it is important to note that relationships between genes are statistical dependencies; therefore causality cannot be directly inferred without further experimental validation.

As MI is always non-negative, genes that are in fact mutually independent with no underlying biological connection will also have a positive (albeit low) value. Therefore, when thresholding a network, it is imperative to select an MI value that represents a statistically significant p-value in order to eliminate false positive connections. ARACNE assigns p-values to MI thresholds using a Monte Carlo simulation using different sample sizes and 10^5 gene pairs [30]. This ensures that for each MI threshold, a reliable estimation of the corresponding p-value has been calculated [30]. For this reason, it is essential to test multiple MI thresholds and check their corresponding p-values, as well as the number of nodes and edges retained in the network before deciding on which threshold to use.

1.1.7 Modularisation approaches

An important step in deciphering and reducing the complexity of an inferred network is through the identification of functional modules. A functional module can be defined as a set of genes that exhibit similar expression, are regulated by a common factor, have functional similarities, or are a combination of them all [25]. There exist numerous modularisation approaches with the aim of tackling this challenge. These include identification of modules based on network topology which consist of identifying highly interconnected sub-networks within the larger networks, such as GLay [33]. For a review of network topology based methods of modularisation, please see Li *et al* [34]. Edge weight scoring methods, which involve identifying modules that are highly co-regulated using correlation and MI have also been developed [35].

1.2 Other types of genome-wide data used in this study

In addition to gene expression data, a significant section of this thesis is dedicated to the analysis of fitness data and ChIP-chip data.

1.2.1 The power of fitness data

The use of fitness data is known to be a powerful way of inferring gene function. This involves phenotypic analysis of mutants missing a particular gene; the inference of gene function is then based on the phenotype exhibited by the mutant [36]. This approach allows the essentiality of a gene to be evaluated. Comparative studies using identical conditions between both fitness and expression experiments have shown that gene essentiality is not necessarily linked to differential expression [37]. Four comparative studies performed by Giaever *et al* revealed less than 7% of genes that exhibited differential expression, also exhibited a statistically significant decrease in fitness [37]. This revealed that a statistically significant increase in gene expression is not necessarily linked to optimal growth under stress conditions, and that the post-transcriptional modification and translation regulation of genes also have a key role in determining cell survivability. This discovery also revealed that genes that show no change in expression may still be essential for cell viability. This is a feature that fitness experiments can capture unlike experiments that measure gene expression.

There are numerous genome-wide approaches that can be applied for characterising gene function using phenotypic analysis of mutant strains, including genetic footprinting [38], random mutagenesis [39] and a 'molecular barcode' approach [36] [37] [40]. Genetic footprinting and random mutagenesis are both relatively rapid, however disadvantages include being unable to recover the mutant strains (genetic footprinting approach) and the time-consuming pairing between the mutated gene and the corresponding phenotype (random mutagenesis approach) [37]. Both these random approaches also suffer from the certainty that some genes will elude detection during the screening process. [36] [37]. To overcome these limitations, a 'molecular barcode' approach was developed, which involved systematically deleting each gene from start to stop codon and replacing it with a *KanMX* gene flanked by two distinct nucleotide sequences which can act as a unique identifier for each deletion mutant [36] [37]. The *KanMX* gene has been shown to not affect the fitness of deletion strains [41]. This approach meant that the phenotype exhibited by the deletion strain could be directly associated to the gene deletion. This method of analysis has proven to be effective in yeast, as strains containing each deletion can be analysed in parallel. A culture containing every deletion mutant can be grown, with samples collected at regular intervals. The essentiality of each gene for cell viability can be determined by quantifying the abundance of the unique molecular barcodes using an Affymetrix Tag3 array containing the respective complementary sequences [37]. Yeast strains containing mutants of genes that are essential for cell viability, will rapidly diminish, as would the molecular barcode associated to that gene, whilst mutations in genes that are non-essential will not affect cellular growth. Therefore, each gene can be ranked by their contribution to fitness. This approach can be applied to yeast strains grown under particular stresses such as limited media or toxic metal exposure. Genes essential for growth under different environmental conditions can be quantified, and can provide information on what genes are needed for metal tolerance, growth in glucose limited media or potentially any type of stress. Using the molecular barcode approach in growth fitness analysis is fundamental in understanding gene function in yeast [42] [43] [40]. Bioinformatic methodologies developed to analyse, normalise and assign fitness scores from raw .CEL data have been developed by one of the pioneers of fitness data, Hillenmeyer *et al* [40]. After extracting fitness scores, sections 1.1.3 – 1.1.7 can be applied to fitness data.

1.2.2 An introduction to ChIP-chip analysis

DNA – protein interactions are characteristic of essential cellular processes including DNA replication, transcription and DNA repair [44]. Chromatin immunoprecipitation (ChIP) is an experimental procedure used to investigate the interactions between protein and DNA, specifically a protein is selectively immunoprecipitated from chromatin and the DNA sequences associated to that protein are determined [45]. ChIP-chip is the combination of ChIP with whole-genome DNA microarrays. ChIP-chip approaches are able to map the genome-wide binding profiles between the protein in question and DNA, examples in the literature include the genome-wide mapping of histone proteins [46], cell cycle transcription factors [47] and broader scope analysis that do not focus on specific proteins but instead proteins that generally show DNA binding association [48]. ChIP-chip protocols within yeast and mammalian cells are relatively standard with very little variation [44]. The first step involves cross-linking the protein of interest with the DNA using formaldehyde fixation. After cross-linking the extract is sonicated to shear DNA fragments to a size typically 1 kilobase (kb) or less [49]. DNA protein complexes are identified from the remaining pool of DNA fragments by using either immunoprecipitation (with a protein specific antibody, or an antibody specific to a tagged protein) or affinity purification using a tag that does not require antibodies [44]. The cross-links between the protein of interest and DNA are reversed by heating and the DNA is purified, amplified and labelled with a fluorescent molecule such as Cy5 [44]. In two-colour arrays, the input DNA before ChIP is used as a reference [50], amplified and labelled using a different fluorescent molecule such as Cy3 [44]. The two probes are combined and hybridised to a DNA tiling array. A genome-wide binding profile representing the *in vivo* binding can then be constructed. Regions of the genome bound by proteins can be detected using computational methods. Currently, the number of methods available that can detect protein binding regions are numerous and diverse. Methods include TileMap [51], TiMAT (http://bdtnp.lbl.gov/TiMAT) and MAT (Model-based analysis of Tiling Arrays) [52] and Hidden Markov Model (HMM) based approaches [53]. Comparative studies between the computational ChIP-chip analysis methods revealed MAT outperformed all existing computational ChIP-chip methodologies [52]. MATs ability to take user defined p-values and to analyse experiments with unequal number of control and test samples has made it a leading computational methodology in the analysis of Affymetrix Tiling Arrays [52].

1.3 The biological systems of relevance

1.3.1 *Saccharomyces cerevisiae* (budding yeast) and *Schizosaccharomyces pombe* (fission yeast)

The completion of the *S. cerevisiae* genome sequence was reported in 1997 [54] [55]. The genome is ~12.1 megabases (Mb) and to date contains 6607 open reading frames (ORFs), of which 5060 are verified [56]. *S. cerevisiae* can be genetically manipulated with ease, and is one of the best studied model organisms across such fields as cell biology, systems biology, molecular biology and biochemistry [56]. *S. cerevisiae* has its own dedicated online data repository called the *Saccharomyces* Genome Database (SGD) which is consulted over 45,000 times a week [56]. SGD contains basic information such as DNA sequence, RNA, encoded proteins and protein structure in addition to a wealth of tools that can be used to query the vast amount of data available, including sequence similarity searches.

Similarly, *S. pombe* is a major organism for the study of eukaryotic cellular processes and a model for human disease [57] [58]. *S. pombe* was the sixth eukaryotic genome to be fully sequenced and annotated [57]. The genome is 13.8 Mb and contains 4824 protein coding genes, 43% of which contain introns. Furthermore, a total of 172 *S. pombe* proteins, are known to have similarity to proteins associated to human diseases [57]. Currently, the number of publications related to *S. pombe* exceeds 10,000, with hundreds more being published every year [58]. In order to cope with the astronomical amount of information available, Pombase was established [58]. Pombase is an online curated database that aims to provide access to the wealth of available *S. pombe* genome information curated data from literature, genomic sequence, high throughput studies and additional fungal genomes [58].

S. cerevisiae and *S. pombe* have been model organisms that have driven scientific research for decades [59] [60] [61]. The contribution of *S. cerevisiae* to the scientific community is not only restricted to cell biology and biochemistry, but also economically in terms of food and beverage industries [59]. Furthermore, yeast has a short life cycle with cells doubling approximately every 100 minutes. Together with the relatively simple conditions required for growth [62], it makes yeast an ideal organism for the generation of omic type datasets. Due to these reasons, yeast is often chosen to be the organism of choice for many omics studies. The focus on yeast is apparent when examining the vast number omics datasets in the public domain. The results presented in this thesis focus on *S. cerevisiae* and *S. pombe*.

1.3.2 Differences and similarities between *S. cerevisiae* and *S. pombe*

The most apparent difference is the way in which the yeast cells divide. As its name suggests, *S. pombe* divides by medial fission [63]. *S. cerevisiae,* instead, divides by budding. Despite the difference in cell division, both yeast species have been model species for studying the eukaryotic cell cycle studies since as far back as the 1980s [60]. It has been estimated that *S. pombe* diverged from *S. cerevisiae* around 330 – 420 million years ago [64], as a result the two yeast species exhibit numerous differences at the molecular level. At 4824, the number of genes in *S. pombe* is substantially less than that of *S. cerevisiae*. Studies have reported that for *S. pombe,* the gene density over the complete genome is one gene every 2528 bp, compared to one gene every 2088 bp for *S. cerevisiae* [57]. The total protein coding sequence that resides within the genome differs at 60.2% and 70.5% for *S. pombe* and *S. cerevisiae* respectively [57]. Another difference is in the number of introns, *S. pombe* contains 4730 [57] whilst *S. cerevisiae* only has 282 introns [65]. Carbon utilisation and energy metabolism is another key difference between the two species [66]. *S. pombe* lacks numerous genes encoding enzymes involved in energy metabolism and production; as a result, entire cycles and pathways that are present in *S. cerevisiae* are absent in *S. pombe*. These include the complete lack of the glyoxylate cycle; the lack of glycolytic paralogues, alcohol dehydrogenases, genes regulating glucose repression and the inability to synthesise glycogen and utilise ethanol as a carbon source [66]. Even though there are numerous differences between the two species, *S. pombe* still has 3281 proteins that are homologous in *S. cereviase* [57].

1.3.3 A review of system biology studies in yeast

1.3.3.1 Existing system biology studies utilising gene expression data

In 2000, a pioneering study by Hughes *et al* [67] constructed and characterised a compendium of expression profiles of 300 mutations and chemical treatments in *S. cerevisiae*. They used a hierarchical clustering approach to reveal that mutants induce expected groups of genes, and that the co-regulation of these genes corresponds to a particular phenotype. They then proved that expression profiles could serve as a means for identifying gene functions, by knocking out genes classified as uncharacterised ORFs, and then comparing the expression profile for the deletion mutant against other mutants within the compendium. By using this approach they were able to identify functions for eight uncharacterised ORFs [67]. This study laid the foundation for functional discovery using gene expression compendiums. Other expression compendiums such as those constructed by Hu *et al*, containing 263 transcription factor (TF) knockouts [68] were used to construct a transcriptional regulatory network. They also used gene ontology annotations to understand the biological functions of TFs in *S. cerevisiae*. However, despite being a comprehensive study, the processing and normalisation of their data did not include background and tip correction [69], nor were p-values corrected for multiple testing [70], which may have restricted the amount of useful information that could have been gained from such a study. As a result, Reimand *et al* reprocessed and reanalysed the data, which was found to outperform the original in every way [71], however this study lacked any kind of network analysis, instead opting to identify differentially expressed genes and determining TF binding sites. Gene expression and network inference studies have also been used to reveal the conservation of genetic modules between human, worm, flies and yeast. One particular network based study, published by Stuart *et al*, predicted that additional genes may be involved in essential biological processes, these candidate genes were then experimentally validated [25].

Up to this point the focus has been on gene expression networks in *S. cerevisiae*. However, *S. pombe* also has a huge wealth of gene expression information, none more so than the gene expression study carried out by the Bähler Lab over the last ten years, which to date contains expression data for over 900 samples including gene knockouts, metal exposure and various stress conditions [72]. They used two clustering methods to identify biologically relevant modules characteristic of gene regulation, these were hierarchical clustering across samples and clustering of a gene correlation matrix. The authors aimed to rank genes based on their variability in gene expression across all available samples, and they identified that the most variable genes encoded protein transport, stress response and carbohydrate breakdown [72]. Similarly to the study performed by Reimand *et al* in *S. cerevisiae*, the methodology employed within the Bähler Lab study did not make use of any network inference techniques. Given the size, comprehensiveness and broad sample range of their expression compendium, it makes it a prime candidate for use in network inference studies.

1.3.3.2 Existing system biology studies utilising fitness data

As *S. cerevisiae* was sequenced five years prior to *S. pombe*, the amount of data available and comprehensiveness of the genomic annotation far surpasses that of *S. pombe*. Consequentially there exist more fitness data within the public domain for *S. cerevisiae*, such as the studies performed by Winzeler *et al* [36] and Hillenmeyer *et al* [40]. The fitness study performed by Winzeler *et al* in 1999 though relatively small, analysing only a third of the genome (~2000 ORF deletions) under two conditions (rich media and minimal media) was a landmark paper as it substantiated fitness data as a valid and informative means of elucidating functional classes of essential and non-essential genes [36]. Soon after the paper was published, the importance of fitness data was recognised and the construction of all tagged deletions for *S. cerevisiae* was completed through the collaborative effort between European and North American labs [36]. As mentioned in section 1.2.1, an important discovery was that a statistically significant increase in gene expression is not always indicative of gene essentiality [37]. The study performed by Winzeler *et al* showed that fitness data was able to take into account the post-transcriptional and translational regulation of genes. Since then, the analysis of genome-wide deletion strains in *S. cerevisiae* has steadily become a forerunner in understanding the yeast system, however, the volume of fitness data pertaining to *S. pombe* is currently very limited [73] [74].

Another noteworthy fitness study was that done by Hillenmeyer *et al* [40]. The authors tested ~6000 heterozygous gene deletion strains and ~5000 homozygous gene deletion strains (~1000 genes are known to be essential, therefore they were excluded from the homozygous analysis.). This fitness study was the largest and most comprehensive analysis done in *S. cerevisiae*. Results indicated that 97% of genes in *S. cerevisiae* are actually essential for growth in one of the 1000+ chemical genomic assays they tested.

This suggested that under normal growth conditions some genes may be non-essential for yeast growth, however when exposed to chemical or environmental stress they contribute to a measureable change in yeast fitness. Using a hierarchical clustering approach, genes exhibiting similar co-fitness profiles (two genes exhibiting similar phenotypic behaviour across all samples) were found to be biologically related [40]. This paper served as a means of validating and introducing the scientific community to the potential applications of the fitness compendium they had constructed. The overall result of their paper contained a general overview of their fitness data compendium and validation using three examples of how co-fitness profiles could be used to functionally classify genes.

A limitation with current network studies based on fitness data is that they are limited in terms of scope, often disregarding large amounts of the dataset to focus only on a specific set of genes. These types of studies include understanding specific pathways such as mapping galactose utilisation in *S. cerevisiae* using a state-space model based methodology [75] or assembling the *S. cerevisiae* ubiquitination system using a point-wise MI network [76]. Though these studies focus on specific pathways, they do further validate how fitness data can be used to gain additional insight into already well understood cellular processes. They however, suffer from a limited scope, considering only a subset of the data thereby only encapsulating the organisation of a select group of genes rather than the global organisation of the yeast system.

1.3.3.3 Existing system biology studies utilising proteomic and metabolomic data

The aim of proteomics is to identify and quantify protein abundance and post-translational modification in cells using high-throughput experiments, which involve the simultaneous measurement of all proteins expressed in cells. High-throughput experimental methodologies used to separate and visualise proteins are based on 2D-PAGE, mass spectrometry and multidimensional separations using micro-capillary liquid chromatography [2]. Several studies in yeast utilising these approaches have been used to understand the functional organisation of the yeast proteome [77] [78] [79]. Metabolomics aim to characterise the dynamic response of low molecular weight metabolites such as lipids amino acids and sugars [80] in response to environmental stress or genetic perturbation [81]. As the metabolome represents the cellular integration of structural components such as the transcriptome and proteome, it therefore provides a functional readout of the cellular state [82], this can be especially useful when testing cellular response to a environmental stress, as the metabolome responds earlier than the transcriptome and proteome [83] [84]. In yeast, metabolomics studies have been performed using electrospray mass spectrometry [85], NMR spectroscopy [81] and gas chromatography / time-of-flight mass spectrometry [86]. Several technologies can be combined in order to achieve a more comprehensive study [87].

1.3.3.4 The current state of network inference in *S. cerevisiae* using integrated datasets

The use of single level datasets is useful in inferring the relationship between different entities, and hence aids in the identification of important biological functions. However, an integrative approach in which multi-level data is integrated into a single network can reveal in further detail cellular processes and mechanisms that could otherwise not be deduced using single level data. The identification of network modules representing a common behaviour over diverse datasets using integrative approach has been previously explored by several groups. These include the integration of genomic and proteomic data [88], transcription regulation and protein-protein interaction data [89], gene expression, protein-protein and protein-metabolite interaction data [90], transcription factor binding, gene expression, protein interactions and phenotypic sensitivity [91]. Though the overall aim of each study varied, the main take home message was that integrating data from several diverse sources provides complementary information which can be used to build more detailed and comprehensive networks representative of yeast. Key biological interactions and relationships can then be hypothesised based on analysis of the network, and then experimentally validated. These studies are all limited due to the lack of comprehensive fitness data included within their network construction. Though Tanay *et al* did include some fitness data, only 30 samples were considered [91], this low number of samples is expected as their study was published in 2003, when fitness analysis was in its early stages. Now that a comprehensive *S. cerevisiae* fitness compendium is available [40], it provides a perfect source for constructing the first fitness based global *S. cerevisiae* regulatory network using an MI based method, as well as the means to construct integrated networks using robust and comprehensive fitness data.

1.4 Why focus on the relationship between ribosome biogenesis and energy metabolism?

In 2011, Clasquin *et al* reported, in a pioneering study, that ribosome biogenesis and energy metabolism were in fact directly correlated in *S. cerevisiae* [92]. This discovery was revealed by applying a metabolomic screening process on genes of unknown function. Candidate genes for the screening were determined by comparative sequence analysis, with the aim of identifying uncharacterised genes which contained domains similar to already known enzymes. The deletion of one particular gene, which they named *SHB17* (previously known as *YKR043c* and reported to potentially catalyse fructose-1, 6-phosphate [93]) lead to the accumulation of seven- and eight-carbon mono- and bisphosphorylated metabolites, particularly those involved in the non-oxidative arm of the pentose phosphate pathway (PPP). Further experiments revealed that SHB17 was an essential sedoheptulose bisphosphatase; its main role was to link glycolysis and the non-oxidative arm of the PPP in one of a series of six key reactions. This newly discovered pathway was named 'riboneogenesis', and consisted of thermodynamically driven pathway that converts glycolytic intermediates into ribose-5-phosphate, independently of NADPH production (see section 1.5.2.2) [92]. This discovery revealed that the process of ribosome biogenesis was in fact dependent on the glycolysis pathway. Further analysis suggested that key enzymes involved in riboneogenesis are in fact correlated to the yeast metabolic cycle, a cycle in which yeast cultures that are nutrient starved synchronise metabolism and cell cycle processes so that the culture cycles between respiration and fermentation (see section 1.5.2.1) [94]. These results demonstrated that the decision to undergo ribosome biogenesis was in fact dependent on many cellular processes, one of the significant of which was glycolysis. This recent discovery means that no comprehensive network based studies have been undertaken with the aim of investigating riboneogenesis in *S. cerevisiae* or *S. pombe*. In addition to revealing the global organisation of both yeast systems, a network approach may identify additional genes with potential roles in riboneogenesis. To date, the conservation of riboneogenesis in *S. pombe* has only been hinted at, with a result that doubly labelled seven-carbon monophosphorylated metabolites are observed in *S. pombe* cells given [6-$^{13}C_1$]-glucose [92]. Given the metabolic differences between *S. pombe* and *S. cerevisiae* (described in section 1.3.2), applying diverse datasets such as fitness and gene expression data with a reverse engineering approach may lead to a further understanding of riboneogenesis in both yeast species, an investigation which to date, has not been tackled.

1.5 Yeast ribosome biogenesis and energy metabolism

1.5.1 Yeast ribosomal proteins

Ribosomes are complex macromolecular machines that are responsible for the production of proteins in every living cell. They are composed of both RNA and protein, to form a large multifunctional complex. Structurally, ribosomes differ between prokaryotes and eukaryotes. In eukaryotes, ribosomes are 80S, each consisting of a small (40S) and large (60S) subunit. The large subunit is composed of 5S, 28S and 5.8S ribosomal RNAs (rRNAs) [95] and the smaller subunit is composed of an 18S rRNA typically between 1800 - 1900 nucleotides in length [96]. There are just under 80 ribosomal proteins (RPs) conserved across eukaryotes [97]. RPs have remained highly conserved during evolution most likely due to their often critical functions in their biogenesis, function and structural integrity of ribosomes [98] [99]. Specifically, RPs assist in shaping the rRNA into the correct tertiary structure and maintaining an optimum configuration [100] as well as rRNA maturation, nuclear export and ribosomal subunit biogenesis. As stated above, RPs are some of the most abundant proteins in both eukaryotic and prokaryotic cells. Typically most RPs are very basic with a pI of >10. Few exceptions include the acidic phosphoproteins P0 – P3 in eukaryotes [101].

Within *S. cerevisiae* and *S. pombe*, most RP genes are duplicated with the protein paralogues being either identical or very similar. However knockout mutations on the two paralogues lead to different phenotypes within *S. cerevisiae* [102]. For example the deletion of *RPL12a* lead to the up-regulation of genes involved in amino acid metabolism, while a deletion of *RPL12b* up-regulated genes which encoded products that localise to the nucleus and repressed genes involved in cell wall synthesis. *Komili et al* suggested that there are different ribosome subtypes in the cell which are distinguishable by the RP paralogues they carry [102]. The results suggested that RPs may be implicated or have regulatory roles in cellular functions not directly related to the ribosome structure. RPs are ubiquitous, abundant and RNA-binding, making them prime candidates for recruitment for functions outside of ribosomal biogenesis and translation. Though they are involved in balancing RNA synthesis and protein components of ribosomes, numerous papers have reported extra ribosomal functions of ribosomal proteins. Numerous ribosomal proteins have been extensively studied including L7, RACK1, L13a and S3 in *S. cerevisiae* and *Drosophila* [101]. A paper published in 1996 detailed more than 30 extra-ribosomal functions of RPs within *E. coli*, *D. melanogaster*, *H. sapiens*, *S. cerevisiae*, *R. rattus* and *M. musculus* [100]. The significance of this review showed a culmination of data identifying that extra-ribosomal functions are conserved amongst eukaryotic organisms and prokaryotic organisms alike. Since then, there have been numerous linking RPs to non-ribosomal functions [98].

1.5.2 The link between ribosome biogenesis and energy metabolism pathways

In yeast, ribosome biogenesis is a complex and resource expensive process requiring the coordinated regulation of three RNA polymerases [103], transcribing ~150 rRNA genes and ~137 RP genes [97]. Furthermore, up to 60% of cellular transcription is dedicated to rRNA transcription and 90% of mRNA splicing is dedicated to RPs during rapid cell growth [97]. The rate of ribosome biogenesis depends on the physiological demands of the cell, such as cell growth and viability. Notably, studies in *S. cerevisiae* within the last few years have reported that linkage between cell growth, cell function, cell viability and ribosome biogenesis are all in fact linked. Examples of these studies include the yeast metabolic cycle [104] and the discovery of riboneogenesis [92].

1.5.2.1 The metabolic cycle – The global co-ordination of cellular processes by mRNA oscillations

Metabolic cycles were observed in *S. cerevisiae* cultures growing in continuous conditions over 30 years ago [105], however not till recently has the underlying regulatory mechanisms been investigated [94] [104]. The yeast metabolic cycle (metacycle) can be described as the genome-wide co-regulation of genes which synchronise diverse cellular processes with cell division and metabolism [106], essentially allowing yeast cells to switch between respiration and fermentation states in a synchronised manner when grown in nutrient limited culture. Synchronised cells were found to oscillate gene expression based on progression through the metacycle [104]. The metacycle contains three key phases, the oxidative phase, and the reductive phase, split into building and charging phases (Figure 1.1).

The oxidative phase is characterised by the massive up-regulation of genes involved in ribosome biogenesis, ribosomal proteins, translation initiation and amino acid synthesis in preparation of protein synthesis [106] [94] (Figure 1.1). The concentration of NADPH and acetyl coenzyme A (acetyl-coA) are at their highest, providing the oxidisable metabolites needed for rapid cellular respiration. An up-regulation of ribosome biogenesis and protein synthesis genes has also been observed in *S. pomb*e however, unlike *S. cerevisiae*, the peak occurs within the G2 phase not G1 phase [107].

The lack of oxygen and low NADPH concentrations marks the entry into the reductive phase and the beginning of DNA replication. This phase is characterised by the up-regulation of genes involved in mitochondrial building and DNA replication machinery (Figure 1.1). The reductive-charging phase follows. The concentration of oxidisable metabolites (predominantly NADPH and acetyl-coA) [106] [94] increases through the up-regulation of genes involved in carbohydrate breakdown (Figure 1.1). At the onset of the oxidative phase, oxygen levels are high and the concentrations of oxidisable metabolites are at their peak, preparing the cells for protein synthesis.

Each phase is defined by the up-regulation of a specific group of cellular processes [94]. The duration of each phase is almost identical [104] [94], however length of a single metabolic cycle remains controversial and has been reported as ~40 minutes [104] and ~300 minutes [94].

Reductive – charging phase

Up-regulation of genes that encode

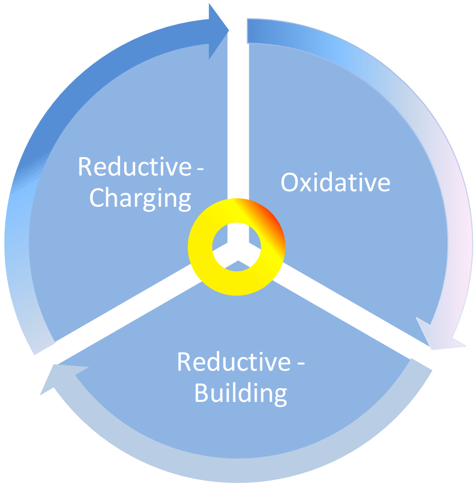
- Non-respiratory modes of energy generation
- Production of acetyl-coenzyme A (acetyl-CoA) and NADPH
- Glycolysis and fatty acid oxidation pathways.
- Genes involved in carbohydrate breakdown
- Vacuolar and proteasomal transcripts
- Ubiqtuination machinery

Characteristics

- Cells prepare for the next respiratory burst

-Concentration of charged oxidisable metabolites reaches its peak

- Cell division is strictly confined to the reductive phases of the metacycle, beginning after DNA replication and ending before the oxidative phase



Oxidative phase

Up-regulation of genes that encode

-Components of the translational machinery, RPs, amino acid biosynthetic enzymes, initiation factors

- Ribosome biogenesis

- Small nuclear RNAs, RNA modification enzymes and machinery, and proteins required for the uptake of sulphur.

Characteristics

- Maximum concentrate of ATP, and highest respiration rates suggesting a tightly coupled link between protein synthesis and energy availability

-Transcript accumulation rises and falls quickly, suggesting RNA is synthesised and degraded rapidly

- G1 phase of cell cycle *S. cerevisiae*

- G2 phase of cell cycle *S. pombe*

Reductive-building phase

Up-regulation of genes that encode

- Nuclear encoded genes for mitochondrial proteins

-Mitochondrial functions including DNA replication, respiration and protein import.

- Spindle pole components
- Histones

Characteristics

- Dedicated to rebuilding mitochondrion
- DNA replication starts at the end of oxidative phase, and ends prior to the charging phase

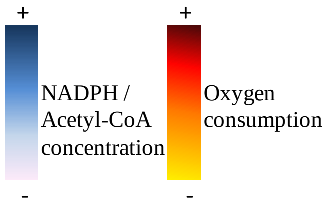


Figure 1. 1 Summary of the metacycle.

A cartoon summarising the culmulative results reported by Tu *et al* [94], Futcher [108] and Reinke *et al* [106]. The metacycle consists of three key phases, the oxidative phase which marks the up-regulation of ribosome and RNA related genes, and utilises stored oxisable metabolites for use during protein synthesis. The reductive phase in general is host to DNA replication and cell division. The reductive-building phase is characterised by the up-regulation of mitochondrial related genes and the formation of spindle poles, as well as the onset of DNA replication. The reductive-charging phase consists of the generation of oxidative metabolites with the up-regulation of glycolysis and fatty acid oxidation. Vacuolar and ubiquitination genes are also up-regulated. Cell division ends before the start of the oxidative phase. The red gradient represents oxygen consumption, and the blue arrow gradient represents concentrations of oxidisable metabolites.

1.5.2.2 Riboneogenesis – Connecting glycolysis and pentose phosophate pathway to ribosome biogenesis

Typically glucose is consumed through two major routes, glycolysis and the pentose phosphate pathway (PPP). Glycolysis is split into two phases, the preparatory phase (where ATP is consumed) and the payoff phase (where ATP is produced). The rate of glycolysis is regulated by numerous enzymes including phosphofructokinase, fructose bisphosphatase and pyruvate kinase which all catalyse thermodynamically favoured reactions. The PPP consists of the oxidative phase and the non-oxidative phase; glucose enters the former through glucose-6-phosphate dehydrogenase and the latter via glycolytic intermediates generating ATP, NADPH and ribose for DNA and RNA synthesis [92].

First reported in 2011 in *S. cerevisiae*, riboneogenesis is a thermodynamically driven pathway which converts glycolytic intermediates into ribose-5-phosphate (R5P) by linking the PPP to the preparatory phase of glycolysis in a series of six thermodynamically driven reactions (Figure 1.2, Table 1.2) [92]. Sedoheptulose-1, 7-bisphophatase (SHB17) is an essential enzyme reported to catalyse the first committed step into riboneogenesis. The production of R5P via riboneogenesis occurs independently of NADPH production, suggesting that riboneogenesis flux is favourable when the cellular demand for R5P exceeds the demand for NADPH or when the oxidative PPP cannot meet the demand for R5P [92].

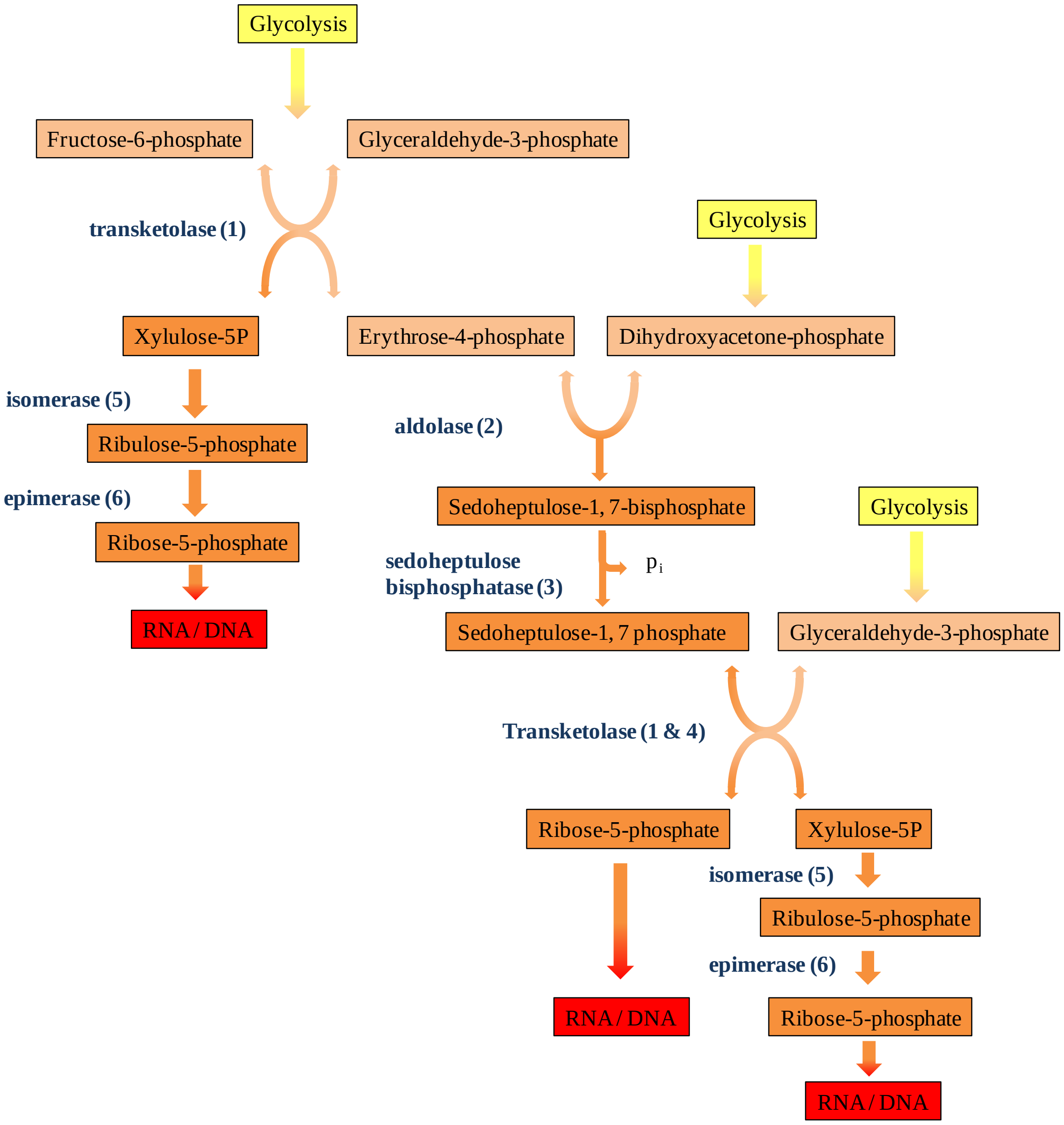


Figure 1. 2 The key steps and enzymes involved in riboneogenesis.
This flow chart has been constructed based on the information presented by Clasquin *et al* [92]. It shows the six enzymatic steps involved in riboneogenesis. Glycolytic intermediates can enter riboneogenesis at three stages (represented in yellow). Enzymes which catalyse each step are represented by the dark blue text; the number in brackets is in reference to the reactions shown in Table 1.2. Of key importance is reaction 3, catalysed by sedoheptulose bisphosphatase (SHB17), this step in riboneogenesis is the thermodynamically driven reaction that commits cells into the riboneogenesis pathway.

Figure 1.2 shows the key reactions in riboneogenesis where the numbers in brackets represent the reactions as detailed in a table 1.2. Glycolytic intermediates in the form of fructose-6-phosphate and glyceraldehyde-3-phosphate enter the riboneogenesis pathway and are catalysed by transketolase TKL1 producing erythrose-4-phosphate (E4P) (Table 1.2, reaction 1). E4P and the glycolytic intermediate dihydroxyacetone-phosphate are catalysed by fructose bisphosphate aldolase (FBA1) to produce sedoheptulose-1, 7-bisphosphate (SBP) (Table 1.2, reaction 2). SHB17 hydrolyses SBP to sedoheptulose-7-phosphate (Table 1.2, reaction 3). Transketolases, TKL1 and TKL2 then convert sedoheptulose-7-phosphate (S7P) to R5P and xylulose-5P (X5P) (Table 1.2, reaction 4). The catalysis of X5P to R5P occurs in two steps. Firstly X5P is converted to ribulose-5-phosphate (Ru5P) by ribulose-5-phosphate epimerase (RPE1) (Table 1.2, reaction 5). Finally, Ru5P is converted to R5P by ribose-5-phosphate ketol-isomerase (RKI1) (Table 1.1, reaction 6). The overall pathway leads to the production of three R5P units [92].

Reaction	Substrate		Product	Enzyme class
1	fructose-6-phosphate + glyceraldehyde-3-phosphate	↔	erythrose-4-phosphate + xylulose-5-phosphate	transketolase
2	erythrose-4-phosphate + dihydroxyacetone phosphate	↔	sedoheptulose-1, 7-bisphosphate	aldolase
3	sedoheptulose-1, 7-bisphosphate	→	sedoheptulose-7-phosphate + Pi	sedoheptulose-1, 7-bisphosphatase
4	sedoheptulose-7-phosphate + glyceraldehyde-3-phosphate	↔	xylulose-5-phosphate + ribose-5-phosphate	transketolase
5	xylulose-5-phosphate	↔	ribulose-5- phosphate	epimerase
6	ribulose-5- phosphate	↔	ribose-5-phosphate	isomerase

Table 1. 2 The six steps of riboneogenesis.
Column 1 represents each step in riboneogenesis as numbered in Figure 1.2. Column 2 shows the substrate(s), column 3 indicates whether the reaction is reversible, column 4 shows the product(s) and column 5 shows the enzyme class responsible for the reaction. Reaction 3, catalysed by SHB17 is the key reaction that commits cells to the riboneogenesis pathway

1.5.2.3 The intricate relationship between the metacycle and riboneogenesis

SHB17 expression is reported to oscillate in tandem with the metacycle [92], with *SHB17* expression peaking together with ribosomal proteins during the oxidative phase. Other key enzymes involved in riboneogenesis have been shown to peak during ribosome biogenesis, including *TKL1* and *RKI1* [92]. This suggests that the up-regulation of *SHB17* and other riboneogenesis genes coincides with the peak demand for ribose phosphate, a signature of ribosome biogenesis.

Riboneogenesis flux is dependent on glycolytic intermediates; therefore, the rate of glycolysis dictates the rate of riboneogenesis. In the metacycle, the up-regulation of glycolysis genes occurs during the reductive-charging phase, which may increase the concentration of glycolytic intermediates leading to rapid flux through the riboneogenesis pathway upon the onset of the oxidative phase, which in turn leads to the up-regulation of ribosome related genes. The discovery of riboneogenesis two years ago, helped explain the oscillating gene expression during the metacycle, and determined that glycolysis has a potentially essential role in ribosome biogenesis.

1.6 Aims and outline of this thesis

This thesis reports a systems biology approach to study two yeast species, *S. cerevisiae* and *S. pombe*. The main goal of this thesis is to construct and interrogate networks derived from large comprehensive multi-level yeast compendia, including fitness and expression data, in both an individual and integrated manner. To accomplish this task, I used an MI based reverse engineering approach. As such, this study marks the first time that Hillenmeyer's fitness compendium [40] has been used to construct a network that encapsulates both the global and local organisation of the yeast system. This approach ensures that the scope of the study remains broad and genome-wide, rather than limiting it to a specific pathway as reported in existing literature. This same analytical pipeline has also been applied to a comprehensive TF knockout gene expression dataset in *S. cerevisae* published by Reimend *et al* [71] and an extensive *S. pombe* expression compendium produced by Bähler lab [72]. By utilising available compendia in this way, I have developed several interesting hypotheses between genes that modularise together and in doing so identified candidate genes which could be used for experimental validation of these hypotheses.

The process of riboneogenesis is particularly focused on within the network based chapters, with the aim demonstrating how network analysis can aid in further understanding this novel pathway in *S. cerevisiae* and determining the degree of conservation of riboneogensis in *S. pombe*. The work presented in this thesis is the first time a reverse engineering network based approach has been used to investigate riboneogenesis in both *S. cerevisiae* and *S. pombe*. Network analysis in *S. pombe* revealed a degree of conservation of the riboneogenesis pathway, and more importantly, the novel results suggested that the gluconeogenesis enzyme FBP1 may substitute for SHB17 as the key enzyme that controls riboneogenesis flux.

In addition, I report several other potentially interesting relationships based on the network analyses which may warrant further investigation in the future. Furthermore, I demonstrate the diverse applications of fitness data and network biology in identifying adverse outcome pathways for metal and metalloid toxicity in *S. cerevisiae.*

I have also investigated the interaction of RPs with chromatin using genome-wide ChIP-chip data in *S. pombe*. This data, provided by a collaborator, was also used in conjunction my constructed *S. pombe* network, with the aim of elucidating the underlying biological organisation of the yeast system and to further aid in understanding riboneogenesis in *S. pombe*.

Finally, I investigated the cytoplasmic nonsense-mediated mRNA decay (NMD) protein, UPF1 in *S. pombe.* This study involved pooling multiple data types with the aim of identifying potential nuclear roles of UPF1 that are unrelated to NMD. This included the analysis of genome-wide ChIP-chip data and expression data from multiple sources, in addition to my own constructed *S. pombe* expression network. Together, they provided compelling evidence that UPF1 is involved in DNA replication.

CHAPTER 2: INFERENCE AND ANALYSIS OF A *Saccharomyces cerevisiae* GENE FITNESS NETWORK

2.1 Introduction

Reverse engineering methods provide means of inferring a network without prior knowledge. The advantage of using a reverse engineering approach is that they can be used on any omic technology. Given the wealth of omic data that has become available over the last decade or so, a reverse engineering approach combined with visualisation or clustering algorithms provide an ideal means of exploring the data. Prior to 2008 the concept of reverse engineering biological pathways from observational data had been focussed on gene expression data and elucidating the transcriptional regulatory network of *S. cerevisiae* [71] [68] [109]. However, recently it has been possible to generate genome-wide libraries of mutant organisms where mutations are marked by a sequence tag then inserted into the genome. Such libraries can be used to test the effect of stress such as chemical exposure on each individual mutant strains in single experiments. The use of fitness data is a powerful way of determining gene function, as mutant yeast strains can be grown, each containing a knocked out or mutated gene and the resulting phenotype can be used to determine gene function. Yeast fitness data has already been validated as a means of gaining additional biological insight into cellular processes, such as targets for chemotherapeutic drugs [43] and temperature response [42].

The aim of the work described in this chapter is to reverse engineer networks representing the relationship between fitness profiles of the different mutant strains using these data. The advantage of using fitness data over expression profiling is that I can directly infer networks based on gene functional association rather than on similarity of transcription. I use the network inference technique ARACNE, its ability to identify non-linear correlations makes it an effective tool in identification of biologically relevant connections [32].

The datasets used in this analysis are two of the largest fitness data compendiums available for *S. cerevisiae.* The first is a publicly available dataset (Hillenmeyer's dataset) that contains fitness data representing exposure to 309 unique chemicals [40]. The second dataset is a still unpublished study developed by Prof. Chris Vulpe (Berkeley University, USA). It represents a smaller set of chemical exposures (11 unique conditions). As both these datasets represent fitness data in *S. cerevisiae*, they should be able to provide complementary information on the activity of different chemical subsets, and would therefore allow the identification of genes that are strongly correlated to each other at the phenotypic level.

This approach demonstrates the usefulness of network inference to understanding cell physiology from genome-wide fitness data, with a particular focus on investigating the links between ribosomal proteins (RPs) and energy metabolism pathways. The data demonstrates the phenotypic link between RPs and energy metabolism pathways is conserved in both *S. cerevisiae* fitness datasets, consistent with current publications on riboneogenesis.

2.2 Methods

2.2.1 The Biological System

The overall aim of this study is to reverse engineer and analyse the structure of biological networks representing phenotypically linked genes. Such links are defined as a correlation between the relative fitness of two yeast strains mutated in two given genes across a range of experimental conditions. In order to achieve this, I used the two previously aforementioned fitness datasets (see Table 2.1 for details).

For Hillenmeyer's fitness compendium, I focused on a subset of the data representing the growth fitness of a population of heterozygous mutant yeast strains at 20 generations, (95% of samples in the whole dataset**)** grown in the presence of environmental and chemical stressors. For clarity, the experimental procedure for the construction and use of the *S. cerevisae* strains library are reported briefly below.

Gene disruption was achieved by deleting each gene from the start to stop codon and replacing it with the *KanMX* deletion cassette via mitotic recombination [37]. The distinct 20-nucleotide sequences flanking the *KanMX* gene act as a unique identifier for each deletion mutant. Such an approach is advantageous over random mutagenesis as the mutant phenotypic reflects the complete loss of the gene. In addition, genes will not elude detection even when a large number of genes are screened [37]. Gene deletions were verified by several polymerase chain reactions (PCRs). Whole genome parallel analysis of *S. cerevisiae* could be achieved due to the unique 'barcode like' flanking sequence linked to each gene deletion. For each experiment (rich medium, altered environmental conditions etc) a culture containing every deletion mutant is grown. Samples are collected at various time points during growth. Quantification of each deletion strain is achieved by hybridising the unique flanking identifier sequences to an Affymetrix Tag3 array containing the respective complementary sequences [110]. The essentiality of a gene is relative to how rapidly the corresponding deletion strain diminishes in the culture, therefore each genes contribution to yeast fitness can be analysed in a single experiment [37].

Hillenmeyer's compendium [40] contains a total of 726 samples with 309 unique experimental conditions analysed using a single experimental replicate design. In order to validate findings with this large dataset and to further investigate metal toxicity (Chapter 8) an independent dataset (provided by Vulpe Lab of Berkeley, California) representing 11 unique conditions across 99 samples was used. Tables containing chemicals used in Hillenmeyer's and Vulpe's study have been constructed and are included on the supplementary CD, in folder 'Chapter 2'.

Dataset	**No. of Deletions**	**Gen**	**Samples**	**Unique Conditions**	**Genome covered**	**Condition summary**	**Replicates**	**Ref**
Hillenmeyer Heterozygous data	~6000	20	726	309	~97%	Chemical and environmental stress conditions	1	[40]
Vulpe Heterozygous data	~4500	15	99	11	~65%	Metals, arsenicals	3-12	NA

Table 2. 1 Summary of datasets used in the fitness analysis.
Detailing the number of deletions, the generation that the readings were obtained, number of samples, and the number of unique conditions.

2.2.2 Analysis Strategy

Figure 2.1 describes in a schematic format the analysis strategy used to reverse engineer and analyse the yeast fitness networks. Initially the dataset is normalised and sample outliers are detected by visual inspection of a PCA plot (Figure 2.1A). The resulting dataset was then used as an input of the well-validated reverse engineering method, ARACNE [30] to infer the structure of the underlying fitness network. No edges have been eliminated using data processes inequality (DPI). Non-statistically significant mutant to mutant connections were eliminated using a threshold of $p<10^{-35}$ (corresponding to an MI $\geq$ 0.15) for the network generated from Hillenmeyer's dataset and $p<10^{-17}$ (corresponding to an MI $\geq$ 0.35) for the network generated from Vulpe's dataset. Different MI thresholds for each dataset were required in order to retain a similar number of edges within each network to allow for comparison and validation analysis.

The network was visualised in Cytoscape using a force driven layout (Figure 2.1B). Force driven layout uses a mechanical model of a network where the edges are represented by forces attracting the nodes with intensity proportional to the MI between each gene pair [111]. The result is that nodes connected by edges with a high MI are located closer together than those with low MI. Groups of nodes connected by multiple edges with lower MI values will also be represented close together in the network [111]. Due to the large network size, I use GLay clustering, a method of modularisation which identifies modules on the basis of connectivity [33]. These GLay identified modules are then mapped onto the parent network (Figure 2.1C). Typically modules of 300 nodes or greater required a further level of GLay clustering in order to comprehensively analyse and annotate. (Figure 2.1D). The functional annotation web-based tool DAVID [21, 22] was used to test whether there is any functional enrichment within each sub-module (Figure 2.1E).

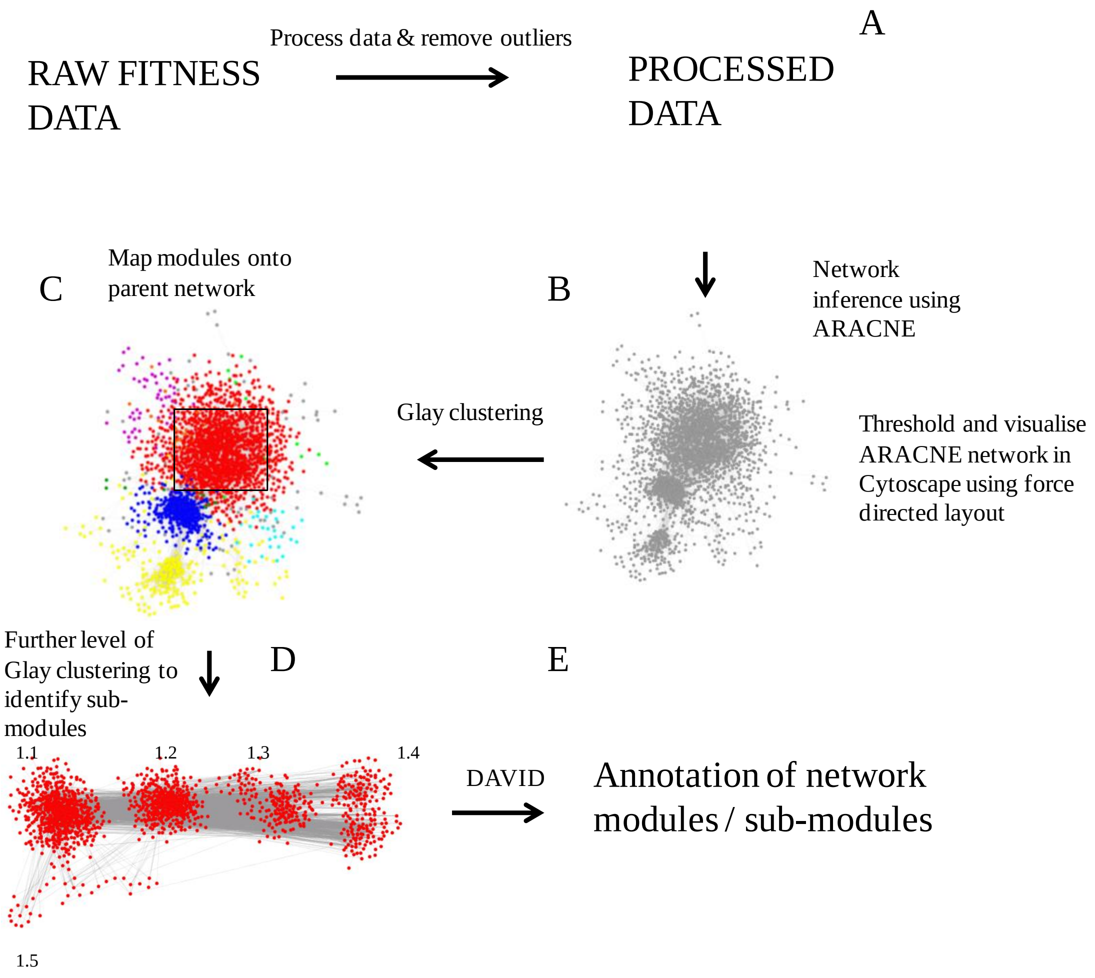

A
Process data & remove outliers
RAW FITNESS DATA
PROCESSED DATA
C
Map modules onto parent network
B
Network inference using ARACNE
Glay clustering
Threshold and visualise ARACNE network in Cytoscape using force directed layout
Further level of Glay clustering to identify sub-modules
D
E
1.1
1.2
1.3
1.4
1.5
DAVID
Annotation of network modules / sub-modules

Figure 2. 1 Workflow for reverse engineering and analysing *S. cerevisiae* fitness networks. A sample flow diagram representing the analytical pipeline for this analysis. The initial step involves processing the fitness data. Once the data has been normalised and fitness values calculated, bad samples have to be removed (Panel A). The network is inferred using ARACNE, the results are thresholded and the network visualised in Cytoscape using a force directed layout (Panel B). Modules are identified using GLay and mapped onto the parent network (Panel C). If the modules are too large, a further level of GLay clustering is done to identify sub-modules (Panel D). Functional enrichment is determined used DAVID (Panel E).

2.2.3 Data processing

2.2.3.1 Public Domain dataset (Hillenmeyer *et al*)

The compendium contains an extraordinary amount of genome-wide fitness data for *S. cerevisiae* encapsulating a diverse set of treatments including drugs approved by the World Health Organization, environmental stresses including depletion of amino acids and vitamins in addition to testing growth responses of cells after exposure to over 300 small molecules [40]. Before any network inference can occur, the dataset must first be processed (Figure 2.1A). The dataset developed by Hillenmeyer *et al* was already available in a processed and normalised format with problematic samples removed, within the supplementary material [40]. The authors used the processing pipeline described below:

1. The raw intensity for each input CEL file is mapped to their associated strain-tags
2. The data is normalised so that each experiment has a mean intensity of 1500
3. Both fitness log ratios and significance values (1 or 0) are calculated for a given set of control and treatment experiments using the normalized data.

The log ratios were calculated using the formula below.

$$log2\left(\frac{avg(c)}{avg(t)}\right)$$

Where *avg(c)* is the average normalized intensity of the tag across the control data and *avg(t)* is the average normalized intensity of the tag across the treatment data. Therefore heterozygous deletion strains that showed increased resistance would have a negative log score, and mutations that conferred sensitivity to yeast would have a positive log score.

The final dataset contained the log ratios representative of fitness, for each gene deletion strain across all 726 chemical conditions.

2.2.3.2 The Chris Vulpe dataset

Similarly to Hillenmeyer's fitness data, high-density Affymetrix Tag3 arrays were used in Vulpe's analysis. This dataset was processed using the pipeline developed by Hillenmeyer *et al*. For consistency, I used the original Perl code developed by the authors, which is available on their supplementary website. In order to make direct comparisons to Hillenmeyer *et al's* dataset, I used only the 15G samples from the Vulpe dataset.

Principal component analysis (PCA) was performed on the processed data to identify any potential outliers (Figure 2.2). PCA identified nine outliers, right of which were the samples for the chemical S-(1,2-dichlorovinyl)-L-cysteine (DCVC), which were done by the same researcher, suggesting a potential bias. The remaining outlier was for a single sample of trichloroethanol (TCEtOH), as highlighted in Figure 2.2. Together, the outliers made up 11% (9/99) of the 15G samples, and were removed, leaving 90 samples.

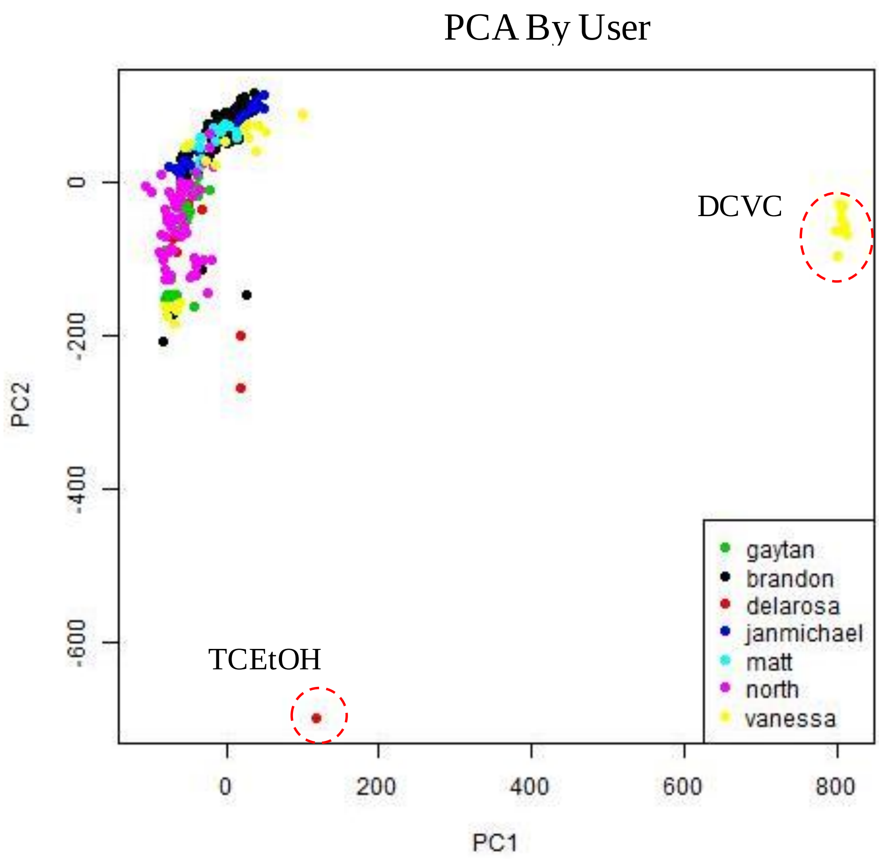


Figure 2. 2 PCA of Vulpe Labs processed fitness data.
Samples are coloured by scientist as shown in the legend. Outlying DCVC and TCEtOH samples which were excluded from the analysis are highlighted with the red dashed circles

2.2.4 Network Inference

The primary aim of this study is to identify gene-to-gene relationships from the fitness data described above. This was achieved using ARACNE [30, 32]. ARACNE infers the interaction between pairs of variables using an information theoretical approach based on mutual information (MI) (Figure 2.1B).

For the Hillenmeyer dataset, statistically significant edges were selected using an MI threshold $\geq$ 0.15 ($p<10^{-35}$) (Table 2.2). This value was chosen arbitrarily; however, it represents a high stringency cut off and retains approximately half the total number of genes. No edges have been eliminated using the data processing inequality (DPI).

For the Vulpe dataset, statistically significant edges were selected using an MI threshold of MI $\geq$ 0.35 ($p<10^{-17}$) (Table 2.3). No edges have been eliminated using the DPI. Again, this threshold was chosen arbitrarily, however it was highly significant and retained approximately the same number of genes allowing comparisons to be made to Hillenmeyer's dataset.

Using said MI thresholds, I retained 2654 nodes (and 55675 edges) within Hillenmeyer's network and 2604 nodes (and 33593 edges) within Vulpe's network.

p-value	Corresponding MI
0.05	0.007972
0.01	0.011134
0.001	0.015658
1.00E-09	0.042802
1.00E-19	0.092564
1.00E-29	0.137803
1.00E-39	0.183042
1.00E-49	0.228281

Table 2. 2 ARACNE p-values and associated MI values for Hillenmeyer's fitness dataset.

P-value	Corresponding MI
0.05	0.0383674
0.01	0.0535852
0.001	0.075357
1.00E-05	0.118901
1.00E-10	0.22776
1.00E-15	0.336619
1.00E-20	0.445478
1.00E-25	0.554336
1.00E-30	0.663195

Table 2. 3 ARACNE p-values and associated MI values for Vulpe's fitness dataset.

2.2.5 Network analysis: Topology

To analyse network topology the Cytoscape plugin NetworkAnalyzer was used [112]. NetworkAnalyzer is a popular tool for network topology analysis and is able to compute the degree, radiality, clustering coefficient, and a variety of other parameters for each node within a network. It also computes edge parameters such as edge betweenness [112].

To determine if the top node hubs identified by NetworkAnalyzer were functionally enriched, I used GSEAPreranked. GSEAPreranked requires two files, a ranked list file and a gene matrix file. The ranked list consisted of genes ordered by their node connectivity as calculated by NetworkAnalyzer. The gene matrix file consisted of a curated list of all the gene ontology (GO) terms within *S. cerevisiae* together with the genes represented by each GO term. By using GSEApreranked it is possible to identify if specific cellular processes are signatures of highly connected nodes. GO terms containing over 800 genes and fewer than 10 genes were excluded from the analysis. GO terms with extraordinarily high number of genes are often very broad terms such as 'membrane' or 'cell surface', genes included within these terms are present within smaller more specific GO terms, therefore excluding GO terms with over 800 genes is not detrimental to the analysis. Results were collected after 1000 permutations. An FDR value of 10% was used to identify the most significant GO enriched in the ranked list. The threshold was chosen on the basis that it is the most likely to generate hypotheses and identify new directions of research [113].

2.2.6 Network Analysis: Modularisation

Networks were modularised using community cluster GLay [33] (Figure 2.1C) to identify modules. The GLay plug-in for Cytoscape allows clustering on the basis of solely connectivity allowing for the decomposition, display and exploration of large networks such as those used in this analysis [33]. GLay defined modules were then mapped onto the force directed layout parent network. Modules larger than 300 nodes underwent a further level of GLay clustering, termed sub-modules (Figure 2.1D). HOPACH was performed on each sub-module to determine if the fitness profile for nodes within, were all positively or negatively correlated across samples.

2.2.7 Ribosomal proteins first neighbour analysis

A list of all known RPs was obtained from the *Saccharomyces* Genome Database (SGD) [56]. RPs were classified into two groups based on their localisation within the cell. 245 were identified as cytosolic and 78 were identified as mitochondrial. Separation of ribosomal factors based on cellular compartmentalisation was required to prevent masking of significant correlations that may be dependent on cellular location

Each group was mapped onto the Hillenmeyer parent network. A first neighbour network was constructed by selecting all the first neighbours of the ribosomal factor group. Network modularisation was done using GLay [33], consistent with the modularisation methodology used on the parent network. Once again modules containing more than 300 nodes underwent a further level of GLay clustering. The Vulpe dataset was used for independent verification.

2.3 Results

2.3.1 A network biology approach identifies clusters of yeast mutant strains with similar phenotypic profiles

The application of the network inference technique ARACNE to the Hillenmeyer's dataset, inferred a network containing 2654 nodes and 55675 edges. In order understand the general properties of the network; I performed an analysis of the network topology using the Cytoscape plugin, NetworkAnalyzer [112]. To analyse the network topology, I constructed two networks based on node degree (Figure 2.3A) and radiality (Figure 2.3B). Networks were visualised using a force directed layout based on edge betweenness and node size was representative of node connectivity. Edge betweenness reflects the control that each node exerts over other nodes in the network. Hence, densely packed clusters of nodes are representative of a high level of regulation. Using a force directed layout led to the separation of nodes into three dense communities (Figure 2.3), suggesting that nodes within each community exert a high degree of control upon each other. Node degree is defined as the number connections to that node, ranging from 1 – 343 (Figure 2.3A). In a biological context the larger the node, the more likely it is to be a gene hub. Radiality is a node centrality index, a high radiality indicates that the node is closer to other nodes (Figure 2.3B); likewise a low radiality means the node is likely to be isolated. In a biological context, radiality can be used to infer the probability that a gene is functionally relevant to other genes, i.e. a gene with high radiality will exhibit regulation over other genes.

The top 50 most connected nodes were identified (see appendix Table A2.1). The full list of node connectivity can be found on the supplementary CD, in folder 'Chapter 2').

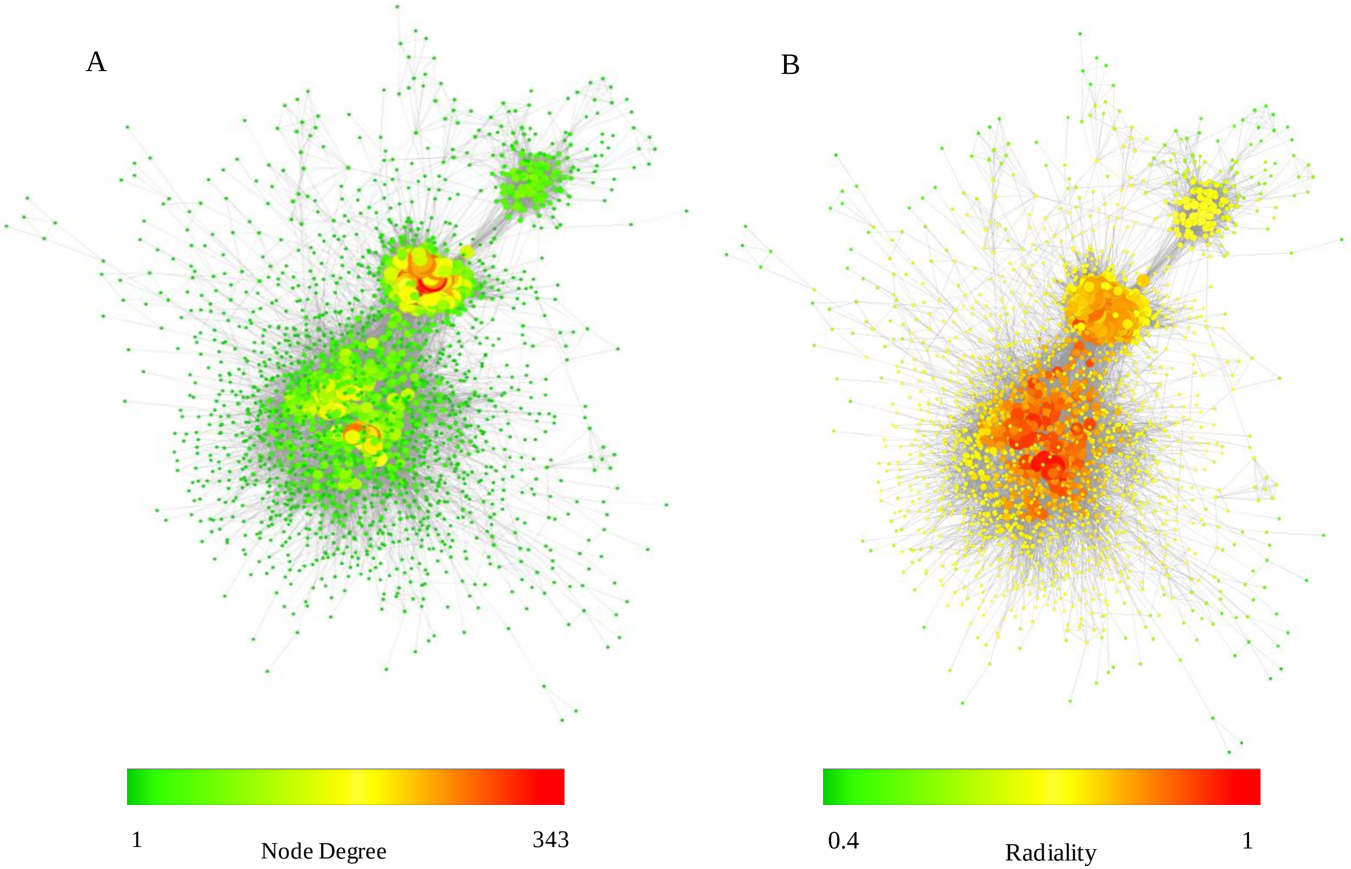


Figure 2. 3 Visualisation of the results of NetworkAnalyzer on the Hillenmeyer network.
Panel A. Node size and node colour is based on node connectivity. Those nodes with higher connectivity are likely to indicate gene hubs. Panel B. Node size is based on node connectivity, node colour is based on radiality of a node. Node radiality is the probability that a gene exhibits regulation over another gene.

To determine if the most connected hubs were functionally enriched, I used GSEAPreranked. Results showed there were no GO terms with a significant positive enrichment score (GO terms that show enrichment at the top of the ranked list). Interestingly, five GO terms were significantly enriched (FDR $\leq$ 0.2) towards the middle / bottom of the ranked list (Table 2.4). GO terms with statistically significant negative enrichment (representing nodes with connecting edges) were mainly proteasome based, suggesting that poorly connected nodes would be enriched in DNA translocase and proteasome functions (Table 2.4). The results did however suggest that functional enrichment may be found using a node parameter based on the betweenness centrality; therefore the GSEAPreranked analysis was repeated using the node parameter, radiality. Once again, GSEAPreranked did not identify any GO terms with a significant positive enrichment score, however 50 GO terms were identified as being significantly negatively enriched (FDR $\leq$ 0.1), the top 20 are shown in Table 2.5 (for the top 50, see appendix, Table A2.2). The results suggested that nodes ranked between positions 500 – 900 by their radiality score (middle of the ranked list), likely encoded ribosome related processes such as translation, transcription from RNA polymerase III (RNAPIII) promoters, proteasome components and ribosome subunits. This suggests that nodes represented as yellow in Figure 2.3B were likely to be involved in the cellular processes shown in Table 2.5.

NAME	DESCRIPTION	CNT	NES	FDR q-val	RANK
GO:0004298	threonine-type endopeptidase activity	11	-1.40	0.065	1281
GO:0015616	DNA translocase activity	10	-1.40	0.084	1165
GO:0010499	proteasomal ubiquitin-independent catabolic process	11	-1.35	0.092	1281
GO:0004175	endopeptidase activity	11	-1.42	0.108	1281
GO:0005839	proteasome core complex	12	-1.44	0.191	1281

Table 2. 4 The top five negative hits for node degree from GSEApreranked ordered by FDR CNT represents for the number of genes, NES stands for normalised enrichment score, FDR q-val is the estimated probability of a false positive and RANK is the location with ranked list.

NAME	Description	CNT	NES	FDR q-val	RANK
GO:0002181	cytoplasmic translation	92	-4.0	0.000	786
GO:0005839	proteasome core complex	12	-3.5	0.000	512
GO:0003735	structural constituent of ribosome	114	-3.5	0.000	846
GO:0030529	ribonucleoprotein complex	163	-3.5	0.000	846
GO:0004175	endopeptidase activity	11	-3.4	0.000	512
GO:0010499	proteasomal ubiquitin-independent protein catabolic ...	11	-3.4	0.000	512
GO:0004298	threonine-type endopeptidase activity	11	-3.4	0.000	512
GO:0000502	proteasome complex	31	-3.3	0.000	551
GO:0022627	cytosolic small ribosomal subunit	41	-3.2	0.000	812
GO:0001056	RNA polymerase III activity	14	-3.2	0.000	661
GO:0034515	proteasome storage granule	20	-3.2	0.000	551
GO:0042797	tRNA transcription from RNA polymerase III promoter	15	-3.2	0.000	869
GO:0022625	cytosolic large ribosomal subunit	50	-3.1	0.000	882
GO:0003899	DNA-directed RNA polymerase activity	23	-3.1	0.000	869
GO:0005666	DNA-directed RNA polymerase III complex	14	-3.1	0.000	661
GO:0000467	rRNA processing	11	-3.1	0.000	672
GO:0005622	intracellular	158	-3.0	0.000	892
GO:0006412	translation	152	-3.0	0.000	989
GO:0006364	rRNA processing	109	-2.9	0.000	855
GO:0005840	ribosome	151	-2.9	0.000	826

Table 2. 5 Results of GSEApreranked on radiality.

The significant negatively correlated GO terms are ordered by FDR value. Results show that yellow nodes in figure 2.3B represents genes that encode primarily proteasome, rRNA processing, translation and ribosome related functions. CNT represents for the number of genes contained with the GO term. NES stands for normalised enrichment score, and is a statistic used for examining gene enrichment results. FDR q-val is the estimated probability of a false positive and RANK is the location with ranked list.

2.3.2 Community analysis of the fitness network identifies highly interconnected modules

In order to assess the biological significance of the inferred networks, I first asked whether it was possible to subset the whole network into a number of distinct network modules. I addressed this question by applying a connectivity based community detection algorithm, GLay [33]. This procedure identified eight distinct modules (Table 2.6) with a very broad size range (11-1590). These localized in distinct areas of the force driven layout visualisation of the parent network (Figure 2.4). Table 2.6 shows the details of each GLay defined module.

Module	**Colour**	**No. of nodes**	**No. of Edges**	**No. of modules**	**Visualised in**
All		2654	55675	8	Fig. 2.4
1	Red	1590	19735	5	Fig. 2.5
2	Blue	541	31908	2	Fig. 2.6
3	Yellow	249	2504	5	Fig. 2.7
4	Purple	53	114	5	Fig. 2.8
5	Light Blue	28	41	1	Fig. 2.9
6	Orange	25	124	1	Fig. 2.10
7	Dark Green	11	11	1	Fig. 2.11
8	Light Green	11	10	1	Fig. 2.12

Table 2. 6 Breakdown of modules identified by GLay clustering in Hillenmeyer's fitness data.

2.3.3 The modular structure of the fitness network reflects functional compartmentalisation.

Having shown that the fitness network inferred by the ARACNE procedure possessed a modular structure, I set to assess whether this also reflects functional organization. This question was addressed by testing each module for functional enrichment (Figures 2.5 - 2.12), using the web based functional analysis tool DAVID [21]. Functional annotations were organised into three groups in order to classify the significance of gene enrichment. Those with a corrected FDR $\leq$ 0.05 are represented in red text, those with a corrected FDR $\leq$ 0.1 are represented in green text, and black text represents no significant enrichment (FDR > 0.1). This nomenclature is maintained throughout the thesis. A corrected FDR of $\leq$ 0.05 indicates that the probability of obtaining a false positive is less than or equal to 5%, which is a respectable threshold used in most scientific studies. In this case, however, it is important to note that even though a group of genes may be classified with an FDR of $\leq$ 0.05, many of these functional annotations in fact have FDR scores far less than 10^{-5}, corresponding to a significance level (the probability of a false positive) far below 5%. The raw DAVID files for each module, including the correct FDR scores can be viewed on the supplementary CD, folder ‘Chapter 2’.

In addition, though some modules contain no significant functional enrichment, inference of phenotypic links remain valid as only significant edges have been retained within the networks (as detailed in the methodology). 31% of sub-modules could be characterised by a specific functional profile, defined by a statistically significant (FDR $\leq$ 0.1). The HOPACH heatmaps for each sub-module (see appendix A2.1 – A2.8).

Module 1 (Figure 2.4, red nodes) was the largest, containing 1590 nodes and 19735 edges. Among the most enriched functions were transport, mitochondrial envelope, ribosome biogenesis and energy metabolism pathways (Figure 2.5). Module 2 (Figure 2.4, blue nodes) has a central position when mapped onto the parent network (Figure 2.4) and is enriched in important biological processes such as DNA replication and energy production (Figure 2.6). Module 2 co-localises with an area of the network that NetworkAnalzyer has identified as having the largest density of hubs (Figure 2.3A). Module 2 also contains a significant number of gene hubs (results of NetworkAnalyzer showed that 97 / 100 most connected nodes belong to module 2). Module 3 (Figure 2.6, yellow nodes), is significantly enriched in ribosomal proteins, ribosome biogenesis and proteasome genes (Figure 2.7). Module 4 (Figure 2.6, purple nodes) is enriched in cell cycle, carbohydrate regulation and ribosome biogenesis (Figure 2.8). Smaller modules, in particular modules 5 – 8 could not be subset with a further level of modularisation, they were however enriched in specific functions such as transcription factors (module 5, Figure 2.9), chromatin and chromosomal functions (module 6, Figure 2.10), nucleotide binding (module 7, Figure 2.11) and zinc binding (module 8, Figure 2.12). What follows is a detailed analysis of the functions that are represented in the three largest modules.

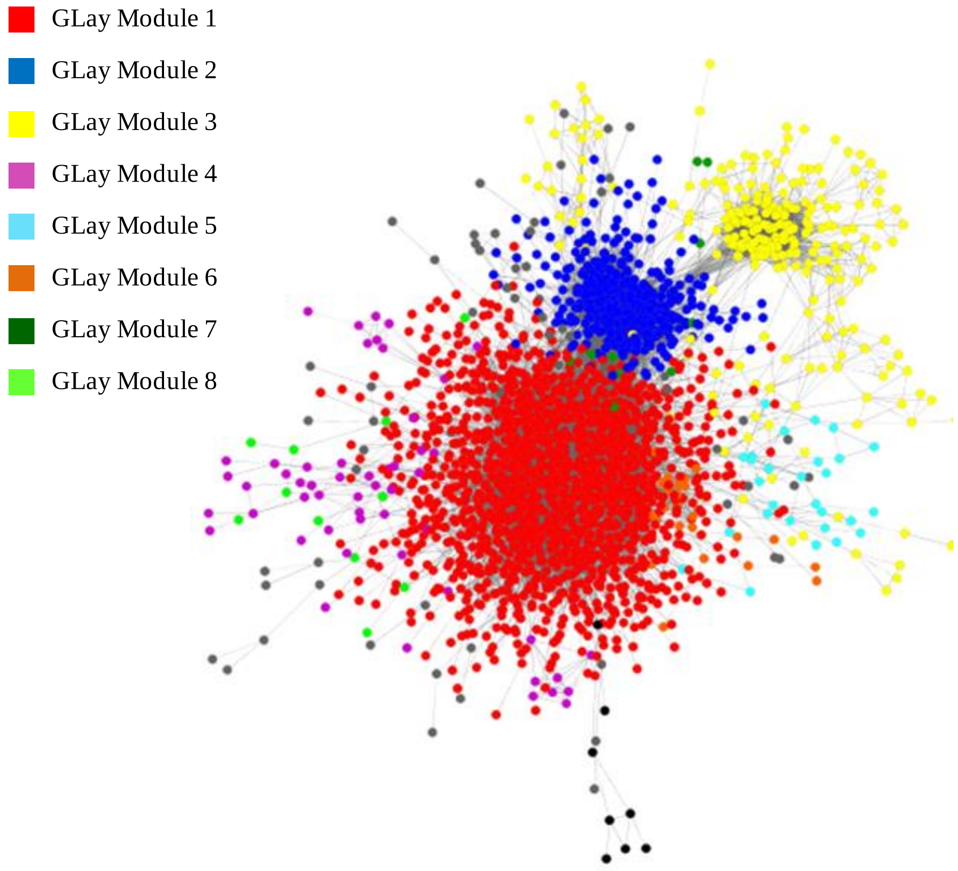


Figure 2. 4 Modules localise within distinct areas of the Hillenmeyer parent network. An undirected network showing the interactions between genes from the Hillenmeyer's *S. cerevisiae* fitness data at a 0.15MI threshold. The network is visualised using a force directed layout, modules defined by are GLay mapped onto the parent network. Node colour represents module (see legend). Edge length is representative of MI value. The accompanying table (Table 2.6) shows the breakdown of each GLay module including the colour, number of nodes, and number of edges within each module.

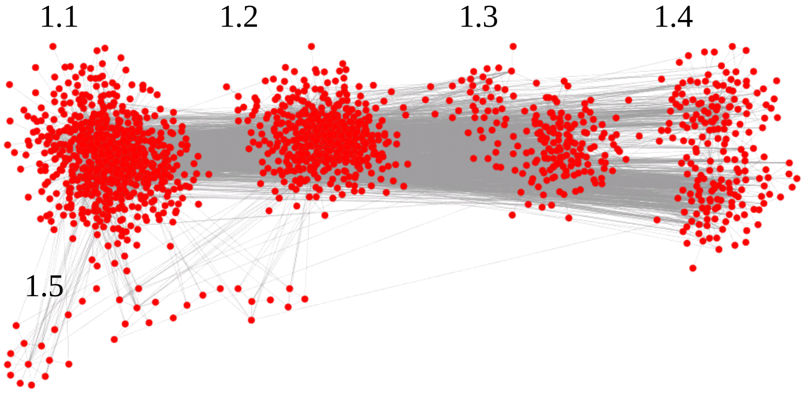
2.3.3.1 Hillenmeyer module 1: Mutations in ribosomal proteins and energy metabolism pathways produce highly correlated fitness profiles

Module 1 is the largest module identified by first level modularisation (1590 nodes and 19735 edges). A second level of modularisation identified five sub-modules (Figure 2.5). Sub-modules 1.1 and 1.2 underwent a third level of modularisation as they contained more than 300 nodes. Functional analysis of each sub-module revealed association between transport and mitochondrial envelope (sub-module 1.1), transport, cell cycle and ribosome biogenesis (sub-module 1.2), vitamin metabolism, hexose metabolism and ribosome biogenesis (sub-module 1.3), endoplasmic reticulum, golgi membrane (sub-module 1.4), and phosphate metabolic process and endoplasmic reticulum (sub-module 1.5).

Noteworthy is sub-module 1.1 which shows a significant enrichment (FDR $\leq$0.05) in mitochondrion envelope and transport related functions (Figure 2.5), indicative of mitochondrial import. Less significant enrichment includes six hexose metabolism genes and twenty ribosome related genes also located within sub-module 1.1. Inspection of the hexose metabolism proteins identified that they were all localised to cytoplasmic energy metabolism pathways and that a subset were directly correlated to RPs. Notably glyceraldehyde-3-phosphate dehydrogenase (*TDH1*), involved in glycolysis and gluconeogenesis, is directly connected to the 60S RP genes, *RPL17B*, *RPL43B* and ribosome biogenesis gene *TMA22*, consistent with the linkage between energy metabolism pathways and ribosomal proteins [92].

Sub-module 1.2 contains ten genes involved in cellular respiration including the succinate dehydrogenases *SDH1*, *SDH3*, *SDH4*. These genes are first neighbours of ribosome biogenesis genes indicating a strong correlation between energy production and ribosome synthesis, consistent with the reductive-charging and oxidative phase of the *S. cerevisiae* metacycle [94] [106]. Notably ribose-5-phosphate ketol-isomerase (*RKI1*), the enzyme responsible for catalysing the final step in riboneogenesis [92] is located within sub-module 1.3, together with genes involved in hexose metabolism and ribosome biogenesis.

Sub-modules 1.4 and 1.5 are both enriched in endoplasmic reticulum (ER) and membrane functions. Unlike other sub-modules, sub-module 1.4 exhibits an anti-correlated relationship between its nodes (Figure A2.1). Interestingly this anti-correlation is between endoplasmic reticulum / protein localisation and membrane proteins, suggesting that despite being located within the same sub-module; when exposed to the same stress, strains containing gene deletions encoding endoplasmic reticulum confer fitness and strains containing deletions in membrane proteins confer sensitivity and vice versa.



Module	Nodes	Edges	Functional analysis
1.1	684	7400	Transport (104), mitochondrial envelope (54), hexose metabolism (6), ribosome (20), mitosis (10), transcription regulation (20)
1.2	481	7434	Transport (31), cell cycle (31), cellular respiration (10), ribosome biogenesis (24), zinc finger, transcription (10)
1.3	199	988	Vitamin metabolism (7), hexose metabolism (5), ribosome biogenesis (12)
1.4	192	745	ER (23), membrane (58), golgi membrane (5) helicase (6), translation initiation (5), mannosyltransferase (6)
1.5	16	19	Phosphate metabolic process (3), ER (3), membrane (6)

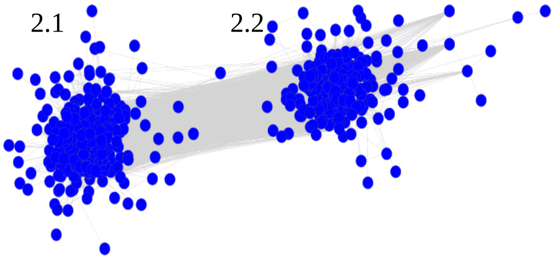
Figure 2. 5 Sub-modular structure of module 1, with accompanying functional analysis. Red text represents an FDR ≤ 0.05, green text represents an FDR ≤ 0.1, and black text represents non-significant enrichment.

2.3.3.2 Hillenmeyer module 2: Phenotypic linkage of mutant strains representing mitochondrial factors and energy metabolism.

Module 2 maps in the centre of the parent network, suggesting it is highly connected to all other modules. This hypothesis is supported by the fact that 97 / 100 most connected nodes are located within sub-network 2. A second level of modularisation identified two highly interconnected sub-modules (Figure 2.6). Functional analysis of each sub-module revealed association between cell cycle, DNA replication and energy metabolism processes (sub-module 2.1) and ribosome biogenesis and energy metabolism pathways (sub-module2.2). The grouping of cell cycle processes and DNA replication together is expected as DNA replication is co-ordinated by cell cycle stage [114].

Energy metabolism processes localised within the mitochondria are also enriched within sub-module 2.1 (electron transport, oxidation reduction, TCA cycle). First neighbour analysis of these genes showed that they were directly connected to nine mitochondrial and cytosolic RPs. The production of acetyl-coenzyme A and NADPH occurs during the reductive-charging phase of the metacycle and is required for ribosome biogenesis during the oxidative phase. [94] [106]. Consistent with this result is the enrichment of 21 cell division genes, cell division is reported to occur during the reductive-charging phase [94] [106]. The correlation between mitochondrional ribosomal factors and cytosolic ribosomal proteins is likely because mitochondrial proteins are synthesised as precursor proteins on cytosplasmic ribosomes before being imported into the intermembrane space [115]. In support of this, I also identified 21 genes functionally annotated as ‘protein transport’, including Translocase of the Inner Membrane 54 (TIM54), an essential component of the TIM22 complex which mediates the import of precursor proteins into the mitochondrial inner membrane. TIM proteins contain a highly conserved zinc finger motif [116] (ten are found within module 2.1). Furthermore the enrichment of four ubiquitin machinery genes within the same module as protein transport genes is consistent with reports that ubiquitination machinery acts a negative regulator in the synthesis and transport of proteins that localise within the mitochondrial inter membrane space [117].

Sub-module 2.2 is enriched in ribosome biogenesis and energy metabolism pathways that are localised within the cytoplasm (Figure 2.6). Seven genes are enriched as ‘alcohol metabolism’, which include predominantly glycolysis genes such as phosphofructokinase (*PFK1*), pyruvate kinase (*PYK2*), pyruvate decarboxylase (*PDC2*), alcohol dehydrogenase (*ADH1*) and the transaldolase *NQM1*. Of particular significance is NQM1, as it catalyses a portion of the non oxidative pentose phosphate pathway (PPP). Deletion of NQM1 has been reported to quadruple the concentration of sedoheptulose 7-phosphate (S7P), a metabolite that is essential for riboneogenesis [92]. The enrichment of 17 ribosomal protein / ribosome biogenesis genes is consistent with reports that flux through glycolysis is essential for providing intermediates for riboneogenesis pathway [92]. Six of the seven alcohol catabolism genes are directly connected to 14 ribosome related genes, suggesting a strong correlation between glycolysis and ribosome biogenesis.



Module	Nodes	Edges	Functional analysis
2.1	300	14404	Cell wall (20), DNA replication (14), cell cycle (42), ubiquitin ligase (5) , electron transport (5), ribosome (12), protein transport (19), oxidation reduction (17), TCA cycle (4)
2.2	241	17504	Alcohol metabolism / dehydrogenase (7), -ve regulation of gluconeogenesis (3), ribosome biogenesis (17), DNA metabolism (6), oxidative phosphorylation (4), zinc binding (17)

Figure 2. 6 Sub-modular structure of module 2, with accompanying functional analysis. Text colour is representative of significance, red: FDR $\leq$ 0.05; green: FDR $\leq$ 0.1; black: FDR > 0.1.

2.3.3.3 Hillenmeyer module 3 – Phenotypic linkage of mutant strains representing RPs and ribosome biogenesis

Module 3 formed five smaller interconnected sub-modules after a second level of modularisation (Figure 2.7). Functional analysis revealed strong phenotypic correlations between yeast strains mutated in ribosome biogenesis and small ribosomal subunit genes (sub-module 3.1), large ribosomal subunit and RNA polymerase III (RNAPIII) (sub-module 3.2), RNAPII and chaperones (sub-module 3.3), proteasome and sexual sporulation (sub-module 3.4) and protein transport and ATP binding (sub-module 3.5). The significant enrichment of ribosomal related functions across all modules is expected, as ribosome biogenesis is a highly coordinated process [97].

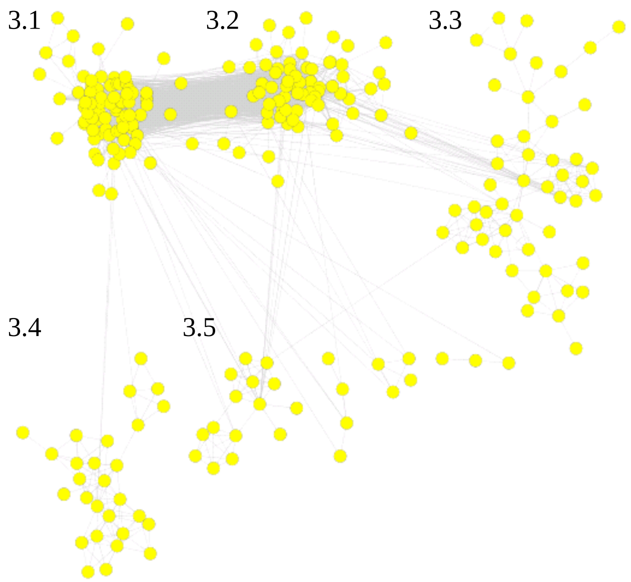
Sub-module 3.1 captures the process of ribosomal biogenesis and small subunit synthesis. Ribosome biogenesis is a multistep process, beginning with the transcription of two RNA polymerases, RNAPI and RNAPIII. RNAPI synthesises the 35S-rRNA primary transcript and RNAPIII synthesises the pre-5S rRNA transcript [118], genes encoding both RNAPs were found in module 3.1. The precursor 35S rRNA is then processed to yield mature 25S, 18S and 5.8S rRNAs [118], consistent with the16 genes involved in the maturation of 5.8S rRNA (FDR: $8x10^{-14}$).

The most significant enrichment in sub-module 3.2 was cytosolic large ribosomal subunit, translation regulation and ribosome export. The significant enrichment of seven genes functionally annotated as RNAPIII complex (FDR: $3.84x10^{-7}$) and 11 genes annotated as 'preribosome, large subunit precursor (FDR: $6.14x10^{-14}$)', including RPL5, suggested that sub-module 3.2 represents the synthesis of the large ribosomal subunit. Sub-module 3.1 is enriched in only small RP genes; sub-module 3.2 is enriched in only large RP genes, suggesting that although the process of ribosome biogenesis is highly coordinated, small and large RPs achieve a further level of coordination between themselves.

There were only two highly significant functional groups within module 3.3. The first was 12 genes related to transcription from an RNAPII promoter and six genes belonging to chaperonin t-complex polypeptide 1 (TCP-1 / CCT) consistent with results from Hillenmeyer *et al* [40]. The six genes in the chaperonin group were all subunits of TCP-1, which assists in the folding of a distinct subset of cellular proteins *in vivo* [119], specifically actin and tubulin [119] and newly translated myosin II heavy chains. [120]. These results suggested that module 3.3 captured the association of TCP-1 to the ribosome and its role in folding of cytoskeleton and chromatin remodelling associated genes, possibly in preparation for the onset of M phase.

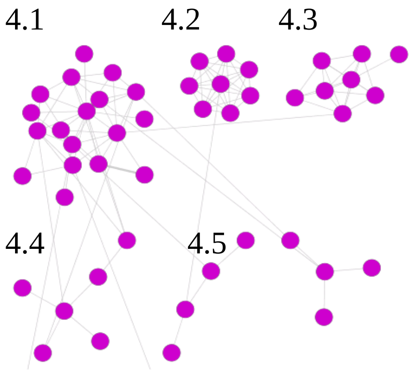
Though module 3.4 contains only 28 nodes; 22 of them are functional annotated as proteasome, consistent with analysis previously done on the same dataset [40]. The yeast proteasome is required for the turnover of proteins and the removal of misfolded proteins. The proteasome is composed of numerous subunits including PRE and PUP components [121], as well as RPN regulatory components, consistent with these reports are the localisation of *PRE2*. *PRE3, PRE4, PRE5, PRE6, PUP1, PUP2, PUP3* and as well the regulatory components *RPN3, RPN6, RPN8, RPN9* and *RPN12* within sub-module 3.4. The presence of this sub-module together with those involved in ribosome biogenesis (sub-modules 3.1 – 3.2) suggested a highly correlated and tightly coupled process in which synthesised malformed proteins translated on the ribosome are rapidly destroyed by the proteasome machinery.

A subset of genes encoding RPs overlap spatially with module 5 in the parent network (Figure 2.4). Interestingly sub-network 5 is significantly enriched in transcription factor TFIID complex (Figure 2.7). TFIID is also known as the TATA binding protein (TBP) [122] and is a transacting factor required by RNAPI, II and III. Reports have shown that RNAPI, II and III require TBP in order recognise specific promoters [123] [124] [125].



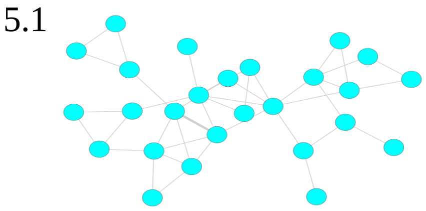
Module	Nodes	Edges	Functional analysis
3.1	78	977	rRNA processing / Ribosome biogenesis (44), small ribosome subunit (26), rRNA related processes (24), translation regulation (11), ribonucleoprotein complex assembly (5), RNA polymerase I (3), RNA polymerase III (3)
3.2	68	630	Cytosolic large ribosomal subunit (26), translation regulation (13), RNA polymerase III (7), ribosome export (6)
3.3	49	124	RNA polymerase II (12), chaperonin / TCP-1 (6)
3.4	28	86	Proteosome (22), threonine protease (11) sexual sporulation (8),
3.5	15	39	Protein / nucleocytoplasmic transport (4), protein complex (3)

Figure 2. 7 Sub-modular structure of module 3, with accompanying functional analysis. Text colour is representative of significance, red: FDR $\leq$ 0.05; green: FDR $\leq$ 0.1; black: FDR > 0.1.



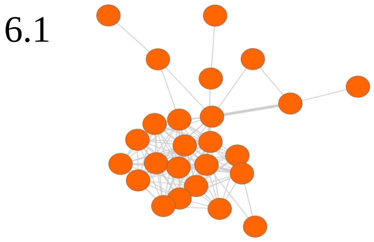
Module	Nodes	Edges	Functional analysis
4.1	18	40	regulation of carbohydrate (3), ER (3)
4.2	8	27	Kinetochore (2)
4.3	8	21	cell cycle phase (4)
4.4	6	5	Organelle lumen (3)
4.5	4	3	Anion transport (1), coenzyme biosynthesis (1), ubiquitin (1)

Figure 2. 8 Sub-modular structure of module 4, with accompanying functional analysis.
Text colour is representative of significance, red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR > 0.1.



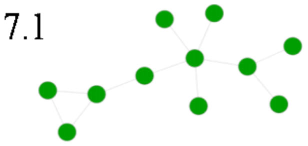
Module	Nodes	Edges	Functional analysis
5.1	26	41	transcription factor TFIID complex (4), glycoprotein (4)

Figure 2. 9 Structure of module 5, with accompanying functional analysis.
Text colour is representative of significance, red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR > 0.1.



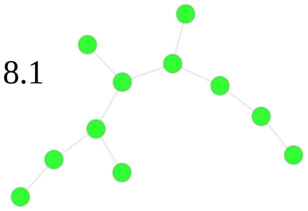
Module	Nodes	Edges	Functional analysis
6.1	25	124	chromatin assembly (3), membrane (7), chromosomal (3)

Figure 2. 10 Structure of module 6, with accompanying functional analysis.
Text colour is representative of significance, red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR > 0.1.



Module	Nodes	Edges	Functional analysis
7.1	11	11	ATPase activity-coupled (5), organelle lumen (4),

Figure 2. 11 Structure of module 7, with accompanying functional analysis.
Text colour is representative of significance, red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR > 0.1.



Module	Nodes	Edges	Functional analysis
8.1	11	10	Zinc (3), ATP binding (3)

Figure 2. 12 Structure of module 8, with accompanying functional analysis.
Text colour is representative of significance, red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR > 0.1.

2.3.4 An independent fitness analysis confirms the functional compartmentalisation of the Hillenmeyer fitness network

The results gathered in this analysis identified that genes can be compartmentalised based on their fitness contribution to cell viability. Interestingly, the modules identified within this study contain genes from multiple biological functions suggesting a strong correlation between different biological pathways such as hexose metabolism and ribosome biogenesis identified in sub-module 1.3 (Figure 2.5). A key question however, is whether within each module, there is a true underlying biological connection between genes from different biological processes or whether the connection is simply a consequence of contributing a similar fitness to *S. cerevisiae* cells.

Therefore, in order to garner further evidence to validate these findings I applied the same analysis pipeline (Figure 2.1) to a fitness dataset developed by a collaborator, Prof. Chris Vulpe (Berkeley University, USA). This procedure identified eight modules, and like the Hillenmeyer network, the sizes of these modules varied greatly (Table 2.7). These localised to distinct areas of parent network (Figure 2.13). Modules underwent a second level of modularisation; each sub-module was annotated using DAVID. For Vulpe's network, I could prove that 36% of sub-modules were characterised by a statistically significant (FDR<10%).

Once again module 1 (Figure 2.13, red nodes) was the largest, containing 912 nodes and 6892 edges. Among the most enriched functions were peroxisome and protein-tyrosine phosphatase (Figure 2.14). Module 2 (Figure 2.13, yellow nodes) was enriched in mitochondrial and cytosolic RPs (Figure 2.15). Module 3 maps to the centre of the parent network and contains a diverse set of functions including endocytosis and energy metabolism processes (Figure 2.16). Module 4 (Figure 2.13, purple nodes) was enriched in cell division, organelle membrane and stress response (Figure 2.17). The enrichment of cell division and organelles suggests a functional overlap with Hillenmeyer module 4. Module 5 (Figure 2.13, light blue nodes) was enriched in amino acid biosynthesis and mitochondrial membrane (Figure 2.18). Module 6 is enriched in transcription regulation and chromosome organisation (Figure 2.19). Modules 7 and 8 were enriched in genes encoding membrane related proteins (Figures 2.20 and 2.21 respectively). The degree of overlap between Hillenmeyer and Vulpe sub-networks is shown in table 2.8 (the raw DAVID files for each overlap can be found on the supplementary CD – Chapter 2). What follows is a detailed analysis of the functions that are represented in the three largest sub-networks. The overlap between modules between the modules within Hillenmeyer's fitness network and Vulpe's fitness network, including the number of genes that overlap and functional enrichment of these overlapping genes are shown in Table 2.8.

Module	Colour	Number of nodes	Number of Edges	Number of modules	Visualised in
All		2604	33593	8	Fig. 2.13
1	Red	912	6892	4	Fig. 2.14
2	Yellow	660	3879	6	Fig. 2.15
3	Blue	653	17938	3	Fig. 2.16
4	Purple	117	179	8	Fig. 2.17
5	Light Blue	37	41	1	Fig. 2.18
6	Orange	20	20	1	Fig. 2.19
7	Dark Green	15	16	1	Fig. 2.20
8	Light Green	13	16	1	Fig. 2.21

Table 2. 7 Breakdown of modules identified by GLay clustering in Vulpe's fitness data

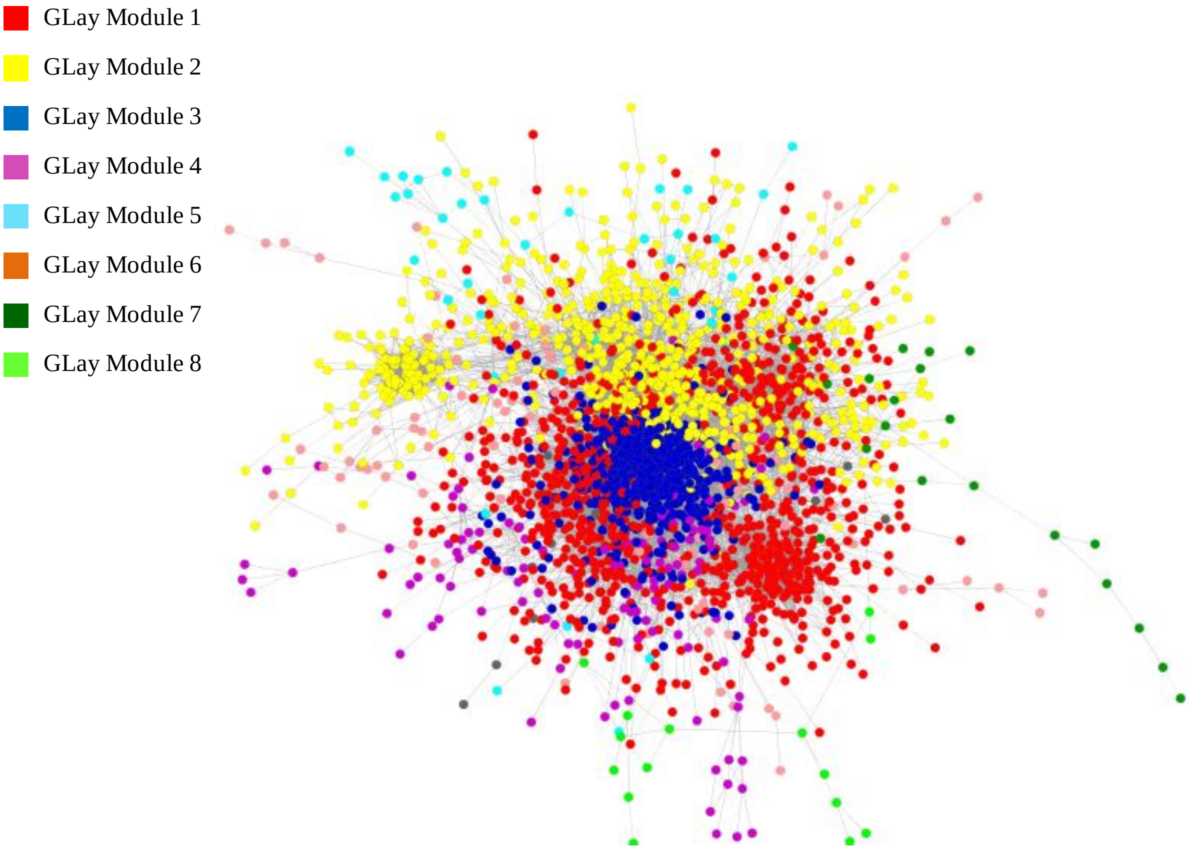
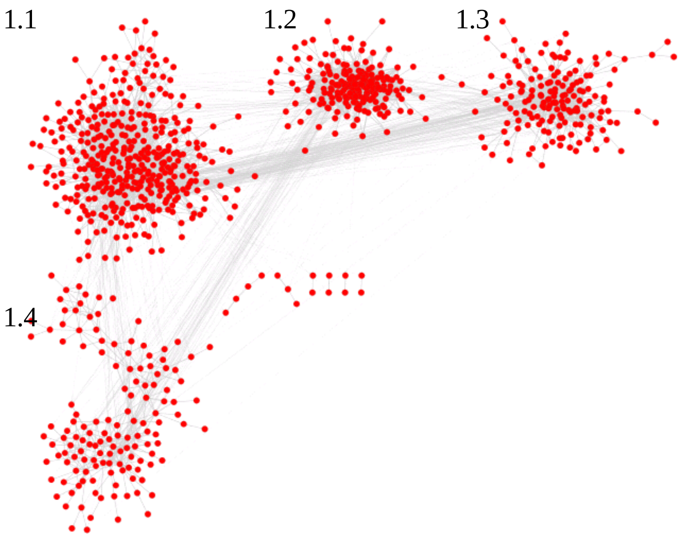


Figure 2. 13 Modules localise within distinct areas of the Vulpe parent network.
An undirected network showing the interactions between genes from the *S. cerevisiae* Vulpe fitness data at a 0.35MI threshold. Force directed layout, with GLay defined modules mapped onto the parent network. Node colour represents module (see legend). Edge length represents MI value. Table 2.7 shows the breakdown of each GLay module including the colour, number of nodes, and number of edges.

2.3.4.1 Vulpe module 1: The phenotypic linkage between ribosomal biogenesis and cytoplasmic energy metabolism pathways is conserved in an independent dataset

Module 1 formed four interconnected sub-modules after a second round of modularisation (Figure 2.14). The functional analysis of these sub-modules revealed association between peroxisome, cell cycle and ribosome biogenesis (sub-module 1.1). The similar fitness profile of genes involved in peroxisome is expected, as strains within this module would be unable to metabolise hydrogen peroxide, leading to cell death [40]. Sub-module 1.2 revealed functional association between RPs and cytoplasmic energy metabolism pathways. Closer inspection of the six glucose metabolism genes showed that dehydrogenase (*MDH1*) and transaldolase *TAL1* both had direct connections to multiple ribosomal proteins (specifically *RPL27A*, *RPL33B*, *MRPL39* and *YMR114C*). *TAL1* was particularly significant due to its role in the non oxidative arm of the PPP. TAL1 is involved in the catalysis of the substrate sedoheptulose-7-phosphate [126]. As described previously, sedoheptulose-7-phosphate is a key metabolite required for riboneogenesis [92]. In Hillenmeyer's module 1, I identified that the transaldolase *NQM1* had direct connections to numerous RPs. Like *NQM1*, deletion of *TAL1* has been reported to quadruple the concentration of sedoheptulose-7-phosphate [92].

Notably, sub-module 1.4 was significantly enriched in protein tyrosine phosphatase SIW14-like. The four genes were oxidant-induced cell cycle arrest (OCA), *OCA1*, *OCA4*, *OCA6* and *SIW14*. *SIW14* involved in protein metabolism and post-translational modification [127]. Both SIW14 and the OCA proteins are required for cell viability upon exposure to redox stresses. Grouping of these genes together was unsurprising, as strains deficient in these genes would be unable to survive in response to redox stress, hence their similar fitness profiles.



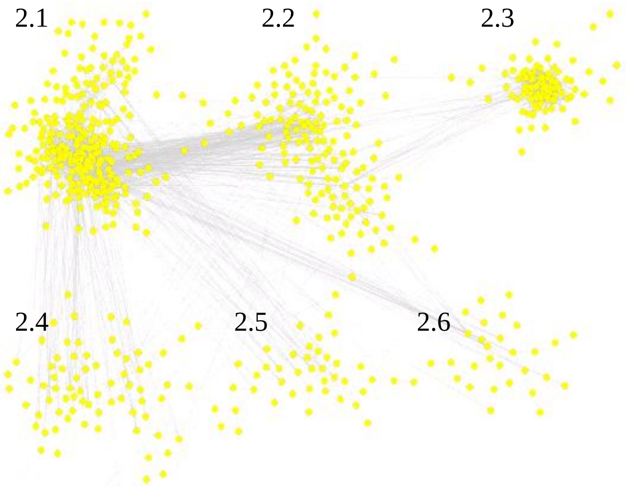
Module	Nodes	Edges	DAVID
1.1	392	2355	Peroxisome (7), cell polarity (7), cyclin (3), oxidoreductase (14), ion transport (8), heat response (17), kinase (15), sporulation (10), cell cycle (22), ribosome / ribosome biogenesis (17), mitochondrion (41), hexose metabolism (8), translation (16)
1.2	204	3014	Antiporter activity (5), membrane (56), glucose metabolism (5), electron carrier activity (5), ribosome (9) DNA repair (6), telomere (3), cytokinesis (4)
1.3	162	780	Kinase (9), NAD biosynthesis (3), protein transport (12), generation of energy (7), cell cycle (7), ATPase (5), temperature response (10)
1.4	139	304	Protein-tyrosine phosphatase SIW14-like (4), membrane (18), zinc finger (10), electron transport (6), ER (10), cell homeostasis (6), ribosome (8)

Figure 2. 14 Sub-modular structure of Vulpe module 1 with functional analysis.
Text colour is representative of significance, red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR > 0.1.

2.3.4.2 Vulpe module 2: Cytosolic and mitochondrial ribosomal proteins

Module 2 forms six modules upon a further level of modularisation. (Figure 2.15), and showed a significant overlap with modules 1 and 3 of Hillenmeyer's data (Table 2.8). Sub-module 2.1 was enriched in genes encoding cell cycle checkpoint, sulphur biogenesis, ribosome and aerobic respiration. Further investigation into the six hexose metabolism genes identified *TKL1* and *RPE1*, key enzymes required in riboneogenesis and *REG1*, another hexose metabolism gene is involved in controlling glucose repression [128].

Interestingly, much like module 3 in Hillenmeyer's network, small and large ribosomal proteins were enriched within their own sub-modules (2.2 and 2.6 respectively), hence the highly significant overlap of 25 genes related to ribosomal functions between Hillenmeyer module 3 and module 2 (Table 2.8) (FDR 4.75x10^{-28}) . Sub-module 2.3 is significantly enriched in mitochondrial ribosome and mitochondrial translation, a feature captured by module 1 in Hillenmeyers data (Table 2.8).

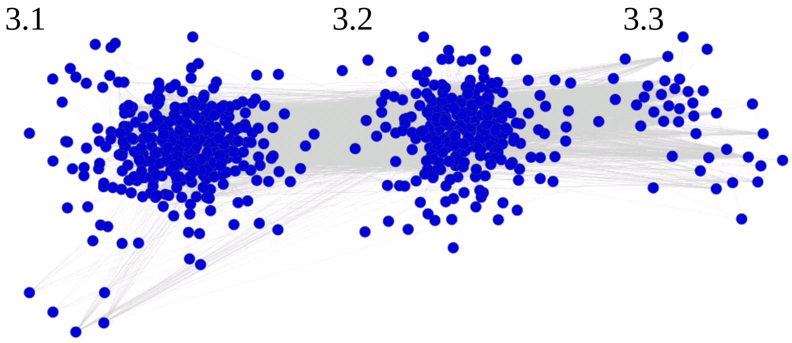


Module	Nodes	Edges	DAVID
2.1	256	1711	Sulfur biosynthesis (10), checkpoint (5), aldehyde dehydrogenase (3), mitochondrial carrier / membrane (4/11), electron transport 4), cytosolic ribosome (11), hexose metabolism (6), RNAPII promoter (5), generation of precursor metabolites and energy (11), aerobic respiration (3)
2.2	148	391	Small ribosome subunit (17) , translation (19), rRNA processes (12), elongator holoenzyme (3), regulation of transcription for RNAPII promoter (3)
2.3	94	1011	Mitochondrial (65), mitochondrion translation / ribosome (33 / 26), AA activation (9), rRNA binding (4), AP-3 adapter (3)
2.4	72	123	RNA polymerase I (3), cell redox homeostatis (3), cell growth (4) energy generation (5), protein transport (4), oxidation reduction (7)
2.5	42	69	Membrane (14), ABC transporter (3), cell cycle (5), response to heat (3), macromolecule synthesis (4), RNAPII (3)
2.6	29	44	Large ribosome subunit (6), ion transport (3), histone modification (3), peroxisome (3), -ve regulation of nucleotide metabolism (3)

Figure 2. 15 Sub-modular structure of Vulpe module 2 with functional analysis.
Text colour is representative of significance, red: FDR $\leq$ 0.05; green: FDR $\leq$ 0.1; black: FDR > 0.1.

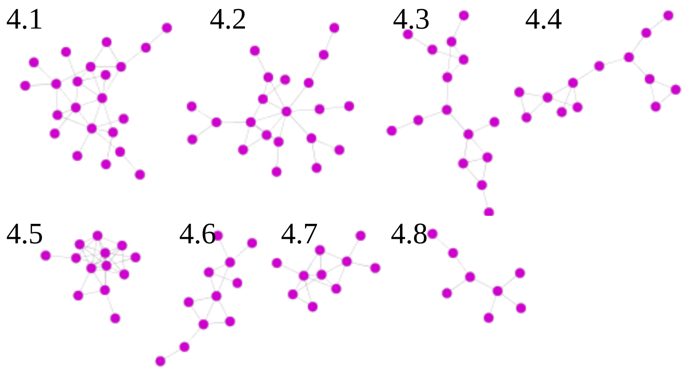
2.3.4.3 Vulpe module 3: The phenotypic linkage of mutant strains representing energy metabolism and cell cycle

Module 3 is defined by three smaller interconnected sub-modules (Figure 2.16). Endocytosis is the only significantly enriched function (FDR: 8.35x10^{-05}), located within sub-module 3.1, which also showed functional association to TCA cycle and mitochondrion. Sub-module 3.2 demonstrated a strong phenotypic link between cell cycle, transport and energy metabolism. This shows remarkable similarity with module 2 from Hillenmeyer's fitness network in both functional enrichment and localisation within the parent network. Both modules are located within the centre of their respective parent networks and both are enriched in a broad range of energy metabolism processes and cell cycle functions.



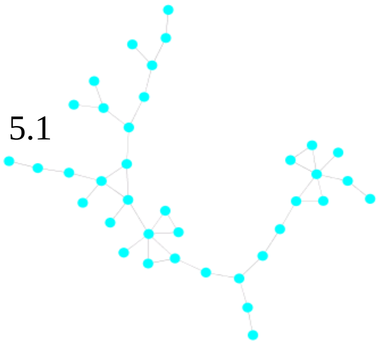
Module	Nodes	Edges	DAVID
3.1	329	5751	Endocytosis (7), TCA cycle (3), tRNA processing (3), cell cycle (23), vacuole (9), vesicle (8), mitochondrion (31), cell wall (10), DNA repair (11), pentose phosphate pathway (3), chromatin organisation (12)
3.2	279	6279	Membrane (74), oxidoreduction (23), Ehrlich pathway (4), starch / glucose metabolism (4), aerobic respiration (5)
3.3	40	142	Transcription regulation (4) ,cell wall (3), macromolecular catabolism (4)

Figure 2. 16 Sub-modular structure of Vulpe module 3 with functional analysis
Text colour is representative of significance, red: FDR < 0.05; green: FDR < 0.1; black: FDR > 0.1.



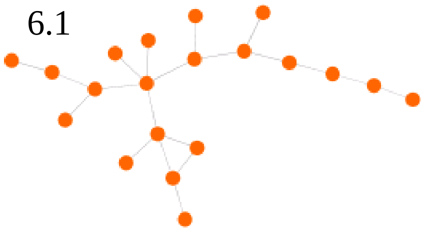
Module	Nodes	Edges	DAVID
4.1	22	31	Cell growth, macromolecular complex organisation (6), cell growth (3), DNA metabolism (4), cell cycle (5), ATP binding (4), membrane (5)
4.2	21	25	endoplasmic reticulum (5), glycoprotein (5), non-membrane-bounded organelle (4)
4.3	15	17	Organelle envelope (5)
4.4	13	15	Ribosome biogenesis (4), organelle membrane (4), mitochondrion (3)
4.5	13	30	Golgi transport complex (4), response to starvation (3), intracellular protein transport (3)
4.6	11	13	Mitochondrion (4)
4.7	10	15	Cell wall biosynthesis (4), sexual reproduction (4)
4.8	8	7	Integral to membrane (3)

Figure 2. 17 Sub-modular structure of Vulpe module 4 with functional analysis.
Text colour is representative of significance, red: FDR $\leq$ 0.05; green: FDR $\leq$ 0.1; black: FDR > 0.1.



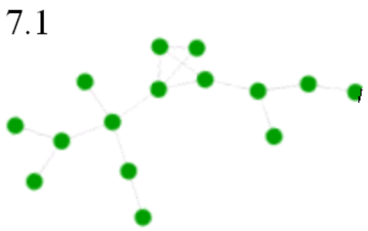
Module	Nodes	Edges	DAVID
5.1	37	41	cellular amino acid biosynthetic process (3), Vacuole (3), membrane organization (4), mitochondrial membrane (4)

Figure 2. 18 Structure of Vulpe module 5 with accompanying functional analysis.
Text colour is representative of significance, red: FDR $\leq$ 0.05; green: FDR $\leq$ 0.1; black: FDR > 0.1.



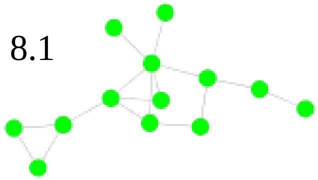
Module	Nodes	Edges	DAVID
6.1	20	20	transcription regulator (6), chromosome (3), zinc ion binding (3)

Figure 2. 19 Structure of Vulpe module 6 with accompanying functional analysis.
Text colour is representative of significance, red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR > 0.1.



Module	Nodes	Edges	DAVID
7.1	15	16	Membrane (7)

Figure 2. 20 Structure of Vulpe module 7 with accompanying functional analysis.
Text colour is representative of significance, red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR > 0.1.



Module	Nodes	Edges	DAVID
8.1	13	16	membrane (7), metal ion binding (3)

Figure 2. 21 Structure of Vulpe sub-network 8 with accompanying functional analysis.
Text colour is representative of significance, red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR > 0.1.

	H_M1	H_M2	H_M3	H_M4	H_M5	H_M6	H_M7	H_M8
V_M1	279 membrane (80) cofactor transporter activity (5)	82 Endoplasmic Reticulum (3) Ubiquitin (4)	8 Cytosolic ribosome (6)	5 Transmembrane (4)	4 Drug response (2)	4 Uncharacterised membrane (4)	3 Nucleus (3)	1 Response to temperature (1)
	17.6% 30.6%	15.2% 9%	3.2% 0.9%	9.4% 0.5%	15.3% 0.4%	16% 0.4%	27.2% 0.3%	9.1% 0.1%
V_M2	153 Mitochondrion (39) Mitochondrion envelope (16)	68 Nuclear chromosome (7) Lipid Synthesis (3)	25 Cytosolic ribosome (23) Regulation of translation (10)	7 Transmembrane (3)	2 Golgi (1) Uncharacterised (1)	2 Uncharacterised membrane (2)	2 Mitochondrion (2)	0
	9.6% 23.2%	12.4% 10.2%	14.1% 5.3%	13.2% 1.1%	7.7% 0.3%	8% 0.3%	18.2% 0.3%	0% 0%
V_M3	204 Rhodenase (5) Histone binding (3)	56 Lipoprotein (4) Cell wall (4)	10 ATP binding(4)	5 Protein transport (2) Aerobic respiration (1)	2 Protein folding (1)	5 Uncharacterised (3)	1 Chromatin organisation (1)	1 Recombinational repair
	12.8% 31.2%	10.4% 8.6%	4% 1.6%	9.4% 0.8%	7.7% 0.3%	20% 0.8%	9.1% 0.6%	9.1% 0.15%
V_M4	29 Cell wall (5) Sexual reproduction(4)	12 Protein transport (3) Endoplasmic reticulum (3)	1 Mitochondrion (1)	1 M-phase (1)	1	0	0	0
	1.8% 24.8%	2.2% 10.3%	0.4% 0.9%	1.9% 0.9%	3.9% 0.9%	0% 0%	0% 0%	0% 0%
V_M5	9 Membrane fusion (6)	3 Nucleotide metabolism (1)	1 Oxidative phosphorylation (1)	0	0	0	0	0

			Proteolysis (1)													
	0.6%	24.3%	0.6%	8.1%	0.4%	2.7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
V_M6	4 Transcription (3)		1 Uncharacterised (1)		0		0		0		0		0		0	
	0.3%	20%	0.2%	5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
V_M7	7 Proteome (2)		0		0		0		0		0		0		0	
	0.4%	46.7%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
V_M8	5 Translation (2) Amino acid metabolism (2)		0		0		0		0		0		0		0	
	0.3%	38.5%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 2. 8 The degree of overlap between modules from Hillenmeyer's network (green) and Vulpe's network (red). Columns represent each of the eight network modules identified in Hillenmeyer's fitness network (green). Rows represent the eight network modules identified in Vulpe's independent dataset (red). Each large cell is split into 3 sections, light blue, green and red. The first section (in light blue) shows the total number of genes which overlap between the two modules, the two most statistically significant functional enrichment from DAVID and the number of genes in each functional category. Text colour represents the corrected FDR, with red $\leq$ 0.05, green $\leq$ 0.1, and black representing no significant enrichment. The % shown in green and red sections beneath each blue section represents the % which the overlap occupies for Hillenmeyer's module (green) and Vulpe's module (red). For example, investigation into the overlap between Hillenmeyer module 3 (column 3) and Vulpe module 2 (column 2) shows that 25 genes overlap, of which 23 are significantly enriched (FDR $\leq$ 0.05) in cytosolic ribosome and translation regulation. The 25 gene overlap occupied a total of 14.1% of Hillenmeyer's module 3 and 5.3% of Vulpe's module 2.

2.3.5 The construction of RP first neighbour networks

The analysis on *S. cerevisiae* fitness data suggested a strong phenotypic linkage between RPs and a diverse array of cellular functions including cell cycle and energy metabolism. This section of results focuses on the first neighbours of ribosomal factors. Ribosomal factors were separated based on their localisation within the cell, either cytosolic or mitochondrial. That data demonstrated that key genes encoding enzymes involved in glycolysis and riboneogenesis are predominantly linked cytosolic RPs. The data also demonstrated that mitochondrial RPs are phenotypically linked to cytosolic RPs despite being compartmentally separated. Finally, I report that cell cycle checkpoint protein BUB1 exhibits a significant co-fitness profile to over 45 cytosolic RPs, and that the correlation is strongest with small RPs only.

2.3.5.1 Mutations in genes encoding cytosolic RPs and energy metabolism enzymes result in highly correlated fitness profiles

Module 1 (Figure 2.22, red nodes) forms three interconnected sub-modules (Figure 2.22). Functional analysis revealed association between homeostatis, transport, and numerous metabolic processes (sub-module 1.1), protein transport, mitochondrial electron transport and hexose metabolism (sub-module 1.2), reproductive process and mitochondrial inner membrane (sub-module 1.3). Module 2 (represented in blue) formed three sub-modules, functional analysis revealed an association between ribosome biogenesis, rRNA maturation and processosome (sub-module 2.1) and cytosolic large subunit and ribosome biogenesis (sub-module 2.2). Module 2 also contained the most edges of any other modules within this network (Table 2.9), representative of the highly co-ordinated process of ribosome synthesis. Module 3 splits into two interconnected sub-modules, functional analysis revealed association between chromatid segregation, spindle pole duplication and

DNA replication (sub-module 3.1), characteristic of the reductive-building phase of the yeast metacycle [94]. Sub-module 3.2 is enriched in metabolic and catabolic processes, as well as ribosomal biogenesis.

Sub-module 1.1 (Figure 2.22) was enriched in five hexose metabolism genes, including transketolase (*TKL1*), glyceraldehyde-3-phosphate dehydrogenase (*TDH1*), triose phosphate isomerase (*TPI1*), *THI3*, involved in thiamine metabolism and *LAT1*, the E2 component of pyruvate dehydrogenase. *THI3* is phenotypically correlated to 41 genes, only two are RPs, (*RPL17B* and *RPL43B*), it is also correlated to mitochondrial RP, *MRPL13*. *TKL1* is phenotypically correlated to a total of 54 genes, of which three are involved in ribosome biogenesis (*LEA1*, *HSH155*, *YBL028C*) and a single gene encoding a cytosolic RP (*RPS9A*). TKL1 is reported to be the primary source of transketolase activity in *S. cerevisiae*, and is co-expressed with sedoheptulose-1, 7-bisphosphatase (*SHB17*) [92]. TKL1 catalyses the entry of glycolytic intermediates fructose-6-phosphate & glyceraldehyde-3-phosphate into the riboneogenesis pathway. Double mutants of both transketolases (*TKL1*, *TKL2*) in *S. cerevisiae* leads to zero flux of metabolites through SHB17, essentially rendering riboneogenesis obsolete [92]. Finally TPI catalyses the conversion dihydroxyacetone-phosphate to glyceraldehyde-3-phosphate, a glycolytic intermediate utilised within riboneogenesis [92].

Sub-module 3.2 (Figure 2.22) is enriched in metabolic and catabolic processes, as well as ribosome biogenesis. Noteworthy was that of the six genes involved in glucose catabolism, five were specific to the glycolytic pathway (*PDC2*, *PFK1*, *ADH1*, *PYK2*, and the hexokinase *YLR446W*). These five genes had 290 first neighbours within module 3 Functional analysis showed that of the 290 genes, 20 were annotated as ribosome biogenesis, and four as RPs (2 cytosolic RPs, 2 mitochondrial RPs).

Module	Nodes	Edges
1.1	222	3644
1.2	178	4059
1.3	35	281
2.1	211	12765
2.2	178	8595
3.1	72	941
3.2	72	733

Table 2. 9 Breakdown of cytosolic ribosomes first neighbours modules for Hillenmeyer's network.

To further investigate the link between RPs and energy metabolism pathways, I identified all the edges between RPs and energy metabolism processes and ordered them by MI (Table 2.10). For each pair of genes a Pearson correlation coefficient was calculated, correlation plots were constructed across all experimental samples from Hillenmeyer's dataset (see supplementary CD, folder 'Chapter 2'). I identified that the highest scoring RP - energy metabolism pairs are often contain the same ribosomal factor, such as ribosomal biogenesis gene *BMS1*, and RPs *RPL22B* and *RPL31B*.

🟥 Cytosolic RP GLay Module 1

🟦 Cytosolic RP GLay Module 2

🟩 Cytosolic RP GLay Module 3

🟨 Cytosolic Ribosomal Protein

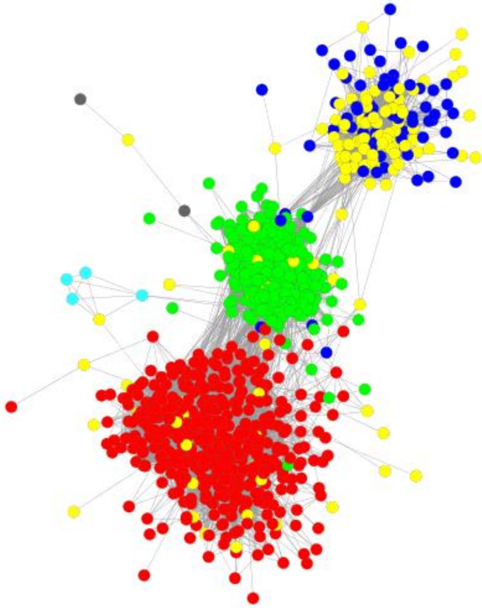
2.1) ribosome biogenesis (43), ribonucleoprotein / small RP subunit (44), small subunit processesome (21), maturation of 5.8s rRNA (15), rRNA export (7), regulation of translation (10), box C/D snoRNP complex (3), ncRNA 3'-end processing, t-UTP complex (5), ribosomal small subunit assembly (4), RNAPI complex (3)

2.2) cytosolic large ribosomal subunit (28), regulation of translation (14), ribosomal large subunit biogenesis (11), ribosome biogenesis (19), multi-eIF complex (4), RNAPIII complex (3), ribosome export (5), chaperone (3)

1.1) cellular ion homeostasis (9), carboxylic acid transport (7), vitamin metabolism (6), metal ion binding (36), cellular amide metabolism (7), alcohol biosynthetic process (5), protein complex assembly (9), mRNA metabolism (11), intrinsic to membrane (56), mitochondrial envelope (17), Glycolysis / Gluconeogenesis (4), hexose metabolism (5), cell cycle (10)

1.2) calcium ion binding (5), biopolymer glycosylation (5), endoplasmic reticulum (11), protein complex assembly (13), exocytosis (5), electron transport / succinate dehydrogenase (3), mitochondrion inner membrane (11),, rRNA processing (10), hexose metabolism (3)

1.3) reproductive proces (5), mitochondrial inner membrane (5), vacuolar transport (4), ribosomal subunit (4), Golgi apparatus (4), transmembrane protein (12), ncRNA processing (3), metal ion binding (5)



3.1) mitotic sister chromatid segregation (7), Golgi vesicle transport,(7) ,cellular protein complex disassembly (5) spindle pole body duplication (3), methyltransferase, (5) DNA replication (9), organelle lumen (38) amino-acid biosynthesis,(7), ribosome biogenesis (13), cell cycle

3.2) Sphingolipid metabolism, (3) , lipid catabolism (4), alcohol / glucose catabolic process (6), cell ageing (5) , GTPase activity (5), aromatic compound biosynthesis,(4) , RNA modification (4), Amino acid transport (3), ribosome biogenesis (15)

Figure 2. 22 An undirected network showing cytosolic ribosomes first neighbours in Hillenmeyer's network.
The network has been visualised using a force directed layout, and modules identified by GLay are mapped on the network. Node colour is representative of GLay module (see legend), edge length represents MI score. Accompanying annotation indicates functional enrichment defined by DAVID. Red text indicates an FDR of ≤ 0.05, and green indicates an FDR of ≤0.1, black represents non significant enrichment.

Ribosomal Gene	**Energy Metabolism Gene**	**MI**	**Pearson correlation coefficient**
GSP2	*IDH2*	0.451	0.744
BMS1	*PYK2*	0.415	0.759
BMS1	*ADH1*	0.367	0.639
BMS1	*FUM1*	0.365	0.716
GSP2	*WCR6*	0.339	0.624
POP7	*ADH1*	0.327	0.444
BMS1	*VMA6*	0.324	0.643
POP7	*VMA6*	0.317	0.588
RNT1	*PYK2*	0.307	0.668
NOB1	*FUM1*	0.286	0.602
RCL1	*FUM1*	0.285	0.641
RNT1	*ADH1*	0.281	0.532
BMS1	*COX12*	0.276	0.546
RNT1	*VMA6*	0.273	0.602
NOB1	*IDH2*	0.267	0.531
GSP2	*FUM1*	0.266	0.486
BMS1	*QCR6*	0.264	0.630
RPS27A	*ADH7*	0.262	0.583
RPL22B	*YOR283W*	0.249	0.579
RPL22B	*ADH7*	0.248	0.669
RPS9A	*SDH1*	0.247	0.339
POP7	*PYK2*	0.244	0.489
RPS6B	*IDH2*	0.241	0.569
RPS9A	*ADH7*	0.241	0.568
POP7	*COX12*	0.237	0.479
POP7	*FUM1*	0.236	0.517
RPL22B	*SDH4*	0.235	0.578
BMS1	*IDH2*	0.227	0.512
RNT1	*QCR6*	0.225	0.511
RPS9A	*LAT1*	0.223	0.387
RPL22B	*COX8*	0.219	0.578
RPS14B	*ATP18*	0.218	0.772
RPL10	*ADH1*	0.218	0.509
GSP2	*KGD2*	0.215	0.607
RCL1	*QCR6*	0.213	0.540
HRR25	*FUM1*	0.210	0.484
RCL1	*ADH1*	0.208	0.465
RNT1	*VMA2*	0.208	0.480
RPS17B	*SOL4*	0.207	0.696
RNT1	*FUM1*	0.205	0.498
RCL1	*IDH2*	0.196	0.481
RCL1	*PYK2*	0.194	0.553
RPS27A	*YOR283W*	0.194	0.501
FCF1	*PYK2*	0.192	0.528
NOB1	*QCR6*	0.191	0.444
POP6	*SDH4*	0.190	0.550
HRR25	*IDH2*	0.188	0.433
RPS27A	*COX8*	0.188	0.514
RPL31B	*FUM1*	0.188	0.520

RPL43B	*LAT1*	0.187	0.544
HRR25	*PDC1*	0.181	0.480
POP5	*TPI1*	0.180	0.489
RPS14B	*THI3*	0.178	0.672
CKA1	*FUM1*	0.178	0.468
POP7	*QCR6*	0.178	0.460
RPS6B	*FUM1*	0.177	0.474
RNT1	*COX12*	0.176	0.439
BMS1	*PDC1*	0.176	0.486
FCF1	*FUM1*	0.176	0.514
RPS9A	*TKL1*	0.175	0.393
RPL31B	*VMA6*	0.175	0.433
RPS27A	*LAT1*	0.174	0.584
RPL7B	*LAT1*	0.171	0.355
RPL31B	*ADH1*	0.170	0.414
RPS25A	*PMA1*	0.170	0.546
RPL15A	*IDH2*	0.170	-0.454
RPS14B	*LAT1*	0.169	0.357
GSP2	*ADH1*	0.169	0.305
RPL38	*SDH3*	0.169	0.585
POP7	*IDH2*	0.168	0.353
RPL43B	*TDH1*	0.168	0.504
POP6	*YOR283W*	0.166	0.500
EMG1	*LAT1*	0.165	0.508
RPL15A	*QCR6*	0.164	-0.469
RCL1	*PDC1*	0.163	0.474
RPL38	*COX8*	0.160	0.483
RPL38	*ADH7*	0.160	0.443
RPL10	*FUM1*	0.160	0.385
RPS27A	*SDH3*	0.158	0.483
RPS9A	*SDH3*	0.156	0.413
RNT1	*RPE1*	0.156	0.376
RPS6B	*QCR6*	0.155	0.422
RPS17B	*KGD2*	0.154	0.553
GSP2	*PDC1*	0.154	0.294
RPL31B	*COX12*	0.154	0.389
RPL17B	*TDH1*	0.153	0.450
RCL1	*VMA6*	0.153	0.389
BMS1	*PFK1*	0.151	0.447
CKA1	*PDC1*	0.150	0.422

Table 2. 10 All the edges between cytosolic ribosomal factors and energy metabolism genes
Table is ordered by MI value, with accompanying Pearson correlation coefficient. MI scores and Pearson correlation coefficients are shown to three decimal places

2.3.5.2 Verification of cytosolic ribosomal proteins link to energy metabolism

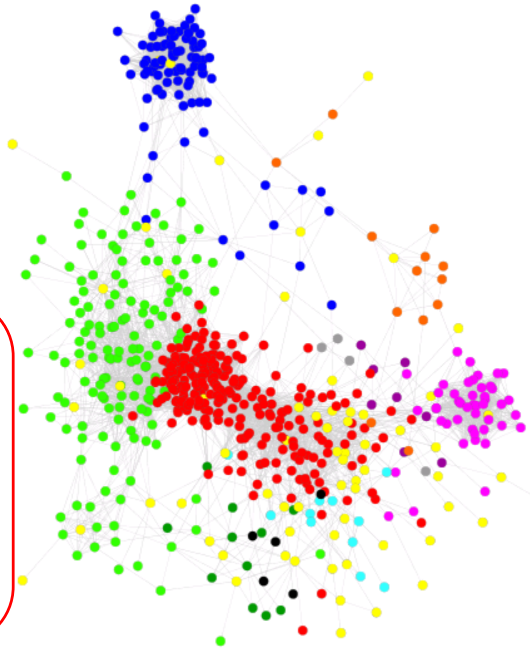
In order to verify the linkage between cytosolic RPs and energy metabolism pathways the analysis was repeated for Vulpe's network (Figure 2.23). The initial level of modularisation identified eight modules as opposed to the three in Hillenmeyer's network, a breakdown of each module is shown in Table 2.11. The results once again suggested that there is a link between cytosolic RPs and cytoplasmic energy metabolism pathways. Functional analysis on each module identified associations between cytosolic ribosomes and rRNA transport (module 1), electron transport chain and large ribosomal subunit (module 2), ubiquitin dependent protein catabolism and translation (module 3), mitochondrial translation and cellular respiration (module 4), cytosolic ribosome and translation regulation (module 5), cytosolic large ribosomal subunit and protein transport (module 6), cytoplasm (module 7), cytosolic ribosome (modules 8 and 9).

The most interesting results were within sub-module 1.2. Sub-module 1.2 contained four genes annotated as glucose catabolic process, these were *EMI2*, *TDH3*, *RPE1* and *TKL1*. In Hillenmeyer's fitness network, I reported that *TKL1* was phenotypically linked to ribosome biogenesis genes, here *TKL1* had 49 first neighbours, three of which were ribosome biogenesis proteins, one was *RPS23*. Sub-module 1.2 also contained glyceraldehyde-3-phosphate dehydrogenase (*TDH3*), the same class of enzyme shown to be first neighbours with genes encoding RPs in Hillenmeyer's network (Figure 2.23). Finally, *RPE1* which encodes ribulose 5-phosphate epimerase. RPE1 catalyses the penultimate reaction in riboneogenesis, specifically the conversion xylulose-5-phosphate to ribulose-5-phosphate [92]. Within sub-module 1.1 there was enrichment five hexose metabolism genes, *GLG1*, *PGM3*, *MDH2*, *PYK2* and the phosphoglycerate mutase *YOR283W*. *PGM3* is a phosphoribomutase which catalyses with inter-conversion of ribose-1-phosphate and ribose-5-phosphate in the PPP and is an alternative route for ribose production rather than riboneogenesis [92].

Surprisingly there is a single cytosolic RP in sub-module 2.1, identified as *RPL2B*. *RPL2B* is first neighbours with every gene within sub-module 2.1, but is not connected to any proteins located in other modules. This suggests that the *RPL2B*, a protein classified as localised to the cytoplasm, shows a similar fitness profile upon deletion to only mitochondrial RPs and mitochondrial proteins in general. This feature is not conserved within Hillenmeyer's network.

Module	**Colour**	**Nodes**	**Edges**
1.1	Red	132	4041
1.2	Red	132	1233
2	Blue	96	1994
3.1	Green	62	397
3.2	Green	62	357
3.3	Green	29	100
4	Pink	50	575
5	Light Blue	21	31
6	Orange	16	28
7	Dark Green	13	19
8	Purple	12	11
9	Black	8	11

Table 2. 11 Breakdown of cytosolic ribosomes first neighbours modules from Vulpe's network



🟥 Cytosolic RP GLay Module 1

🟦 Cytosolic RP GLay Module 2

🟩 Cytosolic RP GLay Module 3

🟪 Cytosolic RP GLay Module 4

🟦 Cytosolic RP GLay Module 5

🟧 Cytosolic RP GLay Module 6

🟩 Cytosolic RP GLay Module 7

🟪 Cytosolic RP GLay Module 8

⬛ Cytosolic RP GLay Module 9

🟨 Cytosolic Ribosomal Protein

1.1) decarboxylase (3), glucan biosynthesis (3), coenzyme metabolic process (6), oxidation reduction (9), glucose metabolic process (5), endocytosis, (5) mitochondrion (9), spore wall biogenesis (3), stress resposne (10), dephosphorylation (3), hexose metabolism (5)

1.2) cytosolic ribosome (23), regulation of translation (13), cytosolic small ribosomal subunit (14), rRNA transport (7), cytosolic large ribosomal subunit (8), nucleolus, ribosome small subunit assembly, rRNA-binding (3), glucose catabolic process (4), amino acid phosphorylation, (8) replication fork protection complex (3), sexual reproduction (8), chromosome (10)

2) electron transport chain, (5) secondary metabolic process (4), iron ion binding (6), large ribosome subunit (6), integral to mitochondrial membrane (25)

8) cytosolic ribosome (3)

9) cytosolic large Ribosome subunit (3)

3.1) translational termination (4), atp-binding (15), cell wall (5), actin cytoskeleton (4), P-loop (5), vacuolar membrane (3), reproductive cellular process (4), mitochondrial (8), protein localization (4)

3.2) regulation of translation (6), stress response (6), Ribosome (4), macromolecule biosynthesis (4), cytoskeleton organization (3), cell fraction (5), integral to organelle membrane (17), mitochondrial envelope (4), ribosome biogenesis (3)

3.3) ubiquitin-dependent protein catabolic process via the multivesicular body sorting pathway (5), cytosolic ribosome (4), mitochondrion (3), zinc-finger (3)

4) Mitochondrion (37), mitochondrial translation (20), ribosome (6) cellular respiration (5), amino acid activation, (5) GTPase activity (3), mitochondrial inner membrane (7), sporulation (3), regulation of translation (3)

5) cytosolic ribosome (10), regulation of translation (6), ribosome biogenesis (7), cell division (3), nucleus (7)

6) cytosolic large ribosomal subunit (3), protein transport (3)

7) cytoplasm (7) cytosolic large ribosomal subunit (2)

Figure 2. 23 An undirected network showing cytosolic ribosomes first neighbours in Vulpe's network

The network has been visualised using a force directed layout, edge length represents MI score. Modulesdefined by GLay are mapped on the parent network. Node colour represents sub-network (see legend). Functional annotation by DAVID for each sub-network is shown in the corresponding coloured box.

2.3.5.3 Mitochondrial RPs are phenotypically correlated to ribosome biogenesis despite being compartmentally separated.

Using the same analysis pipeline as the cytosolic RP first neighbour analysis, mitochondrial RPs were mapped onto both Hillenmeyer's and Vulpe's fitness networks and the first neighbours identified. The first neighbour mitochondrial RP networks for Hillenmeyer's and Vulpe's datasets are shown in Figures 2.24 and 2.25 respectively. Details of the number of nodes and edges are given in Tables 2.12 and 2.13 respectively.

Module 1 of Hillenmeyer's mitochondrial network formed three sub-modules (Figure 2.24). All of which, were significantly enriched in mitochondrial RPs and other proteins localised within the mitochondria. This observation was also conserved in module 1 in Vulpe's mitochondrial network (Figure 2.25). Both modules also contain energy metabolism pathways localised within the mitochondria, such as TCA cycle, and electron transport chain. However in sub-module 1.1 of Hillenmeyer's data, there was enrichment of seven hexose metabolism genes, *PSK1*, *CDC19*, *PMI40*, *PFK27*, *TDH1*, *RPE1* and *TKL1*. First neighbour analysis showed that these hexose metabolism genes had 139 first neighbours within the mitochondrial network. DAVID identified 34 genes encoding mitochondrial proteins as the top hit. *TKL1*, *RPE1* and *TDH1* are key enzymes in glycolysis and riboneogenesis. TDH1 is involved in glycolysis, suggesting that knocking out glycolysis leads to the same phenotype as knocking out an RP.

Within sub-module 3.1 of Hillenmeyer's network there was enrichment of 'negative regulation of glycolysis'. Module 3.1 was also enriched glycolysis related genes *PFK1* and *PYK2,* both of which are tightly regulated enzymes that catalyse thermodynamically favoured reactions [92]. In sub-module 3.1 of Vulpe's network (Figure 2.25) there is enrichment of the electron transport chain and oxidation reduction proteins.

Module	Nodes	Edges
1.1	199	2451
1.2	167	2902
1.3	31	153
2.1	120	4244
2.2	85	2719
2.3	5	9
3	11	8

Table 2. 12 Breakdown of mitochondrial ribosomes first neighbours modules for Hillenmeyer's network.

Module	Colour	Nodes	Edges
1	Red	98	1051
2	Green	90	1944
3	Blue	23	115
4	Pink	16	54
5	Light Blue	14	38
6	Orange	3	3

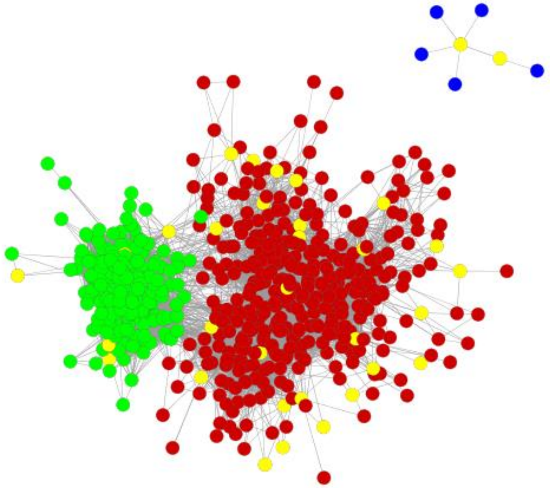
Table 2. 13 Breakdown of mitochondrial ribosomes first neighbours modules for Vulpe's network.

🟥 Mitochondrial RP GLay Module 1

🟦 Mitochondrial RP GLay Module 2

🟩 Mitochondrial RP GLay Module 3

🟨 Mitochondrial Ribosomal Protein



2) Mitochondrial translation optimization (1), targetting protein (1), phosphodiesterase (1)

1.1) mitochondrial ribosome (13), vitamin metabolism (8), er-GOLGI transport (5), negative regulation of organelle organisation (7), HEAT (3), DNA duplex unwinding (4), hexose metabolism (7), cell cycle (9), establishing protein localisation (20)

1.2) mitochondrial ribosome (26), mitochondrial matrix (12), endocytosis (7), oxidoreductase (11), TCA cycle (4), preribosome – large subunit precursor (4)

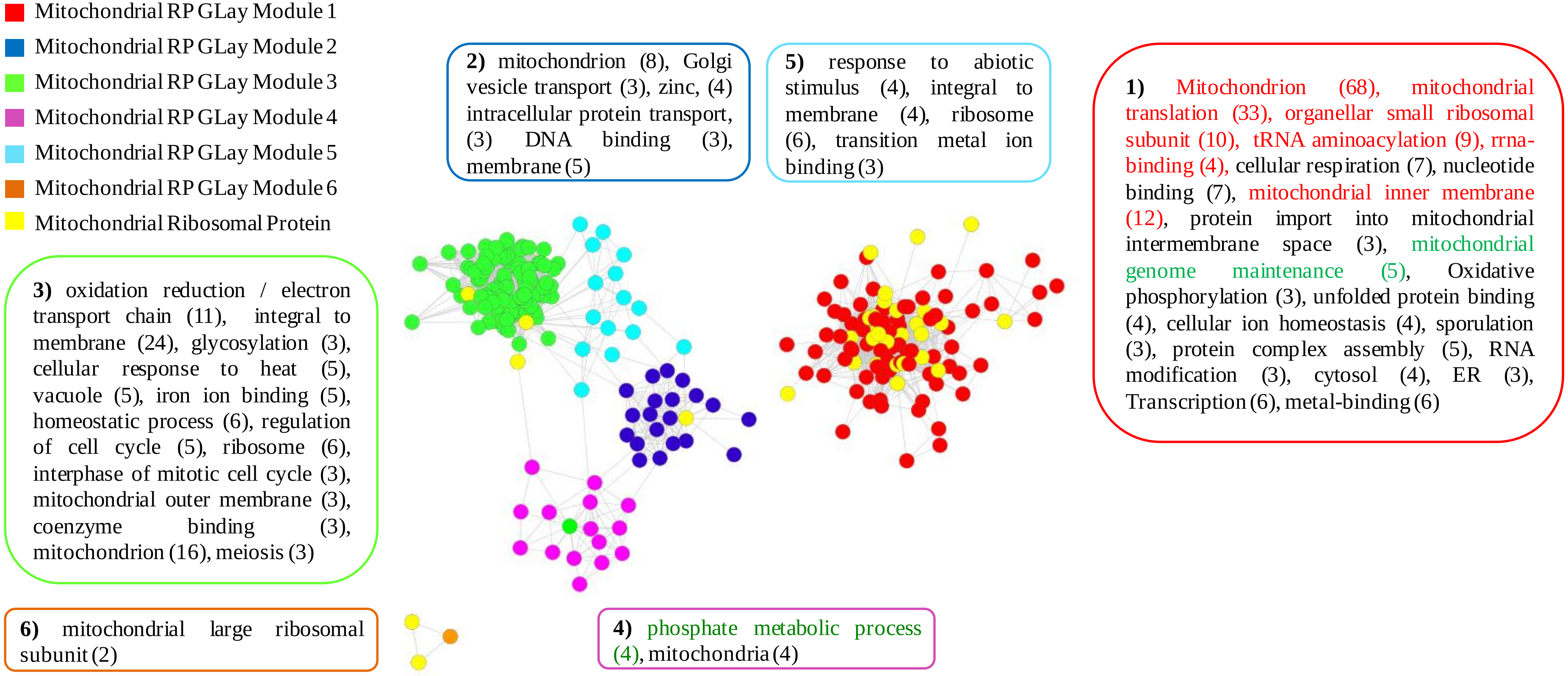
1.3) mitochondrial ribosome (4), integral to membrane (8), nucleotide binding (4)

3.1) sphingolipid metabolism (3), mRNA metabolism (11), ncRNA metabolism (14), negative regulation of gluconeogenesis (5), alcohol catabolic process (5), thiamin biosynthesis (3), ribonucleoprotein complex (3), ribosome biogenesis (13)

3.2) organelle lumen (19), mRNA polyadenylation (3), chromatid segregation (5), DNA replication, (6), electron carrier activity (4), TCA cycle (3),

Figure 2. 24 An undirected network showing the first neighbours of mitochondrial ribosomes from the Hillenmeyer dataset

The network has been visualised using a force directed layout, edge length represents MI score. Modules defined by GLay were been mapped onto the parent network. Node colour represents the module (see legend) with yellow nodes representing mitochondrial RPs. Accompanying annotation indicates functional enrichment defined by DAVID. Red text represents an FDR ≤ 0.05, and green text represents an FDR ≤ 0.1.



🟥 Mitochondrial RP GLay Module 1
🟦 Mitochondrial RP GLay Module 2
🟩 Mitochondrial RP GLay Module 3
🟪 Mitochondrial RP GLay Module 4
🟦 Mitochondrial RP GLay Module 5
🟧 Mitochondrial RP GLay Module 6
🟨 Mitochondrial Ribosomal Protein

2) mitochondrion (8), Golgi vesicle transport (3), zinc, (4) intracellular protein transport, (3) DNA binding (3), membrane (5)

5) response to abiotic stimulus (4), integral to membrane (4), ribosome (6), transition metal ion binding (3)

1) Mitochondrion (68), mitochondrial translation (33), organellar small ribosomal subunit (10), tRNA aminoacylation (9), rrna-binding (4), cellular respiration (7), nucleotide binding (7), mitochondrial inner membrane (12), protein import into mitochondrial intermembrane space (3), mitochondrial genome maintenance (5), Oxidative phosphorylation (3), unfolded protein binding (4), cellular ion homeostasis (4), sporulation (3), protein complex assembly (5), RNA modification (3), cytosol (4), ER (3), Transcription (6), metal-binding (6)

3) oxidation reduction / electron transport chain (11), integral to membrane (24), glycosylation (3), cellular response to heat (5), vacuole (5), iron ion binding (5), homeostatic process (6), regulation of cell cycle (5), ribosome (6), interphase of mitotic cell cycle (3), mitochondrial outer membrane (3), coenzyme binding (3), mitochondrion (16), meiosis (3)

6) mitochondrial large ribosomal subunit (2)

4) phosphate metabolic process (4), mitochondria (4)

Figure 2. 25 An undirected network showing the first neighbours of mitochondrial ribosomes from the Vulpe dataset
The network has been visualised using a force directed layout, edge length represents MI score. Modules defined by GLay have been mapped onto the parent network. Node colour represents module (see legend) with yellow nodes representing mitochondrial RPs. Accompanying annotation indicates functional enrichment defined by DAVID. Red text represents an FDR ≤ 0.05, and green text represents an FDR ≤ 0.1.

2.3.5.4 Cell cycle gene *BUB1* shows significant phenotypic correlation to over 45 genes encoding RPs

The results of the fitness network suggested that cytosolic RPs have similar fitness profiles to genes involved in cell cycle and correct chromosome segregation. This observation is most apparent in the cytosolic RP first neighbour analysis (Hillenmeyer Figure 2.22 sub-module 3.1 and Vulpe Figure 2.23 sub-modules 1.2, 3.1, 4.1 and 5.1). To investigate further, I used the Hillenmeyer network and identified the how many cell cycle genes were first neighbours of cytosolic RPs. Cell cycle genes were split into meiosis (Figure 2.26A) and mitosis (Figure 2.26B). RPs are represented in green and cell cycle represented in red. The analysis found that only a single cell cycle gene, *BUB1*, localised within a module of cytosolic RPs (enclosed within the black box of figures 2.26A and 2.26B). *BUB1* encodes a serine / threonine protein kinase which has an essential role in spindle assembly checkpoint and prevents cell cycle progression in the presence of spindle damage [129].

The mitosis and meiosis networks were combined (Figure 2.26C). 49 of the 66 first neighbours of *BUB1* were RPs (represented as green nodes), with the remaining 17 being involved in translation initiation and cell cycle processes (represented as grey nodes). Isolating solely *BUB1* and analysing the MI score between BUB1 and its adjacent edges, identified two clear peaks indicating that edges could be classified into two distinct groups (figure 2.26D). The genes which were present within the high (0.35MI, p-value: 10^{-77}) and lower, yet still highly statistically significant (0.2MI, p-value: 10^{-45}) peaks were identified. Figure 2.26E shows the breakdown of *BUB1* linkage to genes encoding RPs, edge width represents MI score, and the RP genes have been grouped based on their size (small / large) and function. The results suggest that *BUB1* is more phenotypically correlated to genes encoding small 40S RPs and least phenotypically correlated to large

RPs, representing the first and second peaks in the MI density plot (Figure 2.26D). Finally I calculated the Pearson correlation coefficient between BUB1 and each of its first neighbours across all samples. Mutual information scores and Pearson correlation coefficients between BUB1 and its first neighbours were calculated to three decimal places (Table 2.14). The association of BUB1 with RPs is unexpected, however the results show yeast strains with a *BUB1* mutant, have the same fitness as yeast strains containing a RP mutant, and that the phenotypic correlation is dependent on the size of the RP also.

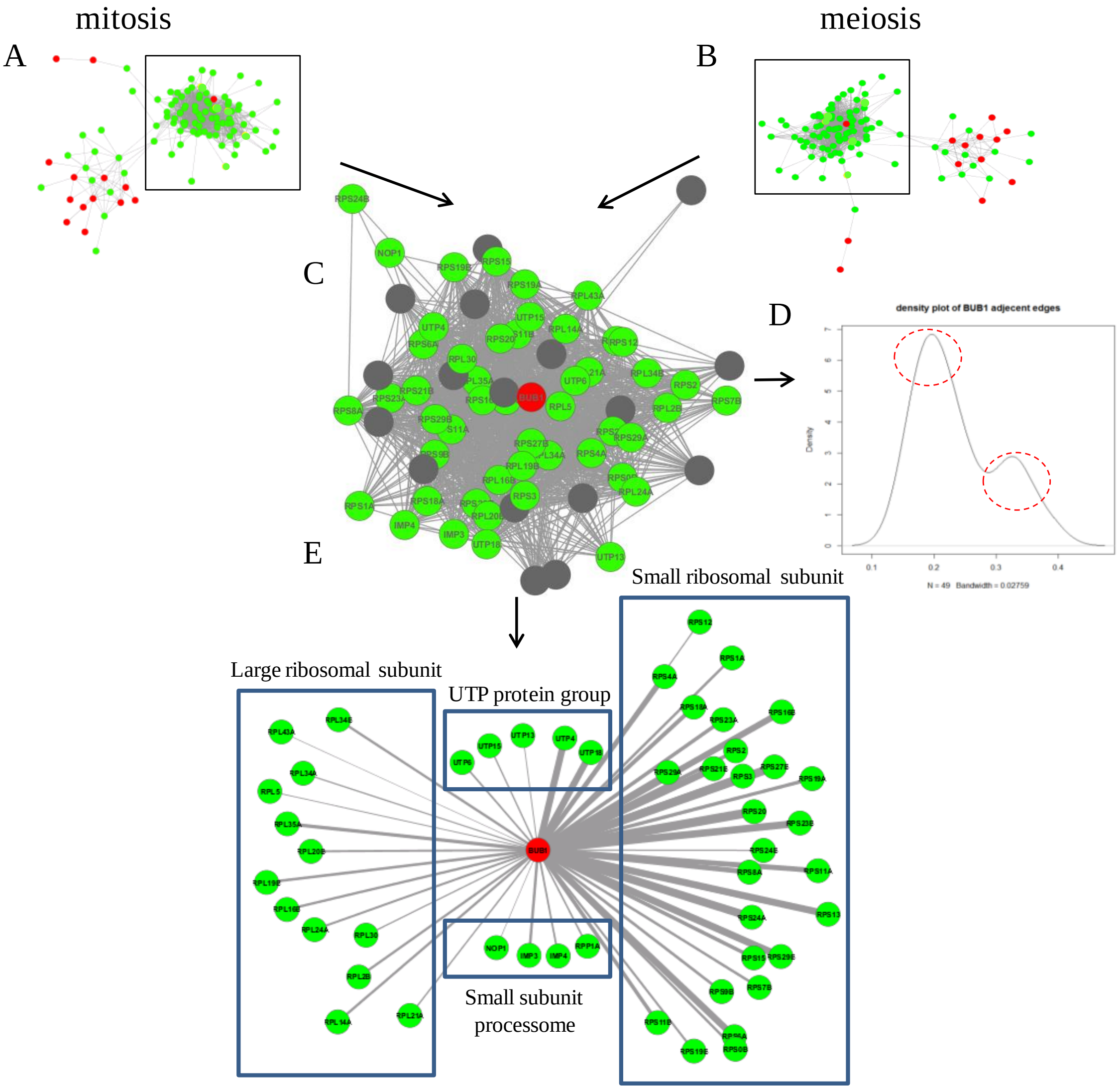


Figure 2. 26 The phenotypic linkage of *BUB1* to cytosolic RPs.
Panel A. The edges between cytosolic RPs (green nodes) and mitosis (red nodes). Network has been visualised using a force directed layout. Panel B. The interactions between cytosolic RPs (green nodes) and meiosis (red nodes). Network has been visualised using a force directed layout. Panel C. *BUB1* is highly correlated to cytosolic RPs in both mitosis and meiosis. Grey nodes are non RP. Panel D. A density plot representing the MI score of all adjacent edges to cell cycle protein *BUB1* shows 2 specific peaks (highlighted by dashed red circles) , Panel E A network representing *BUB1* RP neighbours, edge width represents MI score. Interestingly, the strongest linkage is between *BUB1* and small cytosolic RPs (right most box). The weakest interactions (though still highly significant) are predominantly with large cytosoplasmic RPs (left most box). Processosome components contain a mixture of both high and low correlated genes.

Gene 1	Gene 2	MI	r2_pearson
BUB1	*RPS20*	0.391	0.744
BUB1	*RPS27B*	0.378	0.747
BUB1	*RPS24A*	0.364	0.711
BUB1	*RPS23B*	0.340	0.679
BUB1	*RPS29B*	0.337	0.666
BUB1	*RPS13*	0.332	0.691
BUB1	*RPS21B*	0.330	0.615
BUB1	*UTP4*	0.327	0.679
BUB1	*UTP18*	0.323	0.652
BUB1	*RPS6A*	0.322	0.651
BUB1	*RPS3*	0.317	0.701
BUB1	*RPS16B*	0.309	0.705
BUB1	*RPS11A*	0.279	0.650
BUB1	*RPS8A*	0.278	0.599
BUB1	*RPS4A*	0.267	0.626
BUB1	*RPS18A*	0.250	0.566
BUB1	*RPS29A*	0.244	0.597
BUB1	*RPS11B*	0.238	0.517
BUB1	*RPS23A*	0.238	0.613
BUB1	*RPS15*	0.237	0.593
BUB1	*RPS2*	0.237	0.611
BUB1	*RPS19A*	0.232	0.579
BUB1	*RPL35A*	0.229	0.502
BUB1	*RPS9B*	0.212	0.691
BUB1	*RPL19B*	0.210	0.465
BUB1	*IMP3*	0.209	0.589
BUB1	*RPS1A*	0.207	0.560
BUB1	*RPS7B*	0.206	0.572
BUB1	*RPS0B*	0.203	0.562
BUB1	*RPL16B*	0.202	0.503
BUB1	*RPL34B*	0.202	0.520
BUB1	*RPL14A*	0.202	0.501
BUB1	*RPL2B*	0.200	0.466
BUB1	*RPL24A*	0.194	0.435
BUB1	*RPS19B*	0.193	0.494
BUB1	*RPL20B*	0.191	0.470
BUB1	*IMP4*	0.185	0.508
BUB1	*UTP6*	0.184	0.577
BUB1	*RPP1A*	0.181	0.463
BUB1	*RPL21A*	0.176	0.415
BUB1	*UTP15*	0.175	0.500
BUB1	*RPL30*	0.171	0.467
BUB1	*RPS24B*	0.170	0.483
BUB1	*RPL34A*	0.170	0.413
BUB1	*RPS12*	0.169	0.523
BUB1	*UTP13*	0.161	0.526
BUB1	*RPL5*	0.160	0.437
BUB1	*NOP1*	0.151	0.410
BUB1	*RPL43A*	0.150	0.467

Table 2. 14 All interactions between BUB1 and cytosolic RPs.
Table is ordered by MI value, with accompanying Pearson correlation coefficient. MI scores and Pearson correlation coefficients are shown to three decimal places

2.4 Discussion

The network construction and analysis reported in this chapter is the first time an MI based reverse engineering approach has been applied to Hillenmeyer's genome-wide fitness data. In 2008, the analysis of *S. cerevisiae* fitness data proved that genes previously thought non-essential actually had an essential role in providing tolerance to different stresses [40]. Almost all genes within the *S. cerevisiae* genome (97%) are required for growth after exposure of a specific chemical or stress [40]. The paper proved that using fitness data can potentially bridge the gap in understanding the relationships between genotype and phenotype.

Two important concepts were revealed as a result of this study. The first is that genes which share a common function, form highly interconnected modules, consistent with mutual contribution to cell fitness. For example, mutating an RP would lead to the same phenotype as mutating another RP, as the cells ability to produce a functional ribosome has been affected. The identification of modules enriched in similar cellular functions is consistent with the preliminary hierarchical clustering validation studies done by Hillenmeyer *et al* on the same dataset in which they showed three clusters each enriched with genes encoding functionally similar proteins (proteasome core complex, peroxisome, and chaperonin containing T-complex) [40]. Furthermore, the independent fitness data provided by Vulpe Labs also revealed similar results. These results validate the use of genome-wide fitness data as an informative source of providing biological insight when applied to network inference methodologies.

Secondly, exploration of the fitness networks revealed potentially interesting phenotypic linkages between select groups of genes. For example, the results suggested that RPs are phenotypically correlated to a diverse set of cellular functions, including cell cycle, glycolysis and chromatid segregation. Not only that, but the analytical approach used allows for the identification of which genes specifically from each functional group exhibit the most strongly correlated co-fitness profiles. Surprisingly, the network analysis identified a novel discovery, a strong correlation between the non-essential gene *BUB1* and genes encoding RPs, specifically RPs that constitutes the small ribosomal subunit.

This genome-wide network analysis on Hillenmeyer's fitness data can provide a platform on which to generate hypotheses on a broader scale. In combination with network interrogation techniques and current congruent literature, it is possible to identify potential candidate genes that can be used to experimentally validate hypotheses generated from this network analysis. Below I discuss the most interesting and statistically significant phenotypic correlations suggested by this fitness network analysis.

2.4.1 Ribosomal proteins are required for proper chromosome segregation and cell cycle progression

So far, the results suggest that there is a potential phenotypic link between RPs and cell cycle processes. Indeed, this may be explained by the fact that as cells differentiate, they require an up-regulation in the genes encoding for RPs. However, of all genes involved in cell cycle processes, only *BUB1* appears to be significantly correlated to over 45 RPs. BUB1 is a non-essential protein kinase which controls the checkpoint into anaphase [130]. This spindle checkpoint delays the onset of anaphase in cells which have developed defects in mitotic spindle assembly or if there are adverse attachments of the spindle microtubules to chromosomes [131, 132]. The role of BUB1 in anaphase checkpoint control is conserved across eukaryotes [133] [134]. Knocking out *BUB1* alters chromosome segregation [134]. In *S. pombe* deletion of *BUB1* causes chromosomes to arrest on the spindle during anaphase, leading to chromosome loss [129]. BUB1 requires other spindle checkpoint components into order to fulfil its function, including the MAD protein family and additional BUB proteins [135]. Therefore, it is extraordinary as to why only the non-essential gene *BUB1*, shows such a strong phenotypic correlation to RPs, with almost 75% of its first neighbours being RPs.

In 2004, a study in *S. cerevisiae* revealed the first link between ribosome biogenesis and chromosome segregation [136]. Expression of the ribosome biogenesis gene *RRB1* is induced when the spindle checkpoint is activated however inactivation of *RRB1* leads to abnormal chromosome segregation blocking mitosis at the checkpoint into anaphase [136]. The role RRB1 in controlling the assembly of ribosomal subunits [137] and the transcription of RPs [136] is key to cell cycle progression. The work in this chapter also suggests that there is a close link between ribosome biogenesis and cell cycle progression. Though the exact relationship between these two biological processes isn’t clear, one possible hypothesis is that the lack of a functional ribosome or a delay in ribosome assembly halts cell cycle prior to chromosome segregation. The results from this study demonstrate the phenotypic linkage between *BUB1* and genes encoding RPs. A mutation in either leads to a highly similar phenotype. This raises the possibility that adverse proteins involved in RP synthesis may be a potential mechanism for identifying chromosomal instability [136]. Noteworthy, is that these results demonstrate the novel observation that *BUB1* has a stronger phenotypic correlation to small RP genes, than to large RP genes. A possible counter hypothesis is that yeast strains containing deleted genes encoding small RPs have a different phenotype to those strains containing deletions of large RPs (as suggested by separation of small and large RPs in sub-modules 3.1. and 3.2 in Figure 2.7). As such, it is possible that *BUB1* happens to be more closely correlated to small RPs simply because they have the same fitness contribution when deleted. Many RPs have paralogues, therefore, it is possible that deleting some small RPs isn't fatal to the cell [102], and as *BUB1* is non-essential, they both may have a similar phenotype when deleted. If true, this could mean that the correlation between *BUB1* and genes encoding small RPs is not biologically relevant. However, Table 2.14 lists the RPs that are correlated to *BUB1* and it shows that even RPs that do not have paralogues (such as *RPS20, RPS13 and RPS3*) are strongly correlated to *BUB1,* and that when deleted they are fatal to the cell (as reported on SGD). In light of this, the correlation between *BUB1* and RPs, especially small RPs may in fact be biologically significant. What is undeniable however, is the fact that only *BUB1* is strongly correlated to RPs in general, no other checkpoint proteins are significantly correlated. The importance of why *BUB1* specifically is correlated to RPs that do not have paralogues as well as those that do, can only be validated through experimental techniques. This analysis has identified potential candidate genes in which to test this hypothesis experimentally.

2.4.2 Genes involved in glycolysis may regulate the rate of ribosome biogenesis

The network analyses on the two independent fitness datasets suggested that RPs may be phenotypically linked to genes involved in energy metabolism pathways, in particular glycolysis. For example, in sub-module 1.1 of Hillenmeyer's network I showed that the gene *TDH1*, responsible for catalysing the sixth step in glycolysis, is directly connected to RPs and ribosome biogenesis genes (Figure 2.5). Sub-module 2.2 of Hillenmeyer's network showed enrichment of glycolysis genes and ribosome biogenesis genes (Figure 2.6). The significance of this result is that once again, a subset of RPs localise within modules containing energy metabolism genes instead of with the majority of other RPs.

Keeping in mind the rate of riboneogenesis is dependent on glycolysis intermediates [92], mutating a gene encoding a glycolytic enzyme would affect flux through glycolysis, resulting in a lower concentration of glycolytic intermediates that can be shunted through the riboneogenesis pathway thus causing a decrease in ribosome biogenesis. The results suggest that support this theory, as a mutation in the *TDH1* gene produced a fitness profile that is highly correlated to the large RP genes, *RPL17B*, *L43B* and biogenesis gene *TMA22*.

The translaldolases *NQM1* (Hillenmeyer sub-module 2.1) and *TAL1* (Vulpe module sub-1.2) have a strong phenotypic correlation to RPs either directly in the case of *TAL1* or indirectly for *NQM1*. The results show that although *NQM1* does have any direct connects to RPs, it connects indirectly through the uncharacterised gene known as *YGL242C*. The function of this gene is not known, however these results suggest that it may be involved regulating RPs. Reports by Clasquin *et al* have shown that deletion of aldolase genes, *TAL1* and its paralogue *NQM1*, inhibits the non-oxidative PPP, thereby increasing the concentration of metabolites available for riboneogenesis, as such, the flux through SHB17 is quadrupled [92].

Note worthy is that the energy pathways in Hillenmeyer module 2 (Figure 2.6) are split into two distinct modules, the first contains energy pathways predominantly localised in the cytoplasm (sub-module 2.2) and energy metabolism pathways localised to the mitochondria (sub-module 2.1), however it is only module 2.2 that shows association to ribosome biogenesis genes. Of course, it is expected that ribosomal genes be strongly correlated to energy metabolism pathways, given that eukaryotic ribosome biogenesis and translation is an costly energetic process (requiring ~45% of ATP supplies in mouse [138]), the question is why do RPs have a strong phenotypic linkage to only genes involved in glycolysis or the pentose phosphate pathway. The results are consistent with those presented by Clasquin *et al*, in which they reported that the rate of ribosome biogenesis is dependent on the rate at which glycolytic intermediates are produced [92].These observations were also supported by network analysis done on Vulpe's fitness data.

2.4.3 Limitations of fitness data

The network analysis reported in this chapter has produced numerous results and hypotheses. However, it is important to note that as an MI based method has been used to construct these networks; they do not reveal the relationship between cause and effect between two genes that share co-fitness. It is plausible that two completely unrelated genes may be strongly correlated because they both contribute the same degree of fitness to the cell. For example, an edge may be inferred between two essential genes, simply because they both would cause cell death when deleted. In other words, deletion of both genes leads to the same phenotypic outcome, yet there is no underlying biological connection between them. Conversely, the linkage between the two genes may represent a true underlying biological connection which has yet to be discovered. The only way to reveal the true underlying connection is through experimental validation and this is one of the limitations of MI based networks. Experimental validation is always required in order to strengthen or disprove any hypotheses garnered from MI based networks.

Similarly to expression data, consideration can be given to temporal aspects, for example fitness data can be used to measure the essentiality of a gene within a given time duration, with measurements taken at several times during growth [37]. Though, for the purposes of constructing these fitness networks, experiments in which drug dose or exposure time was varied were treated as independent experiments which were normalised against their respective controls. This pre-processing and normalisation strategy was also used in Hilllenmeyer *et al's* study [40], and the same data was used for this study. In order to study the degree to which specific genes contribute to the survivability of a yeast strain across time, a different analytical approach would be required, such as a multi-class significance analysis of microarrays (SAM), with each class being a reading taken at specific intervals after exposure to a specific stress. Such an approach would identify genes which significantly change their contribution to cell fitness across time. This type of approach is currently limited as the amount of time-course fitness data that focuses on stress response in *S. cerevisiae* is limited [40].

2.5 Concluding remarks

Reported in this chapter is the first time Hillenmeyer *et al's* compendium has been used to reverse engineer a global fitness MI network. Prior to this analysis, network based approaches using fitness data had been restricted in terms of scope and question to elucidating or mapping specific pathways. Though studies of this type are useful for understanding biological pathways, it does involve excluding a large proportion of the data, thereby excluding potentially valuable information. The work presented in this chapter however, utilises a genome-wide approach. The analytical pipeline used in this chapter allowed the Hillenmeyer compendium and Vulpe dataset to be considered as a whole, rather than focusing on a subset of genes. As a result, I was able to construct fitness networks and identify potentially interesting relationships between different cellular functions at the global level. This work also marks the first time a network based approach has been used to investigate riboneogenesis and the links between ribosome biogenesis and energy metabolism. This work identified several statistically significant correlations between genes, which were previously reported by wet lab experiments performed by Clasquin *et al* during their riboneogenesis studies [92]. In addition, this work highlights how applying a MI based reverse engineering approach can identify potential further future areas of research, one example highlighted in this chapter was the novel linkage between the non-essential *BUB1* gene and both essential and non-essential small RP genes. The fitness network constructed in this chapter can be applied to any biological question and can be used to determine gene dependencies prior to wet lab experiments. However, studies using fitness data are limited as *S. cerevisiae* is the only eukaryotic organism in which there is a large volume of fitness data available. Only in recent years has *S. pombe* fitness deletion library been verified [74], therefore *S. pombe* fitness data is not readily available as *S. cerevisiae* fitness data. Though, due to the homology of *S. cerevisiae* with mammals and *S. pombe* [57], these fitness networks may have an important role in identifying potentially interesting interactions in higher eukaryotes.

The next step is to determine if genome-wide transcriptional data can also be grouped into functional modules and whether the linkages observed in the phenotypic data are conserved at the transcriptional level. In the next chapter I report the inference and analysis of a *S. cerevisiae* expression network.

CHAPTER 3: INFERENCE AND ANALYSIS OF A *Saccharomyces cerevisiae* GENE EXPRESSION NETWORK

3.1 Introduction

In chapter 2, I showed how the application of a reverse engineering approach to a genome-wide fitness dataset could be used to discover the existence of functional modules representing highly correlated fitness profiles.

The analysis of these modules identified a number of phenotypically coupled functional processes. In some cases this has yielded interesting hypothesis regarding ribosomal proteins (RPs). Perhaps the most interesting result was the correlation between the phenotypic profiles of strains mutated in cytosolic RPs and strains mutated in energy metabolism enzymes (e.g. glycolysis). In fact, the fitness networks revealed several direct connections between glycolytic enzymes, cytosolic RPs and ribosomal biogenesis genes.

In this chapter, I describe the application of the same network inference approach utilised in Chapter 2, to a compendium of 269 *S. cerevisiae* transcription factor (TF) knockouts, analysed using an Affymetrix expression profiling approach [68]. The overarching aim of this analysis is to test whether I could identify functional modules from a genome-wide transcriptional network and more specifically if any of the modules provide further evidence of the functional associations discovered in Chapter 2.

Indeed, I was able to identify similar functional modules and validate some of the original hypothesis, including the strong association between RPs with energy metabolism. The network analysis also revealed a strong correlation between the expression of retrotransposons and RPs.

3.2 Methods

3.2.1 The biological system

The availability of vast volumes of high throughput microarray data has made inferring and understanding *S. cerevisiae* transcriptional networks a prime focus. The overall aim of this study is to identify and characterise the structure of an underlying regulatory network representing transcriptionally linked genes in *S. cerevisiae.* The expectation is that development of this network may allow for the identification of overlaps between transcriptional and phenotypic coupling (Chapter 2). To accomplish this I selected the one of the most extensive transcription factor (TF) perturbation datasets available for *S. cerevisiae* [68][71]. Using such a comprehensive TF knockout dataset makes it a prime candidate for reverse engineering a genome wide expression network.

The original expression dataset was published by Hu *et al* in 2007 [68] however it was reprocessed and reanalysed in 2010 by Reimand *et al* due to the lack of background correction and print tip correction during normalisation [71]. The reprocessed data was reported to outperform the original in every respect leading to the identification of almost ten times more differentially expressed genes previously reported by Hu *et al* [71]. For this reason, the reprocessed data was used in this study. The dataset was downloaded from ArrayExpress (accession: E-MTAB-109) and it contained 269 TF knockout mutant strains and 6253 genes. The details of the experimental protocol can be viewed in the original Hu *et al* publication [68]. The methodology used to improve the compendium is reported in Reimand *et al's* publication (Methods: 'Microarray data pre-processing and analysis') [71]. The analysis strategy utilised in this study is very similar to the analysis pipeline required for the construction and analysis of *S. cerevisiae* fitness network reported in Chapter 2. Details are given below.

3.2.2 Network inference

The expression data had been normalised used the VSN package [139], normalisation included background and print-tip correction [71], therefore the data only required formatting for ARACNE. As with the phenotypic analysis, the first stage in extracting useful data from the network was first to choose a suitable threshold. Statistically significant edges were selected using a p-value threshold of 10^{-23} corresponding to an MI $\geq$ 0.25 (Table 3.1). This value was chosen arbitrarily, however it does represent an extremely high stringency cut-off and retained approximately half of the total number of genes, consistent with the prior phenotypic analysis. No edges were eliminated using the data processing inequality (DPI). Using the above MI threshold, 3312 nodes and 127528 edges were retained within the network. The network was visualised using a force directed layout.

3.2.3 Network Analysis: Visualisation and modularisation

The network was visualised in Cytoscape [140] using a force directed layout. Network modularisation was done on the basis of connectivity using the GLay community clustering method [33]. Further levels of modularisation were done if the sub-networks were deemed to contain too many nodes (typically $\geq$300) or could possibly yield additional information. Functional analysis of each cluster was done using DAVID [21] [22]. Similarly to Chapter 2, functional annotations were colour coded depending on their corrected FDR (as detailed in section 2.3.3).

3.2.4 Identification of ribosomal proteins' first neighbours

RPs were classified by their localisation within the cell. The list of ribosomal factors for each group was the same as those used in Chapter 2. The analysis pipeline utilised for the mapping and identification of first neighbours for ribosomal proteins is also identical to that detailed in Chapter 2. Briefly summarised, first neighbours for each ribosomal group were identified, visualisation and modularisation was done as described in section 3.2.3 to ensure consistency.

P-value	Corresponding MI
0.05	0.0185077
0.01	0.0258485
0.001	0.0363507
1.00E-05	0.0573553
1.00E-10	0.109867
1.00E-15	0.162378
1.00E-20	0.21489
1.00E-25	0.267401
1.00E-30	0.319913
1.00E-40	0.424935
1.00E-50	0.529958

Table 3. 1 ARACNE p-values and corresponding MIs for the *S. cerevisiae* expression dataset

3.3 Results

3.3.1 The modular structure of the expression network reflects functional compartmentalisation

The underlying hypothesis was that the modular structure of an *S cerevisae* transcriptional network might, at least in part, resemble the functional modules identified within the fitness network (Chapter 2). I therefore applied the same network analysis pipeline to the gene expression dataset. In this analysis, I performed two levels of modularisation using the community detection algorithm GLay. This allowed the structure of each module to be analysed at a more refined level.

A single level of modularisation identified eight network modules (Table 3.2). Module 1, (Figure 3.1, red nodes) was the largest, containing 1608 nodes and 33336 edges. Among the most enriched functions there are glycolysis and ribosomal biogenesis genes (Figure 3.2). Module 2 (Figure 3.1, yellow nodes) maps in the centre of the parent network and is tightly linked to module 1. Functional analysis shows the significant enrichment of cytosolic and mitochondrial ribosomal proteins (Figure 3.3). Module 3 (Figure 3.1, blue nodes) is significantly enriched in stress response genes (Figure 3.4). Module 4 (Figure 3.1, purple nodes) is significantly enriched in transposable elements (Figure 3.5). Module 5 (Figure 3.1, light blue nodes) is significantly enriched in mating pheromone activity (Figure 3.7). Smaller modules 6, 7 and 8 could not be modularised with a further level of GLay clustering, they did however reveal association between transcription and cell cycle (module 6, Figure 3.8), the co-ordination of a small group of transmembrane proteins (module 7, Figure 3.9) and metal ion binding and the endomembrane system (module 8, figure 3.10).

Interestingly, most modules (71%) could be significantly characterised by a specific functional profile (FDR $\leq$ 0.1%). The raw output of the module by module functional enrichment analysis performed with DAVID is available in folder ‘Chapter 2’ of the supplementary CD.

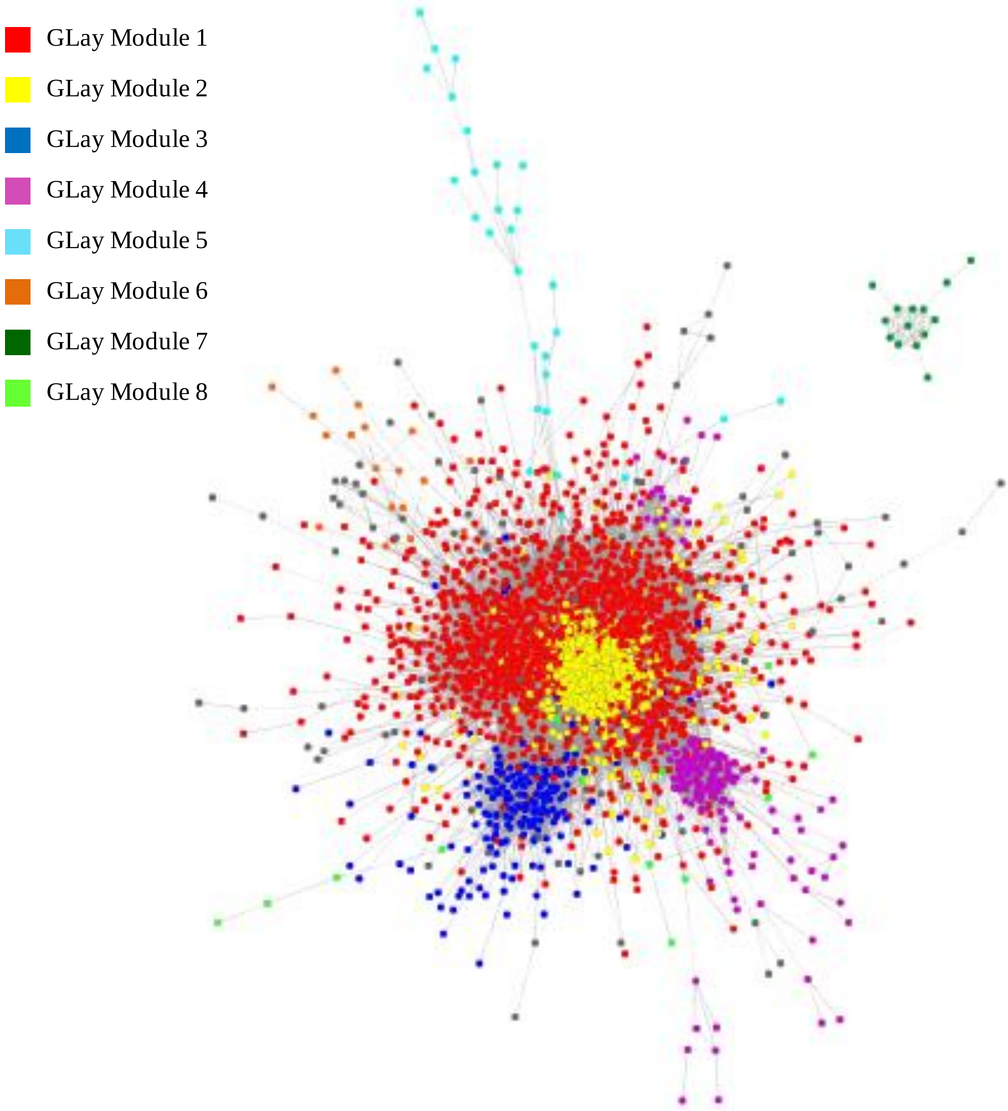


Figure 3. 1 Modules localise within distinct areas of the *S. cerevisiae* expression parent network.

An undirected network showing the interactions between genes from the *S. cerevisiae* transcription factor knockout data at 0.25MI (p-value: 10^{-23}) threshold. Force directed layout, with GLay modules mapped onto the parent network. Node colour represents GLay module. Edge length is representative of MI value. The accompanying table (Table 3.2) shows the breakdown of each module including the colour, number of nodes, and number of edges.

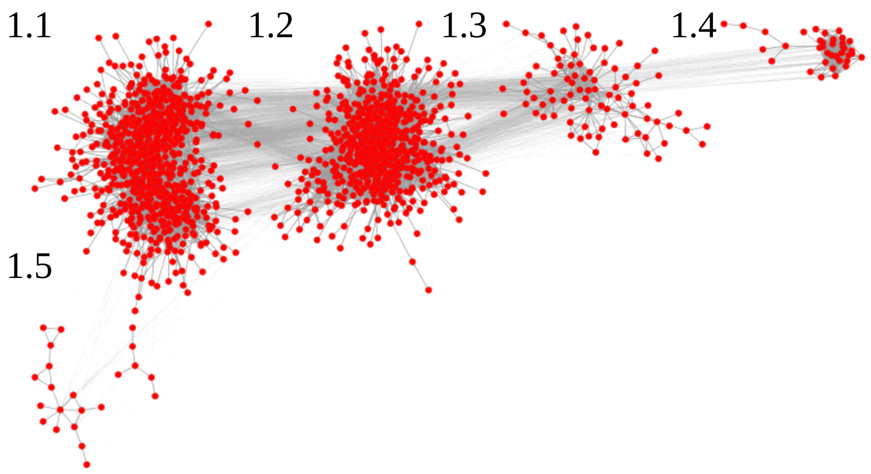
Module	Colour	Number of nodes	Number of Edges	No. of modules	Visualised in
All		3312	127528	8	Figure 3.1
1	Red	1608	33336	5	Figure 3.2
2	Yellow	728	70036	3	Figure 3.3
3	Blue	208	2539	4	Figure 3.4
4	Purple	176	4083	5	Figure 3.5
5	Light Blue	31	31	1	Figure 3.7
6	Orange	20	23	1	Figure 3.8
7	Dark Green	14	43	1	Figure 3.9
8	Light Green	13	12	1	Figure 3.10

Table 3. 2 The breakdown of *S. cerevisiae* expression modules defined by GLay.

3.3.1.1 Module 1: The transcriptional coupling of ribosome biogenesis, glycolysis and cell cycle processes

Module 1 is the largest of detected modules, and contains 1608 nodes and 33336 edges (Table 3.2). Five smaller interconnected sub-modules, identified after an additional round of modularisation, represented the fine structure of this module (Figure 3.2). The functional analysis of the components of this module revealed the association between the expression of ribosomal biogenesis genes, DNA damage and chromatin remodelling (sub-module 1.1), translational regulation and energy metabolism (sub-module 1.2), ribosome biogenesis (including eIF complex) and nuclear export (sub-module 1.3), helicase activity and telomere maintenance (sub-module 1.4) and mitochondrial nucleoid (sub-module 1.5).

Noteworthy is sub-module 1.2 which indicates that 19 glycolysis related genes are significantly correlated to translation regulation and the ribosomal genes from sub-module 1.1, these include enolase enzymes (*ENO1*, *ENO2*, *ERR1*, *ERR2*, *ERR3*), *FBA1*, *PGK1* and triose phosphate dehydrogenase enzymes (*TDH1*, *TDH2*, *TDH3*). Furthermore TDH1 was identified as significantly correlated to RPs in the *S. cerevisiae* fitness analysis. These results are consistent with reports that riboneogenesis is heavily dependent on glycolysis flux [12]. The grouping of ribosome biogenesis genes within module 1 is consistent with reports that ribosome synthesis is highly co-ordinated and tightly regulated [97] [141].



Module	Nodes	Edges	Functional analysis
1.1	754	13863	ribosome biogenesis (102), rRNA maturation (39), helicase (16), ribosome export (15), response to DNA damage (37), chromatin remodelling (27), cytoskeleton (43), telomere silencing (15), aldehyde dehydrogenase (5), hexose metabolism (9)
1.2	658	15836	Translation regulation (72), glycolysis (19), lipid synthesis (48), hydrolase (19), oxidative phosphorylation (12), aa biosynthesis (19), mitochondrion membrane (32), endoplasmic reticulum (114), cell wall biogenesis (29), membrane (301)
1.3	84	232	Ribosome biogenesis (21), multi eIF complex (3), nuclear export (8)
1.4	38	278	helicase activity (16), telomere maintenance via recombination (5)
1.5	16	19	mitochondrial nucleoid (3)

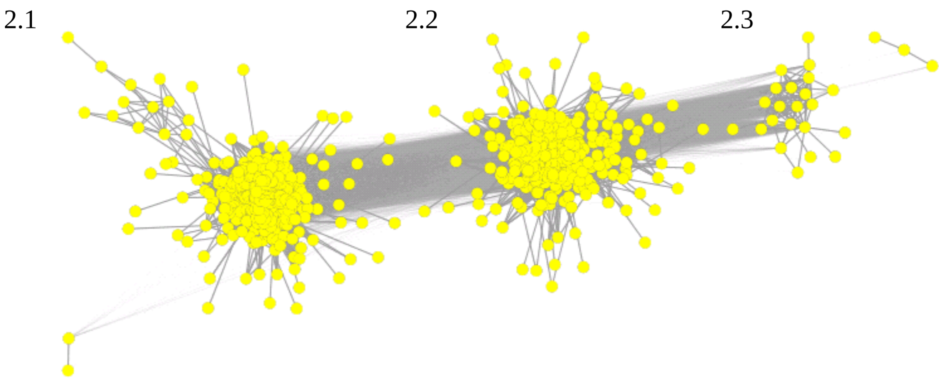
Figure 3. 2 Sub-modular structure of module 1, with accompanying functional analysis
Text colour represents significance of the functional enrichment (red: FDR < 0.05; green: FDR < 0.1; black: FDR >0.1).

3.3.1.2 Module 2: The transcriptional coupling of mitochondrial and cytosolic RPs protein transport and oxidative phosphorylation

Module 2 (Figure 3.3) maps directly into the centre of the parent network (Figure 3.1) and is strongly linked to module 1. Three smaller interconnected sub-modules were identified within this module (Figure 3.3). Similarly to module 1, functional analysis of each sub-module revealed interesting associations between the expression of RPs, and energy metabolism genes.

Mitochondrial RPs are associated to respiratory chain complex IV assembly, an important component of the oxidative phosphorylation pathway [142], golgi vescicle formation, a fundamental pathway involved in protein synthesis and post-translational processing [143] and ubiquitination, known to regulate the import of precursor proteins into the mitochondria [115] [144] (sub-module 2.1). An additional eleven genes representing other components of the oxidative phosphorylation pathway (cytochrome c oxidase and reductase subunits) are linked with cytosolic ribosomal proteins and sugar catabolism (sub-module 2.2). The smallest of the sub-modules (sub-module 2.3) is enriched of protein biosynthesis and transport. Alternative methods for the global analysis of expression data in *S. cerevisiae* also revealed a separation of cytosolic RPs and mitochondrial RPs, suggesting RPs are co-expressed based on cellular location as well as functionality [109].

Noteworthy is the separation of ribosomal biogenesis genes and RP genes (figure 3.2 and 3.3 respectively), representing two different modules of oxidative energy metabolism. Module 1 represents RPs linked to glycolysis, whilst module 2 represents RPs linked to oxidative phosphorylation. The switch between glycolytic and oxidative metabolism is essential for cell differentiation [94].

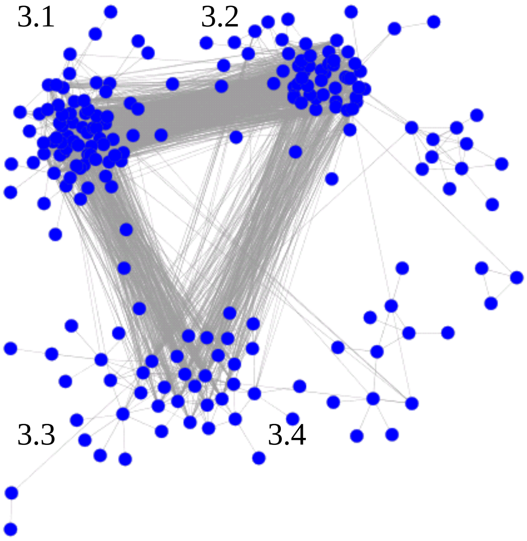


Module	Nodes	Edges	Functional analysis
2.1	352	21877	mitochondrial ribosome (29), respiratory chain complex IV assembly (9), iron transport (13), Proteasome (14), copper, ER (7), golgi vesicle transport (19), ubiquitination (14), lipoprotein, ribonucleoprotein core (7), mitochondrial intermembrane (9),
2.2	349	20300	cytosolic ribosome (100), translation regulation (36), rRNA binding (20), mitochondrial membrane (28), Ribosomal protein L7Ae/ L30e (6), respiratory chain (11), ribosome assembly (18), hexose catabolism (3)
2.3	22	81	protein biosynthesis (6), mitochondrion (3), transport (5)

Figure 3. 3 Sub-modular structure of module 2, with accompanying functional analysis. Text colour represents significance of the functional enrichment (red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR >0.1).

3.3.1.3 Module 3: Stress responses are transcriptional coupled to protein catabolism

Module 3 (Figure 3.4) is defined by four smaller interconnected sub-modules. Functional analysis of the sub-modules identified stress response genes (Figure 3.4). The association between temperature response and vacuolar protein catabolism demonstrated by sub-modules 3.1 and 3.3, temperature response and ubiquitin processing, (sub-module 3.2) and finally endocytosis and vesicle transport (sub-module 3.4). Noteworthy is the significant enrichment of DUP proteins, uncharacterised integral membrane proteins that contain internal duplication due to duplicated genes [145]. The function of these proteins are currently unknown, however my results suggest that they may have a role in heat shock response. The co-regulation of stress response genes was not observed to the same extent in my fitness analysis.

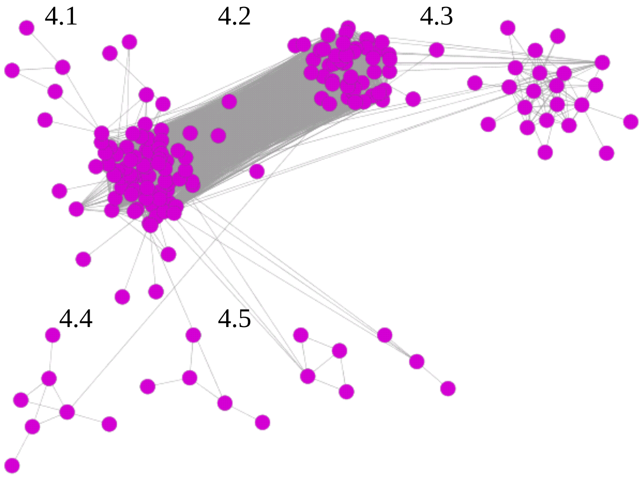


Module	Nodes	Edges	Functional analysis
3.1	78	744	response to temperature (13), vacuolar protein catabolism (11), ion channel activity (3); antiport (3), carbohydrate metabolism (3), cell membrane (9), protein kinase (7), polyamine transport (3)
3.2	70	604	ubl conjugation (11), response to temperature (16), membrane DUP (5), vacuolar protein catabolism (7), heat shock (3), ribosome (6)
3.3	43	130	response to temperature (14), vacuolar protein catabolism (8), membrane DUP (5), glycerol metabolic process (3), oxidoreductase (6)
3.4	12	12	Endocytosis (2), vesicle transport (4)

Figure 3. 4 Sub-modular structure of module 3, with accompanying functional analysis. Text colour represents significance of the functional enrichment (red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR >0.1).

3.3.1.4 Module 4: Transposable elements and translation are transcriptionally co-regulated

Module 4 contains 176 nodes and 4083 edges (Table 3.2). Five smaller interconnected sub-modules represented the finer structure of this module (Figure 3.5). Functional analysis of these sub-modules revealed the association between the expression of transposable elements (TEs) and ribosome frameshifting (sub-module 4.1), RNA mediated transposition and transmembrane (sub-module 4.2), transcription and cell cycle (sub-module 4.3), transcription (sub-module 4.4) and flocculation proteins (sub-module 4.5).



Module	Nodes	Edges	Functional analysis
4.1	86	1488	transposable element / ribosomal frameshifting (28) , oxidoreductase (4), RNAPII (5), mRNA metabolic process, cell cycle (7), translation (3)
4.2	49	873	Transposition - RNA-mediated (19), transmembrane (12)
4.3	22	84	DNA binding (4), cell cycle (4), mitochondrion (3), transcription (4).
4.4	7	8	regulation of transcription - DNA-dependent (4)
4.5	5	4	Flocculation protein (2)

Figure 3. 5 Sub-modular structure of module 4, with accompanying functional analysis. Text colour represents significance of the functional enrichment (red: FDR ≤ 0.05; green: FDR ≤ 0.1; black: FDR >0.1).

Notably, sub-modules 4.1 and 4.2 are connected by over 1600 edges (Figure 3.5), suggesting that despite being functionally similar, the separation of the transposable elements into two separate modules may mean different classes or mechanisms of transposition. TEs are known to influence gene expression at the transcriptional level

[146] [147], however the significant enrichment of ‘ribosomal frameshifting’ suggests that these genes shift the frame during translation, leading to the synthesis of a new protein product which may contain multiple open reading frames [148]. Further investigation determined that each module is enriched in different type of TE (Figure 3.6). Module 4.1 contains predominantly GAG type Tes, with of 19 / 28 being belonging to the Ty1 family, and the remaining 10 belonging to Ty2. Module 4.2 is enriched in GAG-POL type Tes from different families (Ty1: 17 / 19, Ty2: 1 / 19, Ty4: 1/19). Ty1 and Ty2 Tes are closely related [149] explaining why they are the most enriched TE type in the modules. Interestingly there is only one Ty3, and no Ty4 or Ty5 Tes, possibly reflecting the abundance of Ty1 and Ty2 within the *S. cerevisiae* genome [149]. *GAG* genes have a 7-bp frameshift signal located in close proximity to their stop codon, therefore upon translation, a frameshift occurs resulting in the synthesis of a GAG-POL fusion protein [150], hence why GAG genes are associated to ribosomal frameshifting (Figure 3.5).

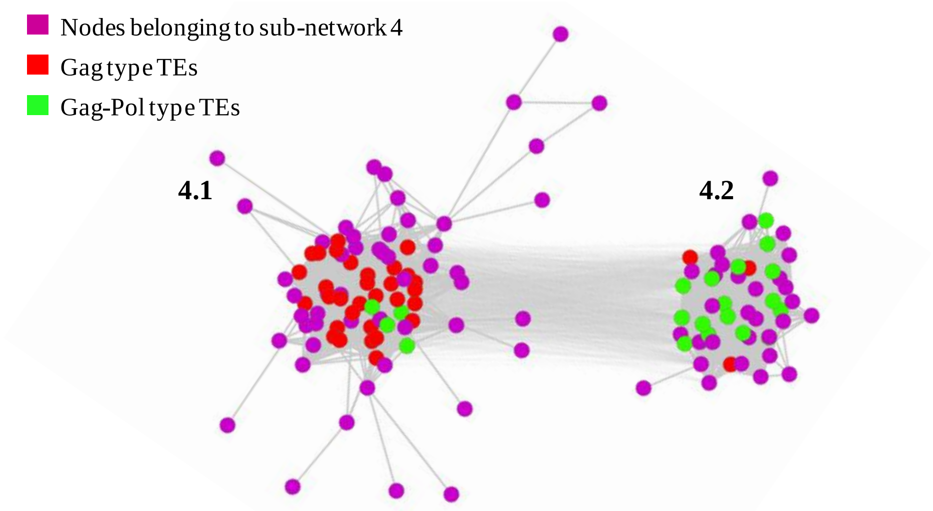
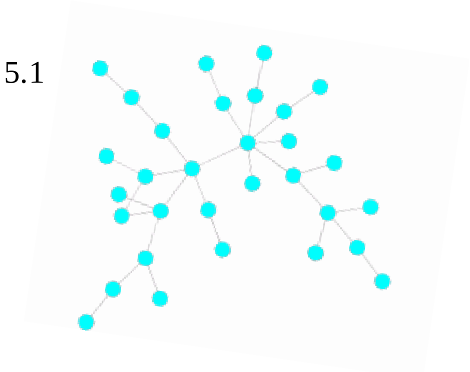


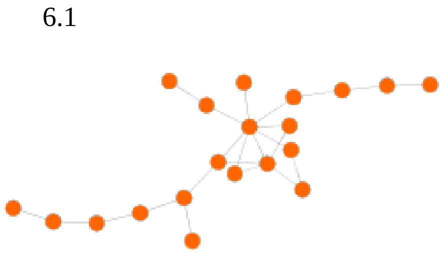
Figure 3. 6 The overrepresentation of GAG and Gag-POL type transposons in modules 4.1 and 4.2.

Nodes are coloured as shown in the figure legend. Module 4.1 has an overrepresentation of GAG type TEs, whilst module 4.2 has an overrepresentation of GAG-POL type TEs. This suggests that TE type can be distinguished by their expression



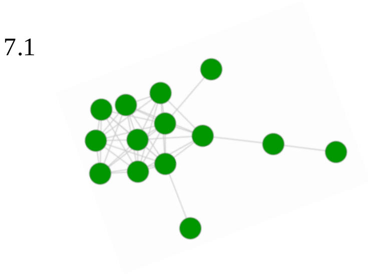
Module	Nodes	Edges	Functional analysis
5.1	31	31	mating pheromone activity (3), transcription regulation (6)

Figure 3. 7 Structure of module 5, with accompanying functional analysis.
Text colour represents significance (red: FDR ≤ 0.05; green: FDR ≤0.1; black: FDR >0.1)



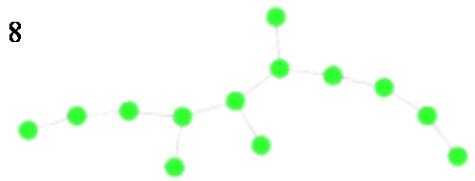
Module	Nodes	Edges	Functional analysis
6.1	20	23	Transcription (5), cell cycle (4), mitochondrion (3), membrane (3)

Figure 3. 8 Figure 3.8 Structure of module 6, with accompanying functional analysis.
Text colour represents significance (red: FDR ≤ 0.05; green: FDR ≤0.1; black: FDR >0.1)



Module	Nodes	Edges	Functional analysis
7.1	14	43	Transmembrane (7)

Figure 3. 9 Structure of module 7, with accompanying functional analysis.
Text colour represents significance (red: FDR ≤ 0.05; green: FDR ≤0.1; black: FDR >0.1)



Module	Nodes	Edges	Functional analysis
8.1	13	12	Metal ion binding (4), endomembrane system (4), non-membrane-bounded organelle (3)

Figure 3. 10 Structure of module 8, with accompanying functional analysis.
Text colour represents significance (red: FDR ≤ 0.05; green: FDR ≤0.1; black: FDR >0.1)

3.3.2 The link between cytosolic RPs, cell cycle and energy metabolism pathways is conserved at the gene expression level

Having demonstrated that cytosolic RPs are correlated to energy metabolism pathways at the phenotypic level, I applied the same analysis to the *S. cerevisiae* transcriptional network to determine if there was semblance to the functional modules identified within the fitness network (Chapter 2). The results suggest there is a degree of similarity between the *S. cerevisiae* fitness and expression networks. Once again, to aid in analysing the structure of each module, two levels of modularisation were performed. The first neighbour network of cytosolic RPs contained 1621 nodes and 112308 edges (Table 3.3). The first level of modularisation revealed three network modules (Figure 3.11).

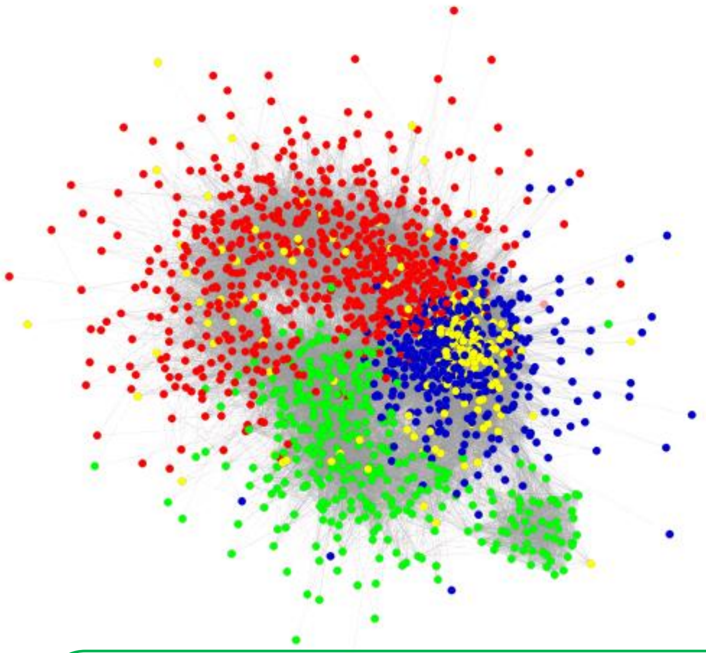
Module	Nodes	Edges
Overall	1621	112308
1.1	411	8848
1.2	290	13356
1.3	8	17
1.4	7	9
2.1	272	13531
2.2	227	15292
2.3	28	187
3.1	165	3786
3.2	141	2176
3.3	58	1427

Table 3. 3 Breakdown of cytosolic RP first neighbour modules identified by GLay.

Module 1 (represented in red) forms four smaller interconnected sub-modules. The functional analysis revealed association between the expression of ribosomal biogenesis, translation regulation and energy metabolism pathways (sub-module 1.1), ribosome biogenesis, tRNA processing and DNA repair (sub-module 1.2), nucleoplasm (sub-module 1.3) and rRNA process and metal ion binding (sub-module 1.4). Module 2 (represented in blue) forms three smaller interconnected sub-modules. Functional analysis revealed association between cytosolic ribosome, small ribosomal subunit and translation regulation (sub-module 2.1), mitochondrial ribosome, mitochondrial energy production and ubiquitin conjugation (sub-module 2.2), and cytosolic ribosome and response to pheromone (sub-module 2.3). Finally, module 3 (represented in green) consists of three smaller sub-modules. Functional analysis revealed association between the expression of translation regulation, membrane and glucose metabolism genes (sub-module 3.1), endoplasmic reticulum and glycolysis / gluconeogenesis (sub-module 3.2) and transposable elements and rRNA processing (sub-module 3.3).

Noteworthy, are the direct interactions between ribosomal proteins and cytoplasmic energy metabolism pathways, including riboneogenesis gene, ribulose-5-phosphate isomerase (*RKI1*) which is directly connected to 54 nodes within sub-module 1.1, 25 of which are related to ribosome biogenesis (FDR 9.8x10^{-23}). Enolase enzymes (*ERR1*, *ERR2*) and pyruvate kinase (*PYK1*) located in sub-module 1.2 all have direct interactions with cytosolic RPs. Interestingly *GPM1* and *TPI1* located in sub-module 2.1 and involved in glycolysis have direct connections to only large ribosomal proteins. *TKL1* and *TAL1*, located in module 3.1 and 3.2 respectively are enzymes essential for riboneogenesis. TKL1 especially, as it is the primary source of transketolase activity in *S. cerevisiae* [92]. Deletion of *TAL1* has been reported to affect flux through SHB17 [92]. *TAL1* expression is anticorrelated with *TKL1* expression throughout the yeast metacycle [92], their localisation within separate modules may represent their anticorrelation. A subset of TEs are direct neighbours of cytosolic RPs (sub-module 3.3), consistent with the role of *GAG* genes causing ribosome frameshifts [150]. Smith *et al* reported that transcriptional silencing in *S. cerevisiae* ribosomal DNA (rDNA) can be caused by Ty1 retrotransposons integrating into rDNA, targeting upstream of the RNAPIII transcribed 5s-rRNA genes [151].

Finally, cytosolic RPs and mitochondrial RPs are located within the same module, however, are separated into different sub-modules (2.1 and 2.2 respectively). Mitochondrial RPs, though functionally similar to cytosolic RPs have different features and primary structures. The assembly of functionally active mitochondrial ribosomes depends on the co-expression of both mitochondrially localised and nuclear localised genes [152], thus explaining the linkage between cytosolic and mitochondrial RPs.



🟥 Cytosolic RP Module 1

🟦 Cytosolic RP Module 2

🟩 Cytosolic RP Module 3

🟨 Cytosolic Ribosomal Protein

1.1) ribosome biogenesis (118), ribosomal large subunit biogenesis (31), preribosome (58), ribosome localization (18), RNA helicase activity (18), RNA modification (41), wd repeat (22), nucleotide-binding (85), Aminoacyl-tRNA biosynthesis (12), translation regulation (31), snoRNA 3'-end processing (11), biopolymer methylation (14), transcription (56), nuclear export (3), Tetratricopeptide region (8), mRNA processing (21), transcription from RNAPI promoter (8), RNA polyadenylation (7), chaperone (5), respiratory chain complex II (3), glucose metabolism (3)

1.2) intracellular organelle lumen (76), ribosome biogenesis (41), ESCRT III complex (4), tRNA processing (11), Small GTP-binding protein (11), Nucleotide excision repair (7), ARF/SAR superfamily (4), RNA degradation (10), Glycolysis (3), hexose metabolism (3)

1.3) nucleoplasm part (4)

1.4) rRNA processing (5), metal ion processing (3)

2.1) cytosolic ribosome (109), regulation of translation (43), cytosolic small ribosomal subunit (6) (45), rRNA binding (15), ribosome assembly (20), Ribosomal protein L7A, rRNA export (14), preribosome (20), zinc (8), glucose metabolism (3)

2.2) mitochondrial (55), mitochondrial ribosome (16), mitochondrial respiratory chain complex assembly (7), cellular protein complex assembly (16), proteosome complex (8), er-golgi transport, metallochaperone activity (4), Ubiquitin conjugation (11)

2.3) cytosolic ribosome (5), response to pheromone (3), metal-binding (3), transport (7)

3.1) translation regulation (28), transmembrane protein (60), atp-binding (37), glucose metabolism (9), lipid synthesis (8), glycoprotein (30), molecular chaperone, cell wall (10), cation transport (13), amino-acid transport, (5) endomembrane system (21),

3.2) endoplasmic reticulum (34), cell wall (19), Glycolysis / Gluconeogenesis (12), golgi apparatus (14), lipid biosynthetic process (14), cell cycle (10)

3.3) transposable element (33), DNA integration (11), rRNA processing (3)

Figure 3. 11 An undirected network showing the first neighbours of cytosolic RPs.
The network has been visualised using a force directed layout, with modules identified by GLay mapped onto the parent network (see legend). Edge lenth represents MI score. Yellow nodes represent cytosolic RPs. Accompanying annotation indicates functional enrichment defined by DAVID. Red text indicates an adjusted FDR of ≤ 0.05, and green indicates a FDR of ≤0.1. Black represents non-significant enrichment.

3.3.3 Mitochondrial RPs are transcriptionally coupled to genes encoding cytosolic RPs, respiratory chain and ubiquitin complexes.

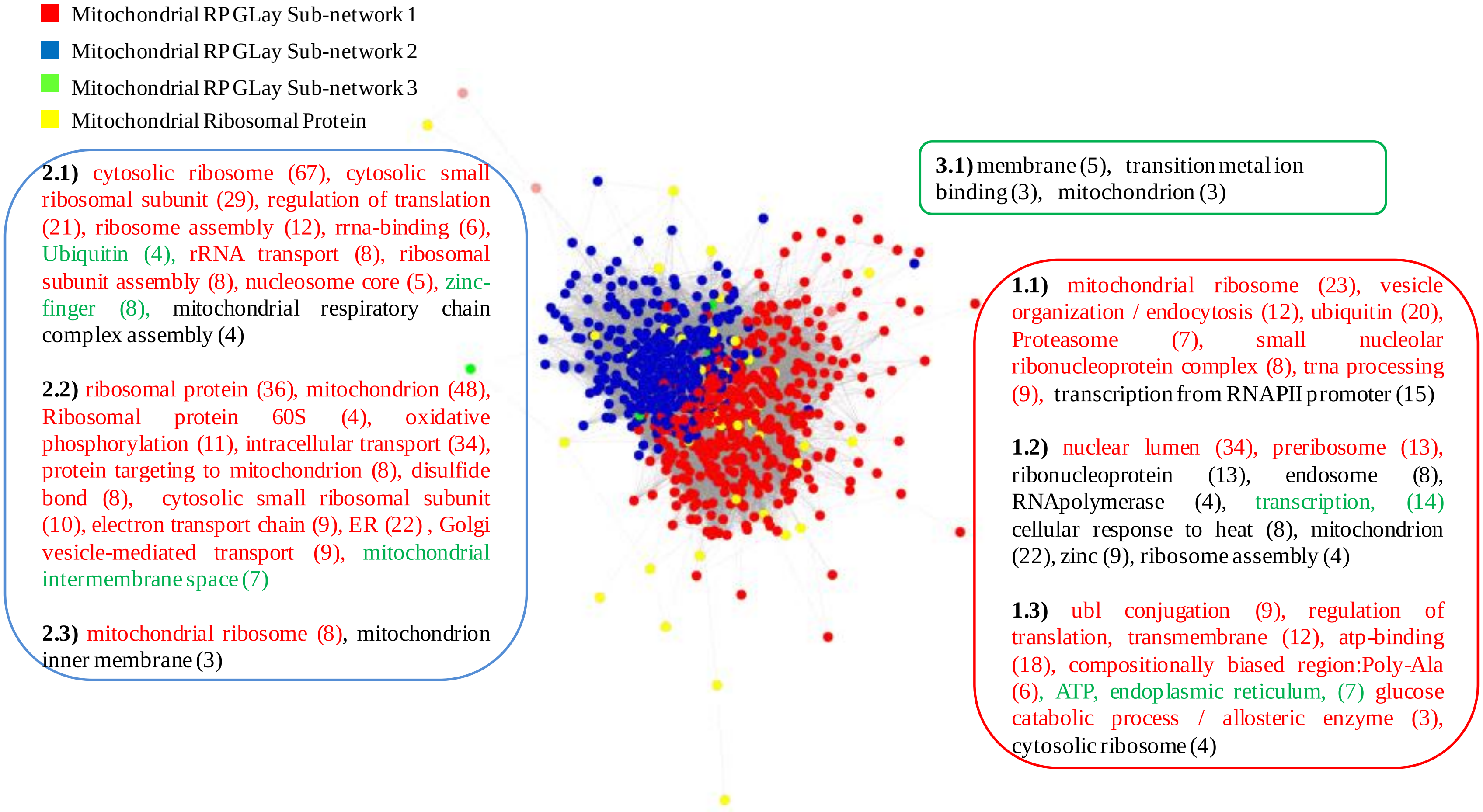
The mitochondrial RP first neighbour network contains 846 nodes and 83958 edges (Table 3.4). The first level of modularisation revealed three network modules (Figure 3.12).

Module 1 (represented in red) forms three sub-modules, functional analysis identified an association between mitochondrial ribosome, ubiquitin and endocytosis (sub-module 1.1), nuclear lumen and preribosome (sub-module 1.2) and ubiquitin machinery, transmembrane, translation regulation and glucose catabolism (sub-module 1.3). Module 2 (represented in blue) forms three smaller interconnected sub-modules. Functional analysis revealed association between the expression of cytosolic ribosomes, translation regulation and ribosomal assembly (sub-module 2.1), mitochondrial ribosomal proteins, protein transport and electron transport chain (sub-module 2.2), and mitochondrial ribosome and mitochondrial inner membrane (sub-module 2.3). Module 3 could not undergo a further level of modularisation, functional analysis identified association between membrane, transition metal ion binding and mitochondrion.

Noteworthy are sub-modules 2.1 and 2.2 which demonstrate the co-expression of mitochondrial RPs and genes involved in the electron transport chain. Passage through the electron transport chain provides the energy for both cytosolic and mitochondrial translation. Furthermore most mitochondrial translation products form part of the membrane embedded centres present within the respiratory chain complexes [153].

Module	Nodes	Edges
Overall	846	83958
1.1	230	10899
1.2	135	3031
1.3	74	1199
2.1	195	14371
2.2	164	5847
2.3	16	77
3.1	7	6

Table 3. 4 Breakdown of mitochondrial RP first neighbour sub-modules identified by GLay.



🟥 Mitochondrial RP GLay Sub-network 1

🟦 Mitochondrial RP GLay Sub-network 2

🟩 Mitochondrial RP GLay Sub-network 3

🟨 Mitochondrial Ribosomal Protein

2.1) cytosolic ribosome (67), cytosolic small ribosomal subunit (29), regulation of translation (21), ribosome assembly (12), rrna-binding (6), Ubiquitin (4), rRNA transport (8), ribosomal subunit assembly (8), nucleosome core (5), zinc-finger (8), mitochondrial respiratory chain complex assembly (4)

2.2) ribosomal protein (36), mitochondrion (48), Ribosomal protein 60S (4), oxidative phosphorylation (11), intracellular transport (34), protein targeting to mitochondrion (8), disulfide bond (8), cytosolic small ribosomal subunit (10), electron transport chain (9), ER (22) , Golgi vesicle-mediated transport (9), mitochondrial intermembrane space (7)

2.3) mitochondrial ribosome (8), mitochondrion inner membrane (3)

3.1) membrane (5), transition metal ion binding (3), mitochondrion (3)

1.1) mitochondrial ribosome (23), vesicle organization / endocytosis (12), ubiquitin (20), Proteasome (7), small nucleolar ribonucleoprotein complex (8), trna processing (9), transcription from RNAPII promoter (15)

1.2) nuclear lumen (34), preribosome (13), ribonucleoprotein (13), endosome (8), RNApolymerase (4), transcription, (14) cellular response to heat (8), mitochondrion (22), zinc (9), ribosome assembly (4)

1.3) ubl conjugation (9), regulation of translation, transmembrane (12), atp-binding (18), compositionally biased region:Poly-Ala (6), ATP, endoplasmic reticulum, (7) glucose catabolic process / allosteric enzyme (3), cytosolic ribosome (4)

Figure 3. 12 An undirected network showing the first neighbours of mitochondrial RPs.
The network has been visualised using a force directed layout, with modules identified by GLay mapped onto the parent network (see legend). Yellow nodes represent mitochondrial RPs. Edge length represents MI score. Accompanying annotation indicates functional enrichment defined by DAVID. Red text indicates an adjusted FDR of $\leq$ 0.05, and green indicates an FDR of $\leq$ 0.1. Black represents non-significant enrichment

3.4 Discussion

In this chapter, an MI based reverse engineering method has been applied to a comprehensive genome-wide *S. cerevisiae* TF knockout dataset. Similarly to Chapter 2, the analytical pipeline used in this study allowed for the utilisation and mapping of the entire dataset rather than a specific subset of genes. The aim was to create a network which encapsulated the response of all the genes within the dataset. This type of network inference approach has not previously been attempted on this dataset. Using network modularisation algorithms it was possible to identify genes that had similar responses across all TF knockouts. Below, I discuss how these network interrogation methodologies were able to identify and elucidate some of the more interesting and statistically significant correlations between groups of genes. I discuss particularly the linkage between RPs and energy metabolism genes and how these results obtained by bioinformatical methods correlate strongly to those reported by Clasquin *et al,* whose results are based on wet lab experiments only **[**92**].**

3.4.1 The highly coordinated expression of genes encoding ribosome factors, glycolysis, and cell cycle

The linkage between RPs, glycolysis and cell cycle was also observed in the *S. cerevisiae* fitness networks (Chapter 2). The metacycle has been reported to link these cellular processes together in space and time, where essential cellular processes and metabolic events occur in synchrony [94]. Figure 3.13 represents a modified schematic of the oscillating nature of genes involved in the metacycle as reported by Tu *et al* [94]. Each cycle contains a reductive non-respiratory phase, split into building (Figure 3.13, green) and charging (Figure 3.13, blue), and an oxidative respiratory phase (Figure 3.13, red). Studies in *S. cerevisiae* identified that cytoplasmic RPs and genes involved in translation have a very similar expression pattern across each metacycle phase, as does 73 / 74 mitochondrial RPs and mitochondrial related genes [94]. Tu *et al* reported that during the metacycle, the expression of energy metabolism genes was at its peak when expression of cytosolic RPs, ribosome biogenesis, translation initiation, and amino acid biosynthesis were at their lowest (Figure 3.13) [94], in agreement with the dynamics of riboneogenesis [92]. Translation is one of the most energy costly processes [154]; therefore the translation machinery would be assembled when there are excess amounts of ATP available. Hence why there is a transient peak in RP expression shortly after (within hours) of the peak of energy metabolism expression which quickly dissipates before the onset of the non-respiratory phase, presumably due to the lack of oxidisable metabolites [94] and glycolytic intermediates [92], This suggests flux through the riboneogenesis pathway immediately dissipates upon the depletion of the previously stored glycolytic intermediates. Genes involved in glycolysis and other carbohydrate metabolism genes peak during the charging phase increasing the concentration of acetyl-CoA and the glycolytic intermediates fructose-6-phosphate, glyceraldehydes-3-phosphate and dihydroxyacetone-phosphate.

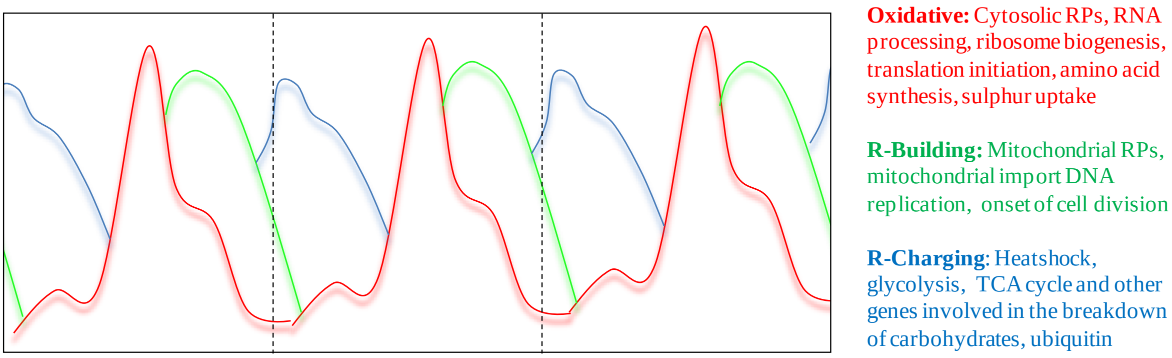


Figure 3. 13 The oscillation of functional groups in the yeast metacycle.
A cartoon based on the data presented in Tu *et al* [94]. Coloured lines represent stages of the metacycle, dashed vertical lines represent a complete cycle. Each phase has a distinct stage within the metacycle which is characterised by the up-regulation of a specific group of genes, represented by the legend to the right of the figure.

DNA replication and cell division genes are up-regulated in the reduction building phase together with mitochondrial RPs. The expression of cell cycle and DNA replication genes during the reductive non respiratory phase of the metacycle may allow cells to negate oxidative damage to the DNA which would occur during the oxidative phase, a feature observed in other species [104]. Many essential cellular functions including respiration, ribosome biogenesis, DNA replication, cell division and glycolysis are all compartmentalised in accordance to the metacycle. This type of cellular organisation would minimise wasted reactions and make efficient use of available metabolites, especially as ribosome biogenesis uses a lot of energy.

The network analysis also highlights a close correlation between cytosolic RPs and mitochondrial RPs. In figure 3.13 the expression mitochondrial RPs and mitochondrial biogenesis peaks shortly after the peak in cytosolic RP expression. This observation may be explained as many mitochondrial precursor proteins are synthesised by cytosolic ribosomes prior to being imported into the mitochondria via specialised translocation machinery. Once imported they are utilised in various critical mitochondrial functions such as mitochondrial biogenesis and energy production [117]. Therefore, if there is a peak in expression of cytosolic RPs prior to mitochondrial RPs, it suggests the precursor proteins targeted for the mitochondria have reached their destination and are now driving mitochondrial processes, including translation localised to the mitochondria.

Though the metacycle was reported decades ago [105], the regulatory mechanisms which control each stage are still being investigated. What is clear is that the duration of each metabolic phase is almost identical [104] [94]. However, currently, the duration of a single metabolic cycle remains controversial, with reports stating a single cycle is ~40 mins [104] or ~300 minutes [94]. So although the results in both the fitness (Chapter 2) and expression data support the reported complementary nature of riboneogenesis and metacycle [92], this study is remains limited in two ways. Firstly, the fact that current literature has yet to establish the finer details and complexities of the metacycle means that interpreting network interactions regarding the metacycle may not be completely trustworthy. Secondly, the lack of experimental work to strengthen the results presented in this chapter. Reverse engineering and network inference methods are advantageous as they provide a means of mapping a biological system without prior knowledge. Numerous bioinformatic tools exist to help simplify, understand and interrogate networks in order to identify and generate hypotheses. Although it is possible to draw support from literature and existing studies to strengthen these hypotheses, the fact remains that causality and whether gene correlations are representative of true biological relationships, can only be obtained by directly testing the hypothesis in question via experimental validation.

3.4.2 The linkage of mitochondrial RPs to cytosolic RPs and ubiquination machinery is conserved at the expression and phenotypic level

The mitochondrial RP first neighbour network shows a close correlation to cytosolic RPs. This is likely because the majority of mitochondrial proteins are synthesised on cytosolic ribosomes before being imported into the mitochondrial intermembrane space [117], hence their localisation within the same sub-network. This is consistent with the significant enrichment of ‘intracellular transport / protein targeting to the mitochondrion’ and other transit peptide functions within the same module as cytosolic RPs (Figure 3.12, module 2).

The co-expression between mitochondrial RPs and ubiquitin-dependent protein catabolic processes, ubl conjugation, proteosome functions within modules 1.1 and 1.3 (Figure 3.12) is a feature shared with the *S. cerevisiae* fitness data (module 3.3, Figure 2.23). Ubiquitin modification is required during the import of mitochondrial precursor proteins from the cytosol to their final destination within the mitochondria (mitochondrial matrix, intermembrane space, inner mitochondrial membrane etc) [115] [144]. A study published in March 2013 investigated the cytosolic biogenesis of proteins whose final location was the intermembrane space of mitochondria. The study identified that ubiquitin-proteasome machinery is responsible for the removal of mis-localised intermembrane space proteins [117]. The proteins that reside within the intermembrane space are critical for metabolic functions, regulatory processes (including mitochondrial transport) and mitochondrial biogenesis [155] [156]. Furthermore the ubiquitin proteasome machinery is utilised in the cytosol also, by acting as a negative regulator in the biogenesis of proteins that localise to the intermembrane space, therefore maintaining protein homeostasis in circumstances in which mitochondria may lose their integrity [117]. These results may aid in understanding the co-expression and co-fitness between ubiquitination and mitochondria related genes.

3.4.3 Future work

Prior to this analysis, the most popular expression dataset used in gene expression network construction for *S. cerevisiae* was the dataset published by Hughes *et al* [67] which has been cited almost 2400 times since being published in 2000. This work used a different, yet still comprehensive TF knockout gene expression dataset that has not been utilised in MI based network analysis. Genome-wide networks of this nature are robust and can be utilised in a number of diverse ways. This chapter primarily focused on the linkage between ribosome biogenesis and energy metabolism. The network interrogation techniques applied in this chapter could easily be shifted to another biological process or pathway of interest. Furthermore, the network constructed in this chapter may act as a foundation for mapping the impact of specific TF knockouts on genome expression. One possible approach would be to construct networks using only TF knockouts of interest, with the networks representing the underlying biological relationships of genes in response to specific TF knockouts.

3.5 Concluding remarks

The work presented in this chapter reported an MI based network constructed solely on TF knockout data from *S. cerevisiae*. One of the key aims of this chapter was to identify and characterise functional modules using this TF knockout dataset. The modules identified in this network are consistent with current biological knowledge and the organisation of the yeast system, such as the modularisation of ribosomal genes due to their highly regulated nature. Interpreting whether the network accurately represents biological processes of higher complexity such as the metacycle is harder to evaluate due to lack of experimental evidence and supporting literature.

This chapter also highlights how considering the entire dataset for network inference can be used to identify potentially new interesting areas of research, one such example, was the ability to distinguish between GAG and GAG-POL TEs based on their localisation within the gene expression network. Furthermore, the characterisation of functional modules in this gene expression network provided a foundation for analysing the dependencies between genes involved in riboneogenesis. Similarly to the fitness network reported in Chapter 2, this chapter marks the first contribution of using a gene expression network to aid in understanding the relationships between genes involved in ribosome biogenesis and energy metabolism.

CHAPTER 4: NETWORK INFERENCE AND ANALYSIS USING AN INTEGRATED FITNESS AND EXPRESSION DATASET

4.1 Introduction

I have shown that reverse engineering biological networks from expression profiling or fitness data can provide insights into important biological processes. Utilising a single type of dataset, may, however limit the robustness of the analysis. In order to further understand a biological system, the integration of multiple types of data is needed, each of which can provide additional information thereby increasing the comprehensiveness and sensitivity of the network [88] [89] [90] [91].

In this chapter, I report the integration of fitness and expression datasets previously analysed in Chapters 2 and 3. The overarching aim is to develop a high level network representing the correlation between gene expression profiles within the modular structure defined within the fitness network developed in Chapter 2.

One of the outcomes of this chapter was the discovery that biological functions represented as modules within the fitness network are also transcriptionally coupled. This suggested that the common behaviour of genes shared across the fitness and expression datasets reflected a specific function shared by the proteins that encode these genes. Several interesting hypothesis were identified during this analysis including, the anti-correlated transcription profiles between genes encoding ribosome biogenesis / cell cycle proteins and stress response, suggesting that ribosome biogenesis and cell cycle processes are repressed during stress response. I also identified that genes encoding ribosomal proteins (RPs) and chromosome segregation modularise together, suggesting that they exhibit co-fitness and co-expression. Finally, I show that applying an extremely high threshold to the integrated network, identifies modules representing the co-expression of genes within specific phases of the metacycle, and I demonstrate that genes encoding glycolytic enzymes are strongly correlated at the phenotypic and transcriptional level to genes involved in rRNA processing and the onset of ribosome biogenesis.

4.2 Methods

Technology development in the last decade has enabled most laboratories to acquire genome-wide functional genomics datasets. Consequentially, amount of publically available genome-wide expression data has increased exponentially. The numbers of growth fitness and proteomic datasets have also increased, although not to the same extent as gene expression datasets [157] [158] [48] [159]. The aim is to integrate expression and fitness data into a single dataset, and then reverse engineer a network representative of the correlation between gene expression profiles, within the modular structure defined by the previous fitness network analysis,

4.2.1 The datasets

In order to construct an integrated network, I used the Hillenmeyer sub-module information identified from my *S. cerevisiae* fitness analysis (Chapter 2, sections 2.3.3.1 – 2.3.3.8). A total of 21 sub-modules were identified within the Hillenmeyer fitness network (Chapter 2, table 2.6). However due to the incredibly small size of the sub-modules identified in module 4 (Chapter 2, table 2.8), it was therefore classified as a single large sub-module. Therefore the modular structure of 17 sub-modules was used in this analysis, which still represented all the sub-modules identified within the fitness data.

The gene expression data used for this study was the pre-processed and normalised TF knockout data published by Reimand *et al* [71] which was used to construct the *S. cerevisiae* expression network. Details of the expression dataset are given in Chapter 3, section 3.2.1.

4.2.2 Analysis strategy

The pipeline required to integrate the fitness and expression datasets, and analyse the network is shown below.

1. For each fitness sub-module, the gene IDs were identified (Figure 4.1A), and used to extract the corresponding gene expression measurements from the TF knockout dataset (Figure 4.1B)
2. The clustering tool HOPACH [15] was used to group expression data across all 269 TF knockout samples (Figure 4.1C). This was done in order to identify groups of genes within each sub-module that exhibited transcriptionally coupled behaviour. Each HOPACH cluster represented a phenotypic outcome based on the genes expressed within it. A total of 90 HOPACH clusters were identified within the 17 fitness sub-modules.
3. Each of the 90 HOPACH clusters were functionally annotated using DAVID [21] to identify if phenotypically and transcriptionally linked genes were representative of cellular processes.
4. The average profile of each HOPACH cluster was calculated. Therefore, each cluster was represented as in a single vector of gene expression measurements.
5. The averages of each HOPACH cluster were merged to create an integrated dataset containing 90 rows (representing each HOPACH cluster) and 269 columns (representing each TF knockout from the expression data)
6. The reverse engineering method ARACNE [30] was used to calculate the mutual information between every HOPACH cluster.
7. The network was thresholded and visualised within Cytoscape [140] using a force directed layout (Figure 4.1D).
8. The network was modularised using GLay [33]

9. Hierarchical clustering (HCL) was performed on each module, to determine if nodes within that module exhibited transcriptional coupling.

The interaction between fitness sub-modules and co-transcription is based on visual analysis. Each node within the network represents a group of genes which exhibit similar growth fitness when mutated, and which are also transcriptionally correlated. A group of nodes within a module represents cellular functions that are transcriptionally co-regulated upon transcription factor knockout, and that when mutated, exhibit the similar phenotype (Figure 4.1F). Applying HCL to each module shows visually the transcriptional coupling of the processes within that module, i.e. if they are correlated or anticorrelated, and provides the foundation for building hypotheses.

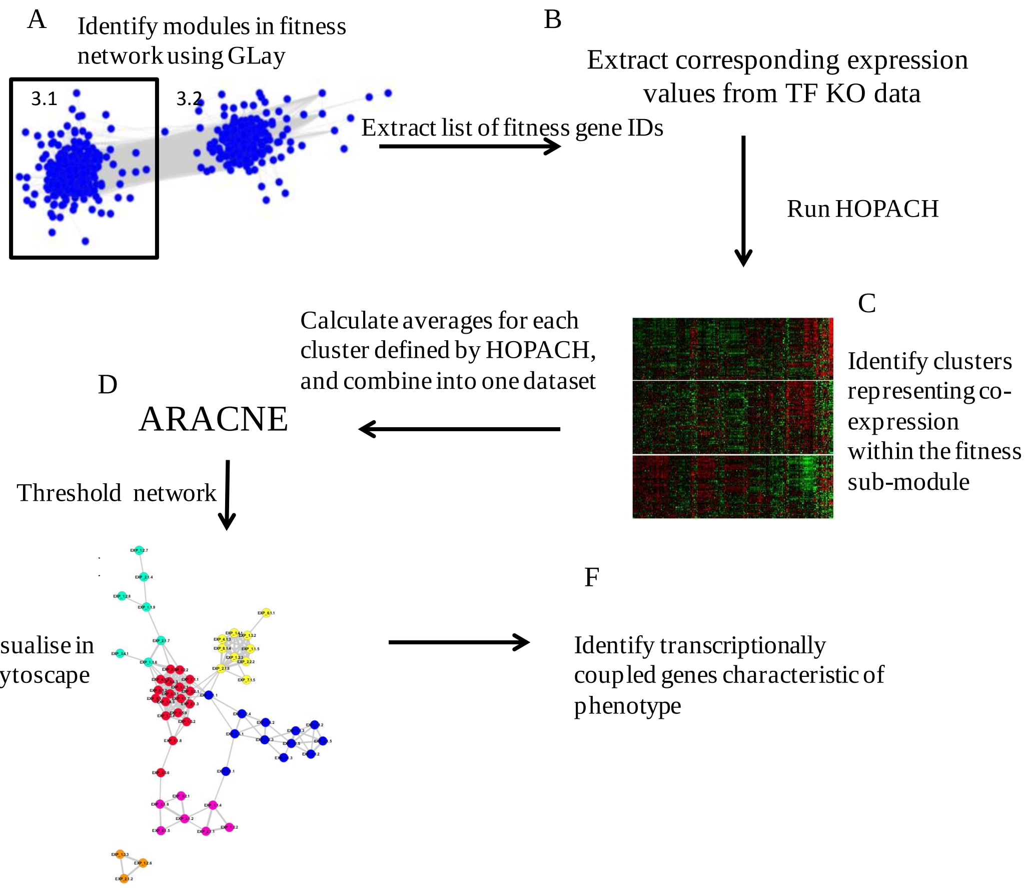

A
Identify modules in fitness network using GLay
3.1
3.2
Extract list of fitness gene IDs
B
Extract corresponding expression values from TF KO data
Run HOPACH
C
Identify clusters representing co-expression within the fitness sub-module
Calculate averages for each cluster defined by HOPACH, and combine into one dataset
D
ARACNE
Threshold network
Visualise in Cytoscape
F
Identify transcriptionally coupled genes characteristic of phenotype

Figure 4. 1 Integration of fitness and expression network data workflow.
A sample workflow detailing how a fitness module would be integrated with the TF expression data. Panel A. Gene IDs within each sub-module of the fitness data are identified. Panel B. Gene IDs are used to extract the corresponding gene expression values from the TF knockout dataset. Panel C. HOPACH is used to identify genes with similar expression across all samples. Each HOPACH cluster is represented by the average expression profile. Therefore each HOPACH cluster represents a group of genes that exhibit a similar phenotype when mutated, and which are also co-regulated. Panel D.The averages of each HOPACH were merged into a single dataset and used in ARACNE. Panel E. The network is thresholded, visualised and interrogated to identify modules. Each node represents a HOPACH cluster, which is the integration of structure properties of the fitness network and the expression measurements of the expression dataset. Panel F. Genes that are correlated at both the phenotypic and transcriptional level are identified.

4.2.3 HOPACH clustering of integrated data

Gene IDs from each of the 17 Hillenemeyer fitness sub-modules were used to extract the corresponding expression values from the TF knockout expression dataset, this strategy therefore integrated the modular structure of the fitness network with the expression measurements of the transcription dataset. HOPACH was used to group the expression measurements of each fitness sub-module to identify genes within similar expression profiles. A total of 90 HOPACH clusters were identified from the 17 fitness sub-modules (Table 4.1). HOPACH was performed in the statistical programming language R [11]. Visualisation of the HOPACH clusters demonstrates that genes with a similar phenotype can be grouped by co-expression, and that several clusters of co-expressed genes are identified from a single fitness sub-module (Figures 4.2 – 4.5).

Fitness module	Gene Count	Found in TF KO dataset	HOPACH clusters	Genes per HOPACH Cluster
1.1	684	678	9	68 / 79 / 84 / 62 / 95 / 62 / 96 / 69 / 63
1.2	481	481	8	109 / 31 / 108 / 73 / 64 / 41 / 21 / 34
1.3	199	198	4	45 / 42 / 31 / 80
1.4	192	191	3	59 / 71 / 61
1.5	16	16	2	4 / 12
1.6	7	7	4	2 / 1 / 1 / 3
2.1	300	298	8	42 / 28 / 54 / 25 / 53 / 21 / 26 / 49
2.2	241	240	3	77 / 68 / 95
3.1	78	77	4	51 / 18 / 3 / 5
3.2	68	67	2	19 / 48
3.3	49	49	5	18 / 12 / 5 / 6 / 8
3.4	28	28	5	4 / 3 / 18 / 2 / 1
3.5	15	15	9	1 / 4 / 1 /2 / 1 / 2 / 1 / 1 / 2
4.1	53	53	5	10 / 12 / 12 / 12 / 7
5.1	28	26	7	8 / 4 / 3 / 2 / 4 / 4 / 1
6.1	25	25	2	18 / 7
7.1	11	11	5	3 / 3 / 2 / 1 / 2
8.1	11	11	5	3 / 2 / 1 / 2 / 3

Table 4. 1 Summary of the HOPACH clusters identified from the fitness sub-modules. Column 1 represents the fitness sub-module from the Hillenmeyer network. Column 2 shows the number of genes within that fitness sub-module. Column 3 shows the number of genes from the fitness sub-module that were identified within the TF knockout expression dataset. Column 4 represents the number of clusters HOPACH identified when it clustered the integrated data. Column 5 details how many genes were identified in each of the HOPACH clusters. Cell colour is a visual aid to identify fitness sub-modules.

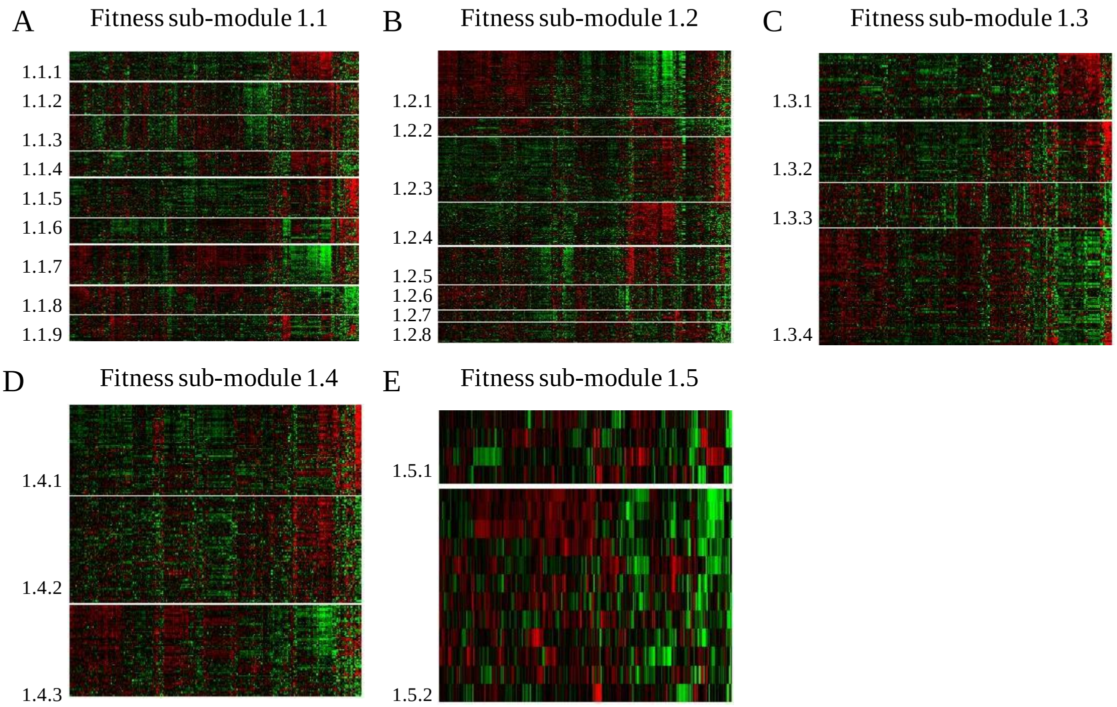


Figure 4. 2 Heatmaps of expression clusters identified by HOPACH from fitness module 1. The HOPACH clusters identified within each fitness sub-module represent a group of genes which exhibit similar growth fitness when mutated, and which are also transcriptionally correlated. The x-axis represents the 269 TF knockouts from the expression dataset. The y-axis represents the genes within that cluster, derived from the original fitness sub-module. Red represents increased gene expression, green represents decreased gene expression. Each HOPACH cluster is annotated in respect to the fitness sub-module it was derived from, and the number of HOPACH clusters within that sub-module. For example, HOPACH cluster 1.1.5, represents fitness sub-module 1.1, HOPACH cluster 5.

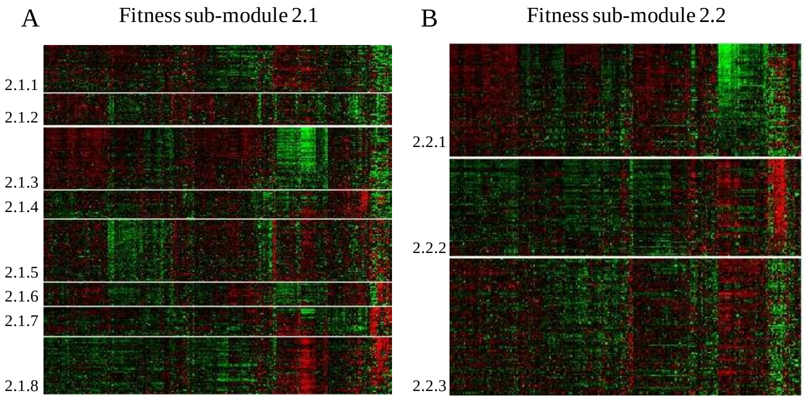


Figure 4. 3 Heatmaps of expression clusters identified by HOPACH from fitness module 2. Each HOPACH cluster is annotated in respect to the fitness sub-module it was derived from, and the number of clusters within that network. The HOPACH clusters identified within each fitness sub-module represents a group of genes which exhibit similar growth fitness when mutated, and which are also transcriptionally correlated. The x-axis represents the 269 TF knockouts from the expression dataset. The y-axis represents the genes within that cluster, derived from the original fitness sub-module. Red represents increased gene expression, green represents decreased gene expression. Each HOPACH cluster is annotated in respect to the fitness sub-module it was derived from, and the number of HOPACH clusters within that sub-module.

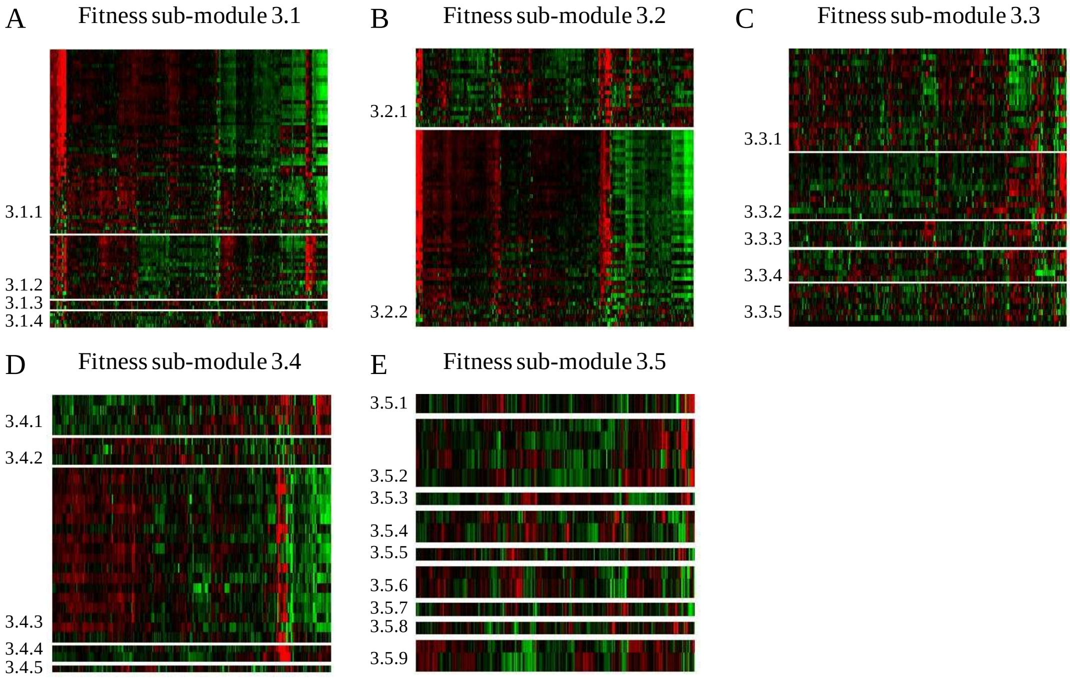


Figure 4. 4 Heatmaps of expression clusters identified by HOPACH from fitness module 3. Each HOPACH cluster is annotated in respect to the fitness sub-module it was derived from, and the number of clusters within that network. The HOPACH clusters identified within each fitness sub-module represents a group of genes which exhibit similar growth fitness when mutated, and which are also transcriptionally correlated. The x-axis represents the 269 TF knockouts from the expression dataset. The y-axis represents the genes within that cluster, derived from the original fitness sub-module. Red represents increased gene expression, green represents decreased gene expression. Each HOPACH cluster is annotated in respect to the fitness sub-module it was derived from, and the number of HOPACH clusters within that sub-module.

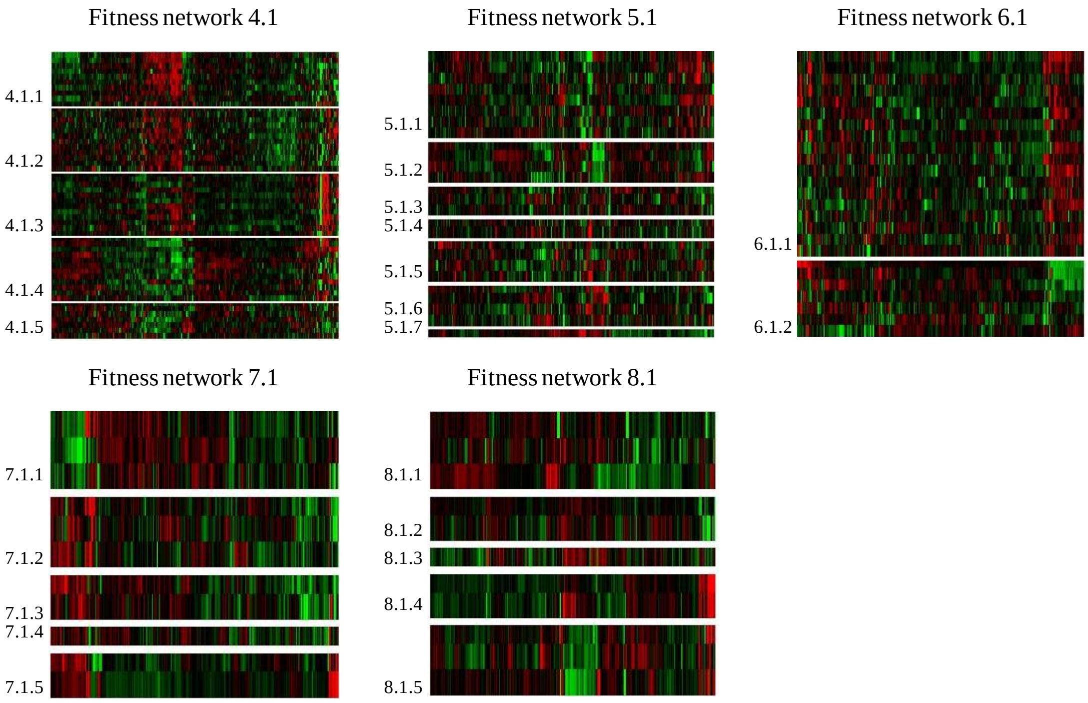


Figure 4. 5 Heatmaps of expression clusters identified by HOPACH from fitness modules 4, 5, 6, 7 and 8 respectively
Each HOPACH cluster is annotated in respect to the fitness sub-module it was derived from, and the number of clusters within that network. The HOPACH clusters identified within each fitness sub-module represents a group of genes which exhibit similar growth fitness when mutated, and which are also transcriptionally correlated. The x-axis represents the 269 TF knockouts from the expression dataset. The y-axis represents the genes within that cluster, derived from the original fitness sub-module. Red represents increased gene expression, green represents decreased gene expression. Each HOPACH cluster is annotated in respect to the fitness sub-module it was derived from, and the number of HOPACH clusters within that sub-module.

4.2.4 Functional analysis of each HOPACH cluster

The web based tool DAVID [21, 22] was used to test whether there is any functional enrichment within each of the 90 HOPACH clusters. Functional annotations were colour coded by their corrected FDR score, as detailed in section 2.3.3. To summarise, functional annotations with a corrected FDR ≤ 0.05 are represented in red text, those with a corrected FDR ≤ 0.1 are represented in green text, and black text represents no significant enrichment (FDR > 0.1).

4.2.5 Network Inference and modularisation procedure

The average profile of each HOPACH cluster was calculated as a means of representing each HOPACH cluster. The average profiles were merged to create a single dataset containing 90 rows (for each HOPACH cluster) and 269 columns (for each TF knockout within the expression data). The integrated dataset was then used in ARACNE [30, 32] to reverse engineer a network based on the structural properties of the fitness network and the gene expression measurements. Statistically significant edges were selected using a highly significant threshold of 10^{-28} corresponding to an MI $\geq$ 0.3 (Table 4.2). No edges were eliminated using the data processing inequality (DPI). The thresholded network contained 57 nodes and 185 edges. The network was visualised in Cytoscape [140], using a force directed layout. Consistent with prior analysis, the network was modularised using GLay [33]. Due to the low number of nodes, only a single level of modularisation was used. For each module, hierarchical clustering (HCL) using Euclidian distance and the average linking method, was performed in MultiExperiment Viewer (MeV) [160], with the aim of identifying if nodes within each module shared similar co-expression profiles.

The final network represents the correlation between expression profiles, within the confines of the modular structure defined by the fitness network. Each node is representative of a group of genes that share similar growth fitness when mutated, which are also co-regulated. Edges between nodes suggest co-fitness and co-expression.

p-value	MI
0.05	0.018508
0.01	0.025849
1.00E-09	0.099365
1.00E-19	0.204387
1.00E-29	0.30941
1.00E-39	0.414433
1.00E-49	0.519456

Table 4. 2 ARACNE P-values and corresponding MIs for expression & fitness integrated network

4.3 Results

4.3.1 Cellular functions exhibit similar behaviour across diverse datasets

Functional analysis of each HOPACH cluster identified cellular functions that are significantly correlated across the fitness and expression datasets. The most enriched functions for each cluster are shown in Table 4.3. Red text indicates an FDR $\leq$ 0.05, green text indicates an FDR $\leq$0.1 and black text represents non-significant enrichment. The full raw DAVID files are available on the supplementary CD, in the ‘Chapter 4’ folder. HOPACH clusters marked with an asterisk in Table 4.3, were eliminated during the network thresholding procedure. I could prove that 20% of clusters exhibited significant enrichment (FDR $\leq$ 0.1). Noteworthy are the statistically significant associations between genes encoding oxidation reduction and membrane (HOPACH cluster 1.1.4); protein transport, membrane and ribosome (HOPACH cluster 1.1.6), cytoplasmic ribosome and ribonucleoprotein (HOPACH cluster 1.2.1), mitochondrial ribosome and aerobic respiration (HOPACH cluster 1.4.3), RNA modification and nucleolus (HOPACH cluster 2.1.6), and transmembrane, cell wall and glycoproteins (HOPACH cluster 2.1.8). Also the separation of small ribosomal proteins and small ribosomal protein biogenesis (HOPACH clusters 3.1.1 and 3.1.2) from large ribosomal protein and large ribosomal protein biogenesis (HOPACH clusters 3.2.1 and 3.2.2). Finally RNAPII core promoter activity and chromatin disassembly (HOPACH cluster 3.3.1), proteasome genes (HOPACH cluster 3.4.3) and chromatin assembly and membrane gene (HOPACH cluster 6.1.1). Interestingly HOPACH cluster 3.3.2 is statistically enriched in chaperone genes, but is excluded when the network is thresholded, suggesting that the six genes within this cluster are share no co-expression and co-fitness with any other cellular processes.

HOPACH Cluster	Gene Count	Functional enrichment
1.1.1	68	intrinsic to membrane (26), cation transport (5), monosaccharide biosynthetic process (3), Zinc finger, C2H2-like (4), endoplasmic reticulum (4)
1.1.2	79	invasive growth in response to glucose limitation (5), cell wall biogenesis (6), manganese binding (4), sporulation, (7) cellular polysaccharide metabolism (4)
1.1.3	84	Zinc ion binding (9), transcription regulator activity (8), anatomical structure homeostasis (3), reproductive developmental process (5), chromosomal part (7)
1.1.4	62	Membrane (28), mitochondrial envelope (10), oxidation reduction (13), response to temperature stimulus (7), vacuole (6), manganese ion binding (3), WD40 (3)
1.1.5	95	glycoprotein biosynthetic process (8), glycoprotein (12), Permease for cytosine/purines (3), regulation of cellular protein metabolic process (9)
1.1.6	62	protein transport (13), intrinsic to organelle membrane (8), structural constituent of ribosome (9), intracellular protein transport (10), vesicle-mediated transport(9)
1.1.7	96	protein transport (21), Chaperone (6), nucleoside-triphosphatase regulator activity (6), ncRNA processing (10), RNA processing (14),
1.1.8	69	cell cycle (17), chromosome (6), centromeric region (6) M phase (9), incipient cellular bud site (3), protein import mitochondrial matrix (9), Spliceosome (3)
1.1.9	63	cellular protein complex assembly (5), Zinc finger (3), cation transport (4), ubl conjugation (3), sexual reproduction (5)
1.2.1	109	protein complex biogenesis (12), iron-sulfur (4), ribonucleoprotein,(14), sulfurtransferase activity (3), cytosolic ribosome (10), ribosomal protein (10)
1.2.2	31	response to temperature stimulus (9), TCA cycle / mitochondrial membrane (5)
1.2.3	108	regulation of translation (10), biopolymer glycosylation (5), biogenic amine biosynthetic process (3), ribosome biogenesis (8), cell wall biogenesis (5)
1.2.4	73	extracellular region, (5) NAD(P)-binding domain (4), exocytosis (3), chromosome organization (8), sporulation (6), chromosome organization (8),
1.2.5	64	cell cycle (16), sporulation (6), nuclear division (5), reproduction of a single-celled organism (5), Meiosis (3), membrane (17), homeostatic process (4)
1.2.6	41	mrna processing (4), nucleoplasm part (5), DNA replication (5), amino-acid biosynthesis (3), mitochondrion (6), metal ion binding (5)
1.2.7	21	organelle membrane (6), regulation of RNA metabolic process (5), zinc (3), membrane (7)
1.2.8	34	nucleotide-binding (12), maintenance of protein location in cell(3), G1 phase of mitosis (3), cell cycle phase (9), GTPase (3), vesicle-mediated transport (7)
1.3.1	45	Transmembrane (11), regulation of cell cycle (4), nuclear chromosome (5), cellular macromolecule catabolic process (6), chromosome segregation (4)
1.3.2	42	asexual reproduction (4), cell wall organization (6), endoplasmic reticulum (8), cell division (6), nuclear envelope (4), golgi membrane (3), cell morphogenesis (4)
1.3.3	31	regulation of phosphorylation (3), energy derivation by oxidation of organic compounds / mitochondrion (7), microtubule organizing (3), membrane (12)
1.3.4	80	Kinase (8), Redox-active center (3), iron-sulfur cluster assembly (3), RNA processing (13), protein ubiquitination (5), mitochondrial matrix (6)
1.4.1	59	Glycosylation (5), endoplasmic reticulum (9), Golgi membrane (7), vacuole (5), glucose metabolic process (3), endomembrane system (8)
1.4.2	71	Helicase, superfamily 1 and 2 ATP-binding (5), intrinsic to membrane (22), cellular response to stress (10), ATPase activity (7), cell division (7)
1.4.3	61	organellar ribosome / mitochondrial ribosome (6), endoplasmic reticulum (11), aerobic respiration (4), generation of precursor metabolites and energy (6)
1.5.1*	4	HMX1, SLM3, CRR1, YDL027C, (transmembrane protein / sporulation) (4)
1.5.2	12	phosphate metabolic process (3), membrane (4)
1.6.1*	2	TAH18, ADE17
1.6.2*	1	OCA4
1.6.3*	1	YML090W
1.6.4*	3	SMK1, YGL217C, YPR077C
2.1.1	42	stress response (4), cofactor binding (5), glycoprotein (6), response to abiotic stimulus (6), sporulation (4), vacuole (4), integral to membrane (13)

2.1.2	28	coiled coil / cytoskeleton (4), transcription from RNA polymerase II promoter (3), cellular protein catabolic process (3), protein transport, membrane (9)
2.1.3	54	chromosome segregation (8), intracellular non-membrane organelle (19), cyto-skeleton (5), mitosis (7), nuclear lumen (10), transcription, DNA-dependent (5)
2.1.4	25	Golgi vesicle transport (5), plasma membrane (6), membrane-enclosed lumen (6), metal ion binding (4), mitochondrion (4), cell cycle (7)
2.1.5	53	M phase of meiotic cell cycle (6), chromosome (8), regulation of transcription (11), chromosome (8), conjugation with cellular fusion (3)
2.1.6	21	nuclear lumen (8), RNA modification (6), ribosome biogenesis (6), nucleolus (6), cell cycle (7), intracellular transport (6), cellular macromolecular complex (3)
2.1.7	26	nucleoside metabolic process (3), protein transport (3), membrane-enclosed lumen (4), organelle membrane (4), cytoplasm (6)
2.1.8	49	Transmembrane (17), cell wall biogenesis/degradation (5), glycoprotein (10), endoplasmic reticulum membrane (6), phospholipid metabolic process (7), vacuole (6), Golgi apparatus part (4), regulation of cellular protein metabolic process (4)
2.2.1	77	nuclear lumen (16), macromolecular complex subunit organization (12), rRNA 5'-end processing (6), respiratory chain (3), ribonucleoprotein complex (14)
2.2.2	68	regulation of protein metabolism (9), wd repeat (5), nuclear export / import (8), tRNA aminoacylation for translation (4), one-carbon metabolic process (4)
2.2.3	95	protein catabolic process(13), response to temperature stimulus (8)
3.1.1	51	cytosolic small ribosomal subunit (26), ribosome biogenesis (30), cleavages during rRNA processing, preribosome (9), rRNA binding (5), regulation of translation (10), rRNA transport (7), ncRNA 3'-end processing (6), ribosomal small subunit biogenesis (10), isopeptide bond (4), zinc (6), transmembrane (4)
3.1.2	18	small-subunit processome (11), maturation of SSU-rRNA (11), t-UTP complex / rDNA heterochromatin (5),
3.1.3*	3	RPS23B, RPC19, YOR146W
3.1.4*	5	NDC1, YOL019W, DEF1, YDL034W, SPT6, (compositionally biased region:Gln-rich)
3.2.1	19	Preribosome large subunit precursor (7), ribosome export from nucleus (5), maturation of LSU-rRNA (3), Initiation factor (4), gtp-binding (3)
3.2.2	48	cytosolic large ribosomal subunit (25), regulation of translation (10), ribosomal large subunit biogenesis (8), rRNA binding (5), DNA-directed RNA polymerase III complex (6), Ribosomal protein 60S, ribosomal large subunit biogenesis (3), zinc finger (3),
3.3.1	18	transcription from RNA polymerase II promoter / chromatin disassembly (8), DNA-directed RNA polymerase II, core complex (3), intracellular transport (4)
3.3.2*	12	Chaperonin TCP-1, conserved site (6),
3.3.3*	5	RPB7, RAD51, RPO26, YNL179C, YPL251W, (DNA-directed RNA polymerase II, core complex)
3.3.4*	6	ALG12, RSC9, COQ6, IDH1, FIT3, FUN12
3.3.5*	8	regulation of transcription (4)
3.4.1	4	NDC1, YOL019W, DEF1, YDL034W, SPT6
3.4.2*	3	PUP1, RPT5, IML2 (proteasome)
3.4.3	18	Proteasome (17), cytosolic proteasome complex (13), endopeptidase activity (14), threonine protease (8), proteasome regulatory particle , lid subcomplex, 26S proteasome subunit P45 (9), sporulation (6)
3.4.4*	2	PRE4, IRC7
3.4.5*	1	PRE2
3.5.1*	1	MAP2
3.5.2*	4	NUP60, SAM4, COG3, KIN3
3.5.3*	1	SSA1
3.5.4*	2	PSD1, MCM2
3.5.5*	1	AAD14
3.5.6	2	AFT2, HRB1
3.5.7*	1	YIL014C-A
3.5.8*	1	YKR078W
3.5.9	2	OPY1, QCR8

4.1.1	10	sporulation resulting in formation of a cellular spore (3), vacuolar protein catabolic process (3), transmembrane protein (5)
4.1.2	12	M phase (3)
4.1.3	12	Glycoprotein (5), chaperone binding, signal (5)
4.1.4	12	organelle lumen (4), mitochondrion organization (3)
4.1.5*	7	cytoskeleton organization (3)
5.1.1*	8	transcription from RNA polymerase II promoter (3), transcription (4)
5.1.2	4	SDO1, TVP18, TAF14, PDR16
5.1.3*	3	YPR089W, PDR5, YPR089W
5.1.4*	2	NAM8, SET6
5.1.5*	4	YBL065W, CDC20 , TAF5, YDR467C (wd repeat)
5.1.6*	4	MAL11, HXT11, YBL044W, MED6
5.1.7*	1	SUR1
6.1.1	18	chromatin assembly (3), integral to membrane (4)
6.1.2	7	transmembrane protein (4)
7.1.1	3	mutagenesis site (3)
7.1.2*	3	RTF1, MTO1, NIF3
7.1.3	2	SIP2, KIP3
7.1.4*	1	MRS1
7.1.5	2	AVT7, TIF1
8.1.1*	3	GLO4, SHU1, YOR052C (cellular response to stress)
8.1.2*	2	PAU13, YOR050C
8.1.3*	1	SPT10
8.1.4	2	HOM3, YHR020W (atp-binding)
8.1.5	3	RNA1, YMR290W-A, BUD16

Table 4. 3 The most enriched functions within each HOPACH cluster.
Red text represents an FDR $\leq$ 0.05, green text FDR $\leq$ 0.1and black text represents non-significant enrichment HOPACH clusters marked with an asterisk (*) were eliminated during network thresholding. The functional annotations within each HOPACH cluster represents a group of genes which share the same function, exhibit similar growth fitness when mutated and which are also transcriptionally correlated.

4.3.2 Community analysis of the integrated network identifies highly interconnected modules

Functional analysis on all 90 HOPACH clusters identified that 20% could be significantly associated (FDR $\leq$ 0.1) to biological functions. The thresholded network contained 57 nodes (HOPACH clusters) and 185 edges, of which 30% (17 / 57) showed significant functional organisation. A single level of modularisation was applied to the network. Six modules were identified; details are shown in Table 4.4. Module size ranged from three nodes to 18 nodes. These modules localised to distinct areas of the force driven layout parent network (Figure 4.6). HCL was performed on each module to determine the transcriptional correlation of nodes within the module.

Module	**Nodes**	**Edges**	**HOPACH clusters within each module**
1	18	95	1.1.7; 1.1.8; 1.2.1; 1.3.4; 1.4.3; 1.5.2; 2.1.3; 2.2.1; 3.1.1; 3.2.2; 3.3.1; 3.4.3; 3.5.9; 4.1.4; 5.1.2; 6.1.2; 7.1.1
2	12	24	1.1.1; 1.1.2; 1.1.3; 1.2.4; 1.2.5; 1.3.1; 1.4.2; 2.1.5; 2.2.3; 4.1.1; 4.1.2; 7.1.3
3	10	30	1.1.5; 1.2.3; 1.3.2; 1.4.1; 2.1.8; 2.2.2; 4.1.3; 6.1.1; 7.1.5; 8.1.4
4	7	10	1.1.4; 1.2.2; 2.1.1; 2.1.6; 3.1.2; 3.2.1; 8.1.5
5	7	6	1.1.6; 1.1.9; 1.2.7; 1.2.8; 2.1.4; 2.1.7; 3.4.1
6	3	3	1.2.6; 1.3.3; 2.1.2

Table 4. 4 Nodes contained within each module of the 0.3 MI integrated network
The number of nodes, edges and HOPACH clusters contained within each module are shown in columns 2, 3 and 4 respectively. The functional annotations for each HOPACH cluster shown in column 4 can be viewed in Table 4.3.

4.3.3 The modular structure of the integrated network reflects functional compartmentalisation

Having shown that genes could be clustered by their behaviour across datasets, I next set out to analyse the interactions between the nodes within each module and whether they were representative of the organisational behaviour of the yeast system. Below is a systematic analysis of the six identified modules.

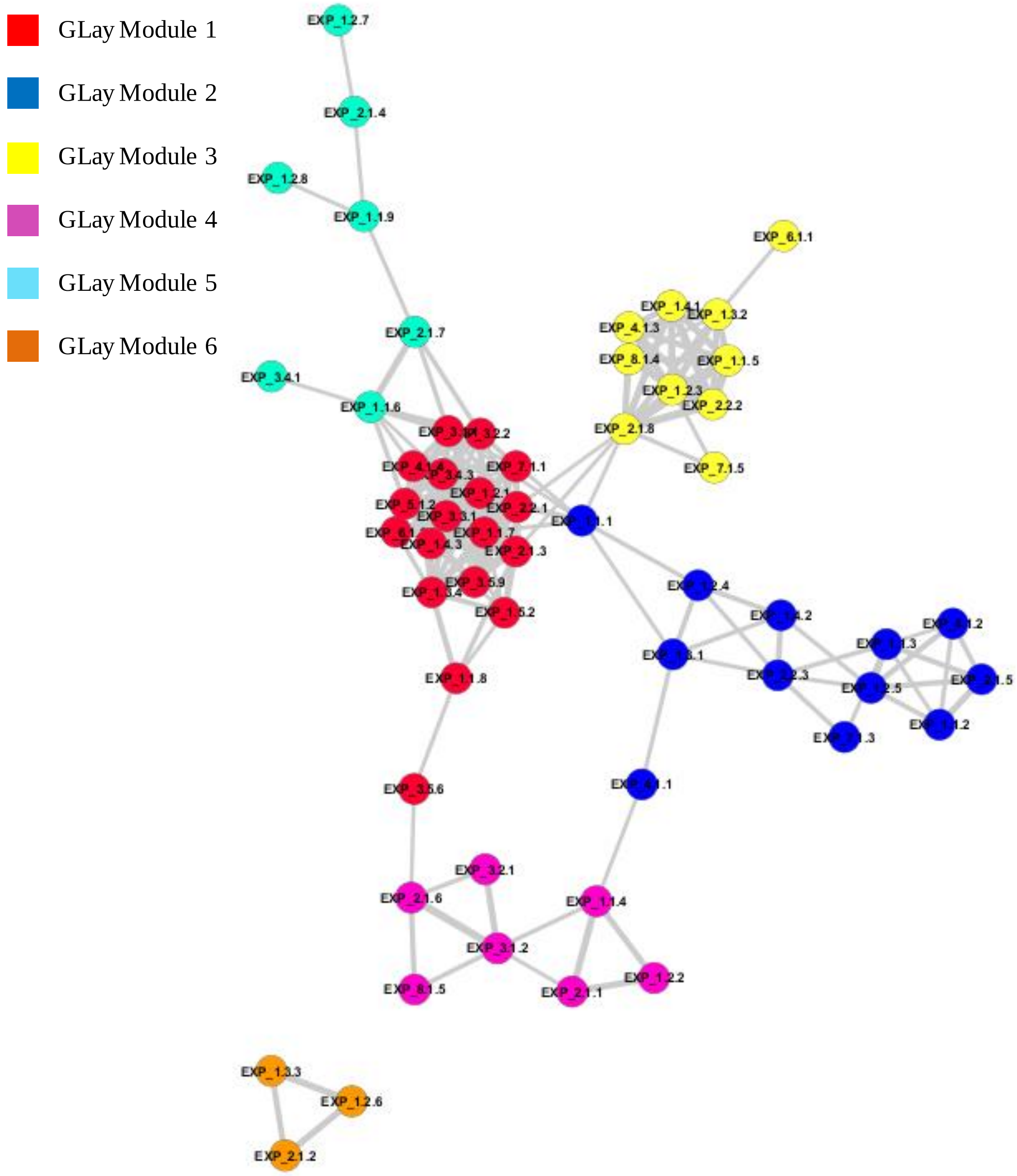


Figure 4. 6 Modules localise within distinct areas of the force directed integrated parent network.

An undirected network showing the interactions between HOPACH clusters. Nodes are labelled to represent their Hillenmeyer fitness module and expression HOPACH cluster. For example, HOPACH cluster 1.1.4, represents fitness sub-module 1.1, HOPACH cluster 4. Each node represents a HOPACH cluster representative of functions which exhibit co-expression and co-fitness (Table 4.3). Table 4.4 shows the HOPACH clusters located within each module.

4.3.3.1 Module 1: Genes encoding ribosomal proteins and post translational protein modification are correlated across datasets

Module 1 contains 18 nodes; suggesting that the functions represented by each node are correlated to each other, both phenotypically and transcriptionally (Figure 4.7). Investigation into the functions represented by these nodes (Table 4.3) identified that nodes significantly enriched in protein transport (node 1.1.7), small cytoplasmic RPs and small RP biogenesis (node 3.1.1), large cytoplasmic RPs and large cytoplasmic RP biogenesis (node 3.2.2), RNAPII transcription (node 3.3.1) and proteasome genes (node 3.4.3). These results are consistent with reports that the coordinate regulation of 150 rRNA genes, 137 RPs, together with RNAPII transcription is required to form a functional ribosome [97]. The significant enrichment of proteasome genes (node 3.4.3) within this module highlights the defence mechanisms against truncated proteins, which aid in the folding of newly synthesised polypeptides and malformed proteins during stress response. Node 1.4.3 is significantly enriched in genes encoding mitochondrial ribosome. There are also multiple nodes representative of cell cycle (1.1.8, 1.3.4, 2.1.3). These results suggest a strong correlation between cell cycle progression / chromosome segregation and ribosome biogenesis. In Chapter 2, I reported that yeast strains containing a mutation in cell cycle checkpoint gene *BUB1* were phenotypically correlated to yeast strains containing mutated RP genes, especially stains containing small 40S RPs mutants. This is reflected in the integrated network, node 2.1.3 is enriched in chromosome segregation, and is directly connected to node 3.1.1, enriched in small RPs.

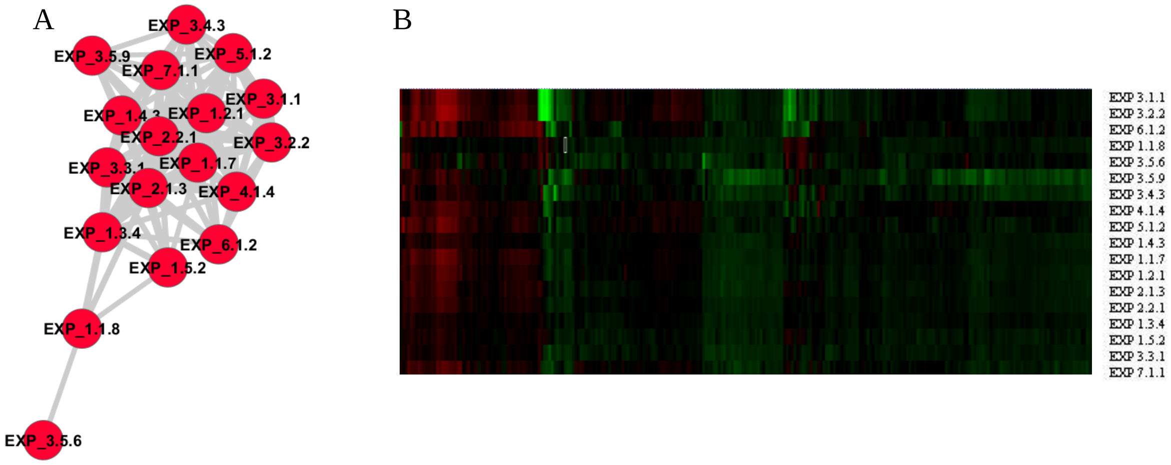


Figure 4. 7 Module 1 visualisation and HCL analysis on the nodes located in module 1. Panel A The structure of module 1, edge length and width represents MI score. Nodes are labelled by HOPACH cluster. Panel B The result of HCL shows that all nodes are transcriptionally correlated. Red indicates an increase in expression, green indicates a decrease in expression.

4.3.3.2 Module 2: Cell cycle, stress response and energy production.

All 12 nodes within this module have a similar expression profile (Figure 4.8), interestingly of the 12 nodes within this module; seven are derived from fitness module 1, the largest module containing a broad scope of cellular processes (Chapter 2). Integrating fitness and transcription data reveals that genes within fitness module 1 are involved in stress response (nodes 1.1.2, 1.4.2, 2.2.3 and 7.1.3) and growth regulation / cell cycle (nodes 1.1.2, 1.1.3, 1.2.4, 1.2.5, 1.3.1, 2.1.5, 4.1.1 and 4.1.2) (Table 4.3) are all co-expressed. This suggests that under stress conditions (particularly glucose starvation), the rate of cell growth is controlled. Consistent with observations that yeast cells starved of nutrients do not continue proliferating, but enter a regulated growth state until nutrients become available [161] [162].

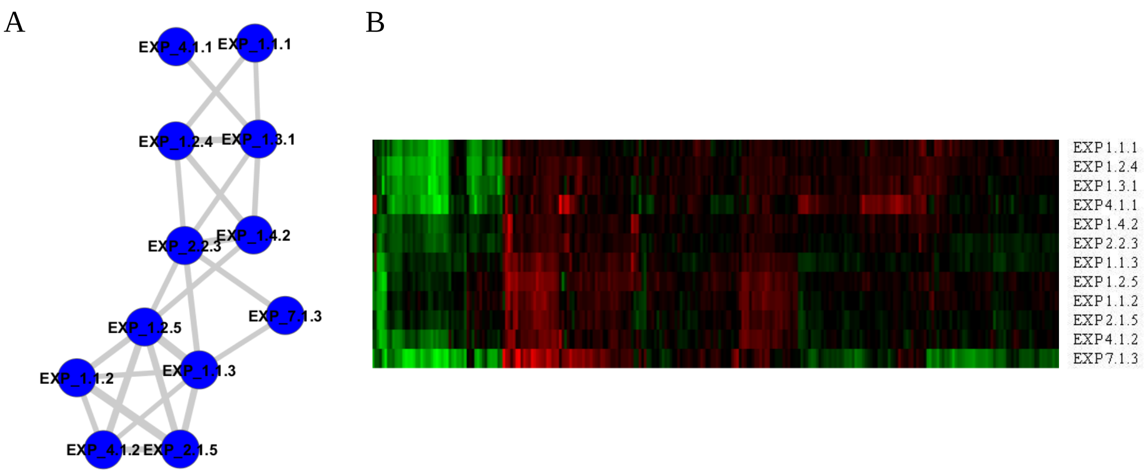


Figure 4. 8 Module 2 visualisation and HCL analysis on the nodes located in module 2
Panel A The structure of module 2, edge length and width represents MI score. Nodes are labelled by HOPACH cluster. Panel B The result of HCL shows that all nodes are transcriptionally correlated. Red indicates an increase in expression, green indicates a decrease in expression.

4.3.3.3 Module 3: Glycosylation behaviour is conserved across transcription and fitness data

Module 3 contains ten nodes (Figure 4.9), six of which are enriched in glycosylation / glycoprotein genes (1.1.5, 1.2.3 1.4.1, 2.1.8, 2.2.2, 4.1.3) (Table 4.3 & raw DAVID tables on supplementary CD). Glycosylation is a co- and post- translational form of modification; and serves in numerous cellular functions including stabilising proteins and roles in cell – cell adhesion [163]. Cell adhesion is essential in regulating cell growth and cell cycle [164] [165] which is consistent with the enrichment of cell cycle processes such as cell wall biogenesis (nodes 1.2.3 and 2.1.8), cell division (1.3.2) and chromatin assembly (6.1.1). Glycosylation can be N-linked and O-linked [166] and is dependent on the relatively small hexosamine biosynthetic pathway [166] [167]. The hexosamine pathway diverges from glycolysis at fructose-6-phosphate, consistent with glycosylation functions grouped with glycolysis and glucose metabolism genes (nodes 1.2.3, 1.4.1). N-linked glycosylation of proteins occurs co-translationally in the ER, and further remodelling of N-glycans takes place in the golgi apparatus [166], represented by nodes 2.1.8, 1.3.2, 1.4.1. The observed linkage between glycosylation and energy metabolism pathways is consistent with reports that N and O-linked glycosylation may have roles in responding to cellular nutrient availability and in metabolic diseases [166].

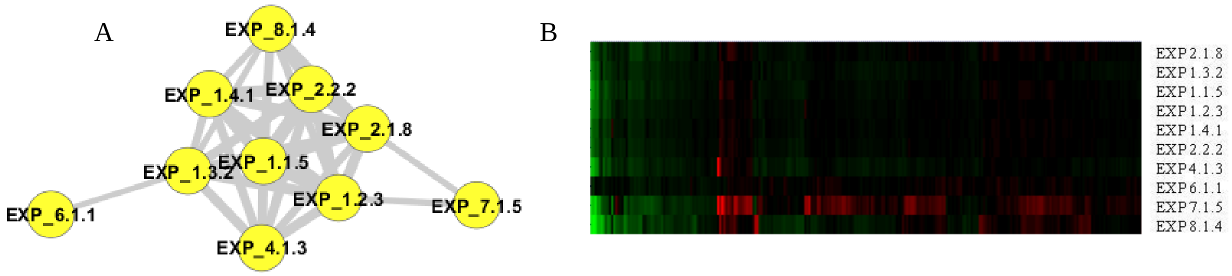


Figure 4. 9 Module 3 visualisation and HCL analysis on the nodes located in module 3
Panel A The structure of module 3, edge length and width represents MI score. Nodes are labelled by HOPACH cluster. Panel B The result of HCL shows that all nodes are transcriptionally correlated. Red indicates an increase in expression, green indicates a decrease in expression.

4.3.3.4 Module 4: The anti-correlation between stress response genes and ribosome biogenesis.

Module 4 contains seven nodes (Figure 4.10A). Figure 4.10B shows a clear transcriptional anti-correlation between stress response genes (nodes 1.1.4, 1.2.2, 2.1.1) and RNA processing / ribosome biogenesis genes (nodes 2.1.6, 3.1.2, 3.2.1). These results are consistent with reports that repression of ribosomal biogenesis and RNA metabolism genes are a feature of stress response [161]. Gasch *et al* reported that a number of diverse cellular processes are induced in response to environmental stress, including carbohydrate metabolism (nodes 1.2.2), cell wall modification, vacuolar and mitochondrial functions (nodes 1.1.4, 2.1.1), cellular redox (nodes 1.1.4, 2.1.1), protein folding, and cell wall reactions [161]. Many of the genes induced during a stress response have been shown to protect the cell [162]. The integration of fitness and transcriptional data demonstrates and further substantiates the anti-correlated expression profile between genes encoding ribosome related functions and stress responses.

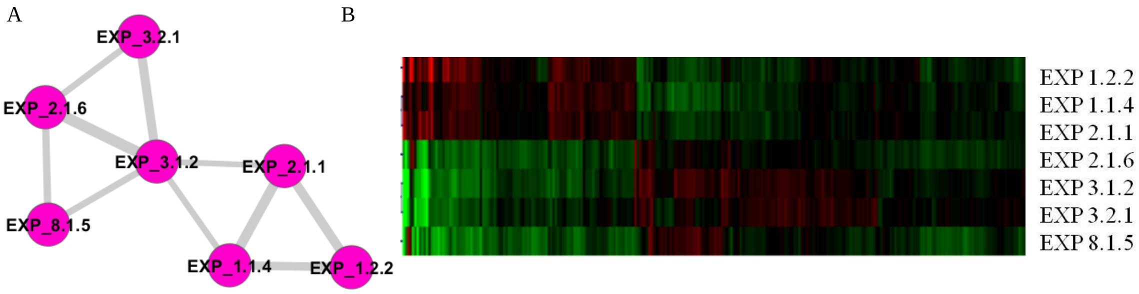


Figure 4. 10 Module 4 visualisation and HCL analysis on the nodes located in module 4. Panel A The structure of module 4, edge length and width represents MI score. Nodes are labelled by HOPACH cluster. Panel B The result of HCL shows that all nodes are anti-correlated transcriptionally. Red indicates an increase in expression, green indicates a decrease in expression.

4.3.3.5 Module 5: The inverse transcriptional relationship between mitochondrial precursor protein transport and zinc ion binding

Module 5 contains seven nodes (Figure 4.11). There is an anti-correlated transcriptional relationship between two groups of nodes (Figure 4.11B). The first group are enriched in protein transport, ribosomal constituents and protein targeting to mitochondria (nodes1.1.6, 2.1.7) (Table 4.3, raw DAVID files on supplementary CD) suggesting these nodes encode genes required for import of precursor proteins into mitochondria and mitochondrial ribosome assembly [115]. The second group consists of genes encoding protein complex assembly, ubiquitin conjugation and sexual reproduction (node 1.1.9), zinc and RNA metabolism (node 1.2.7), cell cycle and nucleotide binding (node 1.2.8), vesicle transport and mitochondrion (node 2.1.4) and chromosome segregation (node 3.4.1). The results suggest that genes in the second group are expressed when genes in the first group are repressed and vice versa. In support of this, ubiquitin proteins have been reported to negatively regulate the synthesis of proteins that localise to the mitochondrial inner membrane [117] and mitochondrial import is known to be cell cycle dependent and is restricted to the reductive-building phase of the cell cycle [94].

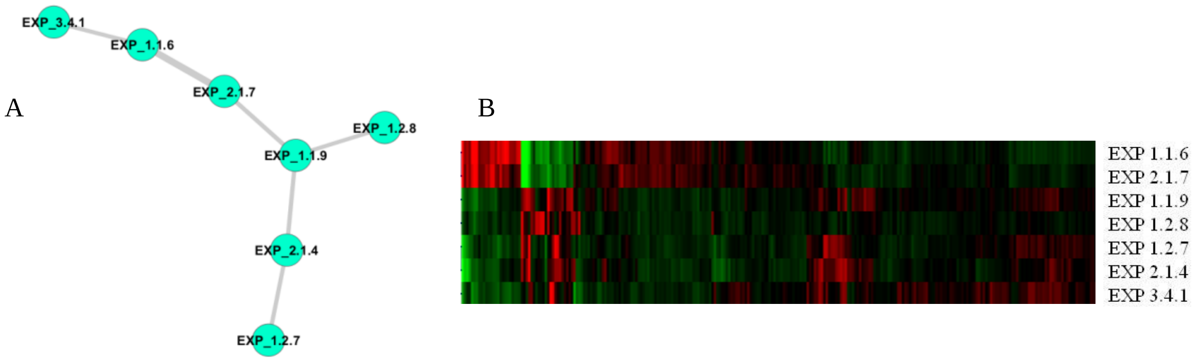


Figure 4. 11 Module 5 visualisation and HCL analysis on the nodes located in module 5. Panel A The structure of module 5, edge length and width represents MI score. Nodes are labelled by HOPACH cluster. Panel B The result of HCL shows that two groups of nodes are anti-correlated. Red indicates an increase in expression, green indicates a decrease in expression.

4.3.3.6 Module 6: The co-expression and co-fitness of cytoskeleton, mitochondria and DNA replication genes.

Module 6 contains three nodes, 1.2.6, 1.3.3 and 2.1.2 (Figure 4.12). Nodes 1.3.3 and 2.1.2 are enriched in microtubule and cytoskeleton function, node 1.2.6 is enriched in mRNA processing, DNA replication and mitochondrion genes. The cytoskeleton plays an important role in the transport of vesicles and organelles, as well as cell division. The co-expression between cytoskeleton components and mitochondria energy generation (node 1.3.3), is likely because cytoskeleton motor proteins require the hydrolysis of ATP to convert chemical energy into mechanical movement, in order to move cell organelles [168] and proteins [169]. Alternatively the correlation between DNA replication, mitochondria, microtubules and cytoskeleton may be explained by cell cycle linked motility in which during the S-phase, mitochondria undergo linear movement from mother cells to daughter cells via actin cables [170] [171] .

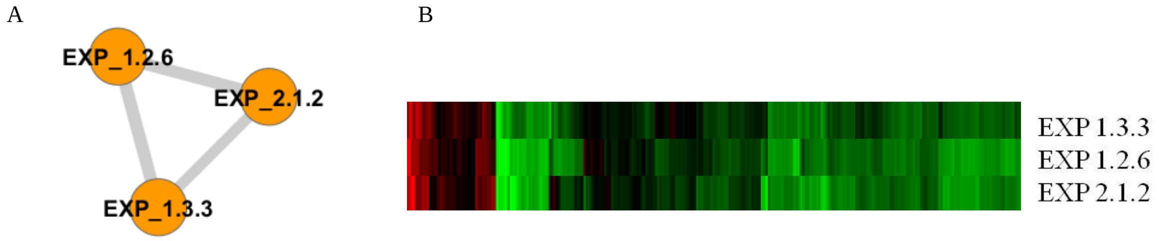


Figure 4. 12 Module 6 visualisation and HCL analysis on the nodes located in module 6. Panel A The structure of module 6, edge length and width represents MI score. Nodes are labelled by HOPACH cluster. Panel B The result of HCL shows that all nodes are transcriptionally correlated. Red indicates an increase in expression, green indicates a decrease in expression.

4.3.4 The strongest edges between nodes are representative of the metacycle

To study what the strongest interactions between cellular processes were, I thresholded the network using an extremely high p-value (10^{-57}), this retained edges with an MI $\geq$ 0.6 (Table 4.2). The aim was to identify network modules that represented the most significant functions correlated across both expression and fitness data. The resulting network contained 15 nodes and 15 edges. Modularisation was not required, as the network formed five unconnected modules (Figure 4.13A). Each module contained between two and five nodes. By using the functional analysis data obtained previously (Table 4.3) I was able to functionally classify each module (Table 4.5). Interestingly, modules showed remarkable similarity to phases within the metacycle. The metacycle is an example of how the regulation of groups of genes linked in space and time leads to physiological changes in yeast. The oxidative phase is characterised by the co-ordinate up-regulation of rRNA processing, translation machinery, RPs, ribosome biogenesis, and sulphur uptake (modules

3 and 5) [106] [94]. The reductive-building phase is characterised by the up-regulation of mitochondrial proteins, DNA replication, spindle pole components and histones (module 2) [106] [94], finally the reductive-charging phase is characterised by the co-ordinate up-regulation of glycolysis, ubiquitination machinery vacuolar and proteasomal transcripts, carbohydrate breakdown and cell division (modules 1 and 4) [106] [94]. HCL using the Euclidian distance metric and average linkage identified that each module exhibited the same expression profile (Figure 4.13B), further substantiating that genes involved in each metacycle phase are co-ordinately regulated together. Modules 1 and 3, representing the reductive-charging phase and the onset of the oxidative phase respectively exhibited similar expression profiles, suggesting that genes involved in the early stages of ribosome biogenesis such as rRNA maturation and processome assembly are closely co-regulated with genes involved in glycolysis and NADP production. Module 5, which represents genes involved in translation initiation and ribosome biogenesis, represents the latter part of the oxidative phase, and therefore not correlated to Module 1. This is consistent with reports that ribose-5-phosphate production via riboneogenesis is dependent on glycolytic intermediates [92], the production of ribose-5-phosphate is essential in synthesising rRNA required for ribosome assembly. These features were also identified in my analysis on the fitness and expression networks (Chapters 2 and 3).

Module	**No. of Nodes**	**Most common functions**	**Phase of metacycle**
1	5	glucose metabolic processes (glycolysis / gluconeogenesis), cell wall biogenesis, NADP production, cell cycle phase, alcohol catabolic process and cytokinesis	Reductive - charging
2	4	organellar ribosome / mitochondrial ribosome, respiration, protein transport, chromatin remodelling	Reductive building
3	2	RNA modification, ribosome biogenesis, nucleolus, small-subunit processome, maturation of SSU-rRNA, t-UTP complex / rDNA heterochromatin,	Onset of oxidative phase
4	2	sister chromatid segregation, DNA metabolism, ubquitination, ATP, Redox-active center, cytokinesis	Reductive – building / charging
5	2	ribosome biogenesis, regulation of translation, RNA transport, rRNA processing, preribosome,	Oxidative

Table 4. 5 The modules with the strongest interactions correspond to phases of the metacycle

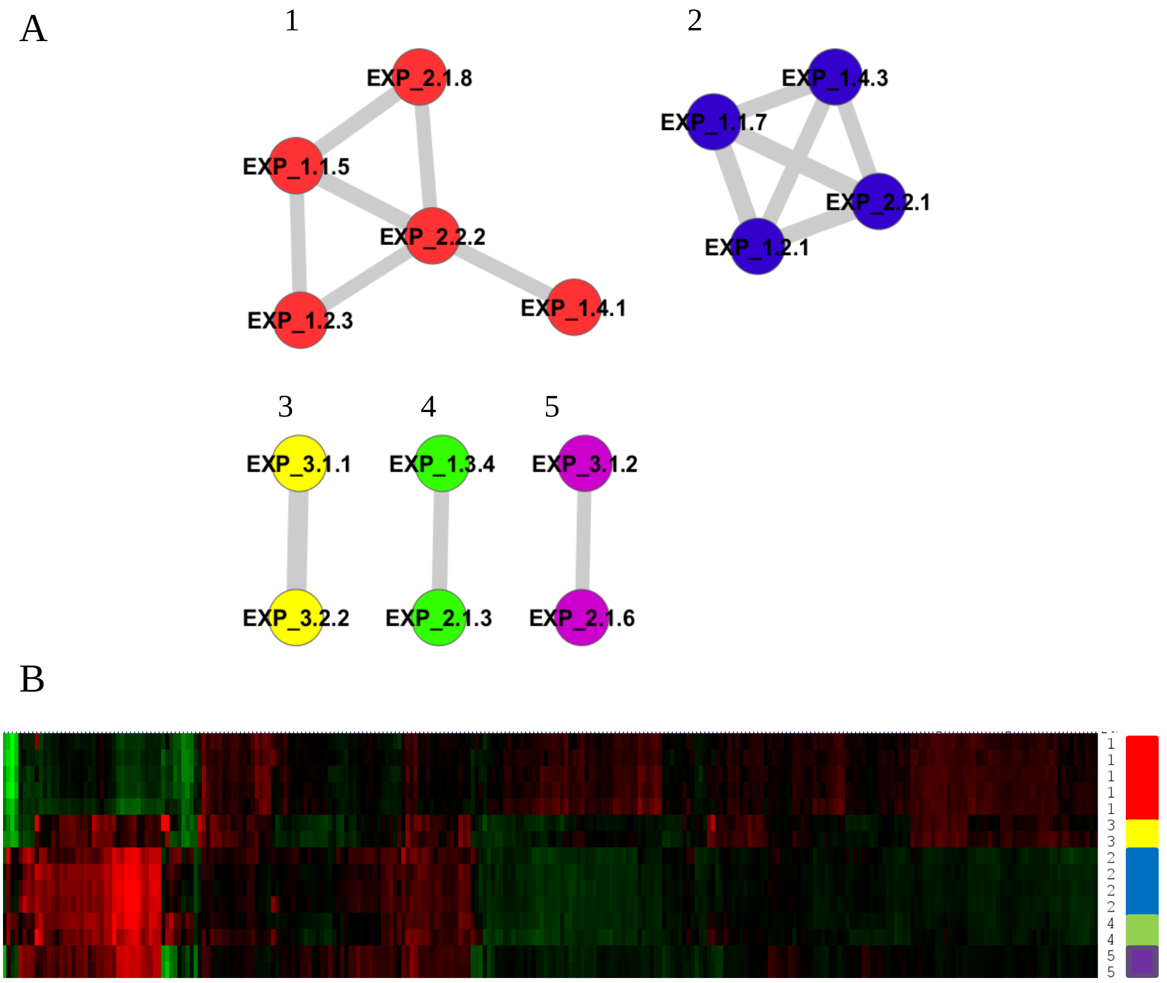


Figure 4. 13 Visualisation of the most significant integrated network edges (MI > 0.6)
Panel A Five modules were identified using a threshold of an MI threshold of 0.6, each characteristic of a stage within the metacycle. Panel B. A heatmap representing the clustering (using Euclidean distance metric and average linkage clustering), of the expression measurements within each module. Red indicates an increase in expression, green indicates a decrease in expression. The legend is shown on the right; each colour represents module number as shown in Figure 4.13A. Nodes within each module all exhibit the same co-ordinated expression in response to TF KO knockouts, and are clustered separately from each other.

4.4 Discussion

The most important aspect of research presented in this chapter is the integration of a fitness compendium and a TF knockout expression dataset. By integrating these two types of data, the aim was to overcome the shortcomings presented by using each dataset individually and create a comprehensive network representing the global and local organisation of the yeast system. As proven by Giaever *et al*, expression data alone is unable to take post-transcriptional modification and translational regulation into consideration which means genes that show little to no change in expression may in fact be essential for cell viability [37]. Conversely, genes that show a statistically significant up-regulation in gene expression aren't necessarily essential for cell viability [37]. Fitness data on the other hand is limited by the possibility of inferring false positive edges between genes that have similar fitness profiles, but have no underlying biological link (as discussed in section 2.4.3). Therefore, by integrating a comprehensive fitness compendium with a gene expression dataset, it was possible to minimise the weaknesses posed by using each dataset individually.

4.4.1 The power of applying reverse engineering methods to integrated datasets

The aim of this study was to identify and functionally annotate representative clusters of genes that exhibit correlated behaviour across fitness and expression datasets. HOPACH clustering and subsequent functional analysis demonstrated that merged fitness and expression data could be functionally categorised. Comparing the functional modules identified in this analysis with those identified in existing integrative network studies such as Tanay et *al* [91] and Ideker *et al* [88], shows an overlap of functional modules between the studies, including protein transport (node 1.1.6), stress response (node 1.2.2), mitochondrial ribosome (node 1.4.3), ribosome biogenesis (node 3.1.1), ribosomal proteins (node 3.2.2) and proteasome (node 3.4.3). Though the datasets and methodologies vary between studies, it does however demonstrate the power of using integrated datasets for network inference. Single level datasets, such as gene expression, protein binding or fitness data, when used in combination, provide complementary information, which increases the broadness and comprehensiveness of the network.

The integrated network identified nodes (HOPACH clusters representing groups of genes) that exhibited similar behaviour across the fitness and expression datasets. Modularisation and visualisation of the network identified six modules, which when investigated, suggested linkage between genes involved in several biological processes. For example, in the *S. cerevisiae* expression network (Chapter 3), I identified a module that was dedicated to stress response. The results from that analysis suggested that in response to heat shock, there is a co-regulation of ubiquitination machinery and vacuolar catabolism genes (Chapter 3, section 3.3.1.3). In the *S. cerevisiae* fitness analysis I identified that proteasome genes modularised together with ribosomal proteins (Chapter 2, section 2.3.3.3). This integrated analysis revealed that stress response genes are inversely transcriptionally correlated to ribosome biogenesis genes (module 4, Figure 4.10) which suggest potential a cellular mechanism involved in adapting to stress response. The down regulation of genes encoding ribosome biogenesis proteins in response to the up regulation of genes involved in stress response has already been reported [161], however this serves as an example as to how combining datasets can increase the scope and sensitivity of biological networks. In Chapter 2, I discussed that yeast strains mutated in the cell cycle checkpoint gene *BUB1* showed similar fitness to strains mutated in RPs, the correlation in fitness was demonstrated to be higher with small 40S RPs especially. The analysis of this integrated network provided additional evidence between the linkage of chromosome segregation and RPs, as the node representing chromatid segregation was directly connected to the node representing small RPs, however there was no edge identified connecting chromatid segregation to large RPs. Integrative approaches using diverse datasets not only have the potential to determine groups of genes that exhibit significantly correlated behaviour across data sources, but also to characterise genes of unknown function [91].

4.4.2 Using integrated networks to characterise unknown genes

As of July 2013, the *S. cerevisiae* genome contains 761 uncharacterised ORFs, defined as ORFs that are likely but not confirmed to encode a protein [172]. Therefore integrated networks hold great significance as they have the ability to predict the function of uncharacterised ORFs and suggest potentially new directions for experimental research [91]. One way of inferring functions to uncharacterised genes is by analysing the topology of an integrated network. Specifically, analysing nodes that connect larger modules together may infer possible functions. In a biological context, it would highlight potential genes that link cellular processes to each other [91]. This concept is demonstrated with module 1 (Figure 4.14). Nodes that bridge module 1 (represented in red) to other modules include nodes 1.1.6, 2.1.7, 2.1.8, 1.1.1 and 2.1.6, and are highlighted by green dashed circles. Functional analysis of these nodes (Table 4.3) reveals that genes encoding protein transport proteins, or organelles involved in protein transport (endoplasmic reticulum and golgi body) link larger modules together. This is consistent with existing integrative studies, which report signalling and transport genes form bridges between modules [91]. Therefore, if uncharacterised genes are found to localise between larger modules, it may suggest it is involved in protein transport or signalling.

The main take home message is that merging diverse datasets into a single dataset can provide complementary information, resulting in a more detailed and comprehensive high level network, which can provide a great deal of insight into complex biological processes and the nature in which they are controlled.

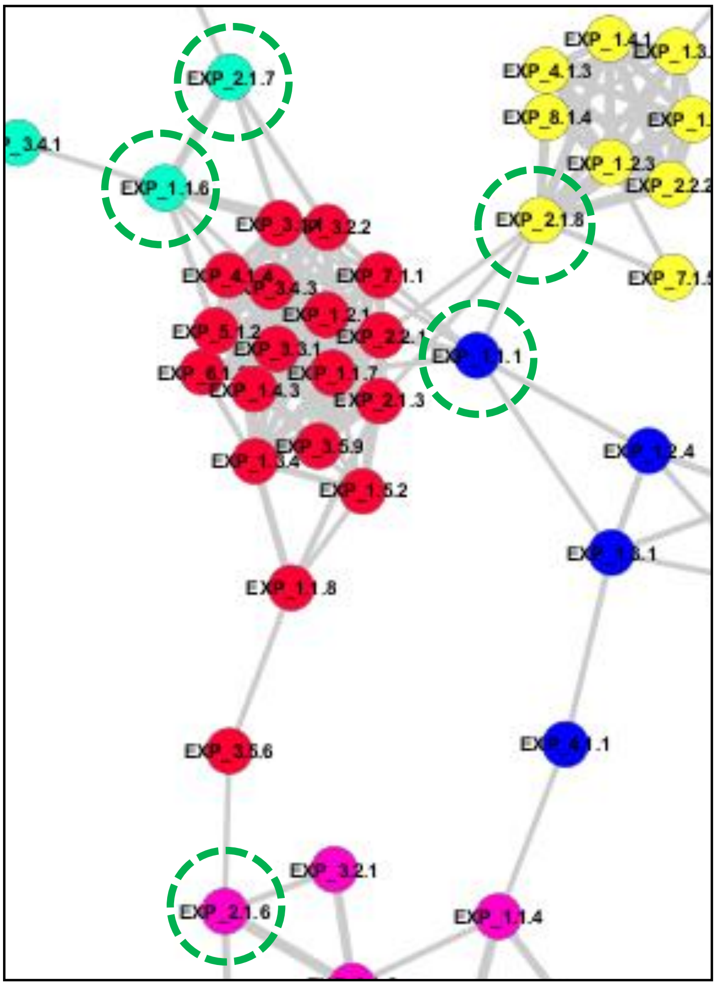


Figure 4. 14 A subset of the integrated network, showing nodes that bridge act as a bridge to module 1.
Module1 is represented in red, the nodes that act as a bridge as highlighted by green circles. Identifying nodes that form bridges between sub-modules can be used to infer genes which link biological processes together, functional analysis of these nodes show a common enrichment in protein transport related functions.

4.5 Concluding remarks

The integration of different types of experimental data has the potential to detect modules and relationships that would not otherwise have been possible using a single level data. Integrating fitness and expression data together is advantageous as they complement each other and overcome the shortcomings of using them individually. The construction of an MI based network using Hillenmeyer's fitness compendium and Reimand's gene expression data in an integrated manner had not previously been attempted, nor had any integrative study using the entirety of Hillenmeyer's fitness compendium. The modules identified in this analysis represent essential cellular processes including cell cycle, stress response, ribosome biogenesis, energy production and post translational modification. These processes cover the biological changes that would be required to adapt to cellular stress, as expected given Hillenmeyer's fitness data is based on cell exposure to 300+ stresses. Furthermore, the integration of TF KO data establishes which TFs are required for controlling the regulation of the genes present within each module, which has the potential to be developed further in future studies. This study also demonstrated that the integration of multiple datasets increases the statistical sensitivity of the network and allows for module detection at finer detail. It is this increase in sensitivity that makes it possible to infer functionality to uncharacterised genes within *S. cerevisiae* genome.

CHAPTER 5: INFERENCE AND ANALYSIS OF A *Schizosaccharomyces pombe* GENE EXPRESSION NETWORK

5.1 Introduction

The analysis so far has focused on *S. cerevisiae* fitness and expression data. Results have shown that networks are representative of functional compartmentalisation and that links between genes encoding energy metabolism and RPs are conserved across fitness and expression data. Studies up to this point had been restricted to *S. cerevisiae,* therefore I decided to expand the analysis to another yeast species, *Schizosaccharomyces pombe. S. cerevisiae* and *S. pombe* are believed to have diverged over 300 million years ago and exhibit numerous differences at the molecular level [57] and especially in regards to energy metabolism (as discussed in section 1.3.2). Like *S. cerevisiae, S. pombe* is a model species and has been well annotated since being fully sequenced in 2002 [57], making it an ideal species for comparison studies against *S. cerevisiae.*

In this chapter, I report the network inference of an *S. pombe* expression network using the same approach utilised in Chapters 2 and 3. The compendium used in this analysis was constructed by Bähler Lab and is publically available on their website. The compendium contains expression data for hundreds of environmental and genetic perturbations [72]. The first aim of this chapter is to infer the expression network for *S. pombe* and identify functional modules. The second aim is to determine if functional modules are shared between *S. pombe* and *S. cerevisiae*. The final aim is to identify if any modules provide evidence of functional associations between ribosomal proteins (RPs) and energy metabolism, and whether these functional associations are the same as those observed in *S.*

cerevisiae. Although the conservation of riboneogenesis in *S. pombe* has been suggested [92], to date, there has not been a comprehensive study analysing riboneogenesis in *S. pombe.*

In this chapter, I report, that indeed functional modules can be derived from the *S. pombe* genome-wide expression network, and that the modules identified share a functional overlap to those found in *S. cerevisiae.* The most interesting result was the data suggested that the riboneogenesis pathway is conserved within *S. pombe*, but that the pathway may differ by a single enzyme. The data suggests that in *S. pombe,* the gluconeogenesis protein FBP1, may replace SHB17 as the enzyme responsible for thermodynamically driving riboneogenesis.

5.2 Methods

5.2.1 The dataset

The overall aim of this study is to identify and characterise the underlying regulatory network representing transcriptionally linked genes in *S. pombe.* To accomplish this task I used a publicly available *S. pombe* expression compendium containing over 1162 samples across 344 unique experimental conditions. [72]. Conditions included starvation, drug treatments and stress response [72]. A total of 5250 open reading frames were analysed in this study, however the dataset was thresholded to remove genes with 80% or more missing values across the samples bringing the total number of genes to 4254 (~88% of total genome). Though raw ratio signals were reported to be normalised to wildtype [72], visualisation of the data through box plots identified a handful problematic samples. A total of 119 samples were removed, therefore the final dataset contained 4524 genes and 1043 samples.

5.2.2 Network inference

The analysis pipeline used in this study is very similar to Chapters 2 and 3. Networks were inferred using the reverse engineering method ARACNE [30] [32]. No edges were eliminated using the data processes inequality (DPI). Statistically significant edges were selected using a p-value threshold of 10^{-78}, corresponding to a 0.25 MI value or greater (Table 5.1). Though the cut-off was chosen arbitrarily, it represented an extremely high threshold and approximately half the total number of genes were retained, consistent with the thresholds chosen for the *S. cerevisiae* expression analysis. Using the above threshold, a total of 2289 nodes and 135581 edges were retained within the network.

5.2.3 Network visualisation and modularisation

The network was visualised in Cytoscape [140] using a force directed layout. Consistent with prior network analysis, the GLay community clustering method [33] was used to identify highly connected modules. A second level of modularisation was used if a module was deemed too large (typically $\geq$ 300 nodes) or if modules were likely to reveal additional functional associations. Functional enrichment of each module was done using DAVID [21, 22]. Consistent with prior chapters, functional annotations were colour coded by their corrected FDR score, as detailed in section 2.3.3.

5.2.4 Identification of ribosomal protein first neighbours

Genes encoding RPs were classified into three groups based on their cellular location and annotations from the curated *S. pombe* database, Pombase [58]. The three groups were cytosolic RPs, mitochondrial RPs and ribosome biogenesis. The reason why cytosolic ribosomal factors were split into cytosolic RPs and ribosome biogenesis as opposed to just the two groups used in the *S. cerevisiae* analysis, was to thoroughly investigate which genes correlated with each ribosomal group. As prior to this study, there had not been a network based analysis focussing on the linkage between RPs and energy metabolism genes. I identified 139 genes that encoded cytosolic RPs, 70 that encoded mitochondrial RPs and 36 genes that encoded ribosome biogenesis genes. First neighbour networks for each ribosomal group were constructed. Visualisation and modularisation was done as described in section 5.2.3 to ensure consistency.

P-value	Associated MI
0.05	0.005715
0.01	0.007982
0.001	0.011226
1.00E-09	0.030685
1.00E-19	0.063118
1.00E-29	0.09555
1.00E-39	0.127983
1.00E-49	0.160415
1.00E-59	0.192848
1.00E-69	0.22528
1.00E-79	0.257713

Table 5. 1 ARACNE p-values and associated MIs for Bähler's data compendium.

5.3 Results

5.3.1 The modular structure of the *S. pombe* expression network

The hypothesis was that the modular structure of the *S. pombe* expression network should to some degree resemble the functional modules identified in the *S. cerevisiae* expression network. Here, I investigated the *S. pombe* expression network, identifying modules of highly interconnected nodes characteristic of transcriptional correlation, applying the same analysis strategy used in Chapter 3. Functional analysis using DAVID revealed that the modular structure of the *S. pombe* expression network reflects functional compartmentalisation, and that there is some overlap with the *S. cerevisiae* expression network (Chapter 3). 83% of sub-modules could be significantly functionally characterised by a specific functional profile (FDR $\leq$ 0.1).

The first level of modularisation identified nine network modules (Table 5.2). Module 1 (Figure 5.1, red nodes) is the largest, containing 1041 nodes and 63723 edges. The most enriched functions include genes encoding ubiquitination machinery, ribosome biogenesis and RNA processing (Figure 5.2). Module 2 (Figure 5.1, yellow nodes) shows significant enrichment in protein biosynthesis, nitrogen biosynthesis and cytosolic ribosome (Figure 5.3). Module 3 (Figure 5.1, blue nodes) is significantly enriched in protein folding, proteasome and transposable elements (Figure 5.4). Module 4 (Figure 5.1, purple nodes) is significantly enriched in cytokinesis and mitotic cell cycle (Figure 5.5). Module 5 (Figure 5.1, light blue nodes) is significantly enriched in mitochondrial energy production (Figure 5.6). Modules 6 through 9 were too small to undergo a further level of modularisation, they were however enriched in WTF proteins (a family of proteins with unknown function, often encoded at the end of long terminal repeats within the genome) and WD repeat proteins (module 6, Figure 5.7), endoplasmic reticulum and metal ion binding (module 7, Figure 5.8), thiamine biosynthesis (module 8, Figure 5.9) and finally cell surface and transmembrane proteins (module 9, Figure 5.10).

Module	**Colour**	**Number of nodes**	**Number of Edges**	**No. of sub-modules**	**Visualised in**
All		2289	135581	9	Figure 5.1
1	Red	1041	63723	4	Figure 5.2
2	Yellow	696	45181	3	Figure 5.3
3	Blue	153	756	5	Figure 5.4
4	Purple	108	680	5	Figure 5.5
5	Light Blue	43	171	3	Figure 5.6
6	Orange	35	74	1	Figure 5.7
7	Dark Green	9	13	1	Figure 5.8
8	Light Green	8	10	1	Figure 5.9
9	Black	8	14	1	Figure 5.10

Table 5. 2 Breakdown of modules identified by GLay for the *S. pombe* expression network

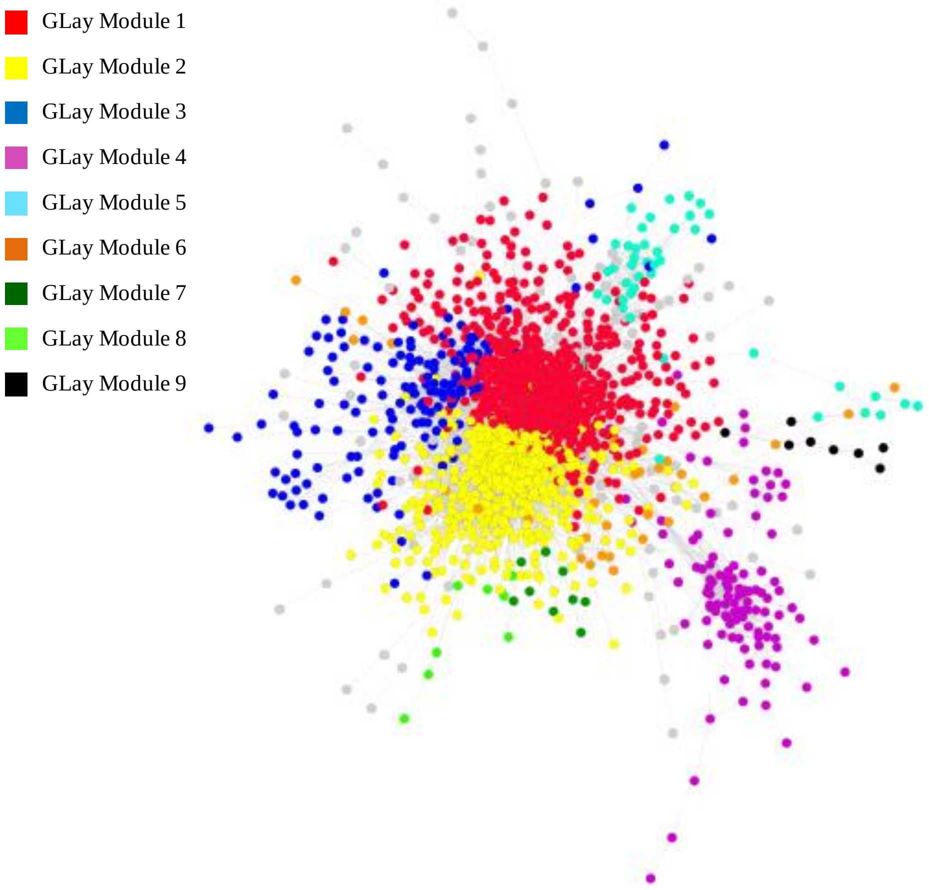


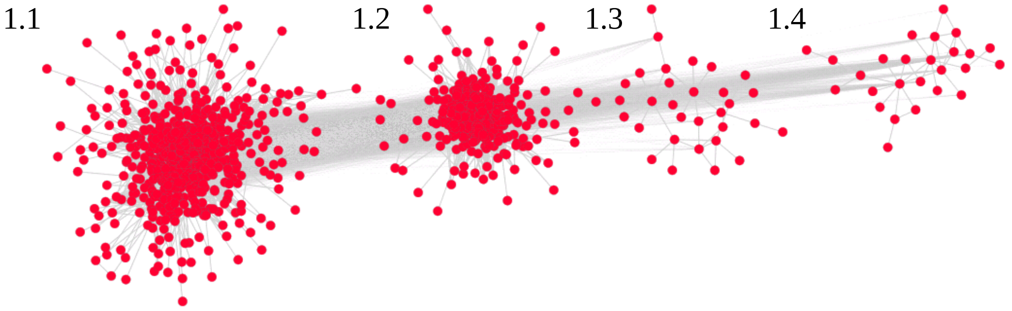
Figure 5. 1 *S. pombe* expression network showing modules mapped onto the parent network at 0.25MI threshold (p: 10^{-78})
The network has been visualised using a force directed layout. Node colour represents GLay module. Edge length is representative of MI value. The accompanying table (Table 5.2) shows the breakdown of each GLay cluster including the colour, number of nodes, and number of edges within each cluster.

5.3.1.1 Module 1: Cytoplasmic energy metabolism pathways and ribosome biogenesis

Module 1 was the largest module detected, and contained 1041 nodes and 63723 edges (Table 5.2). Four interconnected sub-modules were identified after a second round of modularisation (Figure 5.2). A summary of the most significantly enriched functions for each sub-module is shown in Figure 5.2 (Red text represents an FDR ≤0.05, green text represents an FDR≤ 0.1, and black text represents non-significant but enriched functions within the each sub-module). The raw DAVID files for each sub-module are available on the supplementary CD, in folder ‘Chapter 5’. The functional analysis of each sub-module revealed associations between genes encoding membrane, stress response and ubiquitination machinery (sub-module 1.1), ribosome biogenesis and RNA processing (sub-module 1.2), golgi apparatus and RNA splicing (sub-module 1.3) and nucleotide and zinc ion binding (sub-module 1.4).

The most striking observation is that module 1 was significantly enriched in glycolysis, pentose phosphate pathway (PPP) (sub-module 1.1) as well as ribosome biogenesis (sub-module 1.2). Specifically, sub-module 1.1 is significantly enriched of 11 aldo / keto reductase genes, a family of enzymes containing monomeric NADPH-dependent oxidoreductases involved in glycolysis and PPP [173]. The first neighbours of the 11 aldo / keto reductase genes, revealed a network containing 410 nodes, of which 62 were enriched in ribosome biogenesis functions (FDR $3.1x10^{-17}$, enrichment score 9.26). In sub-module 1.1 there was enrichment of six ribitol dehydrogenase enzymes, these enzymes are involved in the PPP. First neighbour analysis identified that of these six ribitol dehydrogenase genes, *SPCC736.13* and *SPAC521.03* (short chain dehydrogenases) have significant correlations with ribosomal biogenesis genes. Together *SPCC736.13* and *SPAC521.03* have over 364 first neighbours, of which 71 are ribosome biogenesis (FDR $2.7x10^{-27}$, enrichment score 15.17). The localisation of ribosome biogenesis genes to a single module is expected as ribosome biogenesis is a highly coordinated process [174].

Noteworthy is the enrichment of stress response genes in sub-module 1.1 and ribosome biogenesis genes in sub-module 1.2. The transcriptional coupling between these two processes is likely to be anti-correlated as suggested by my integrated analysis in *S. cerevisiae* and existing literature [161]. .



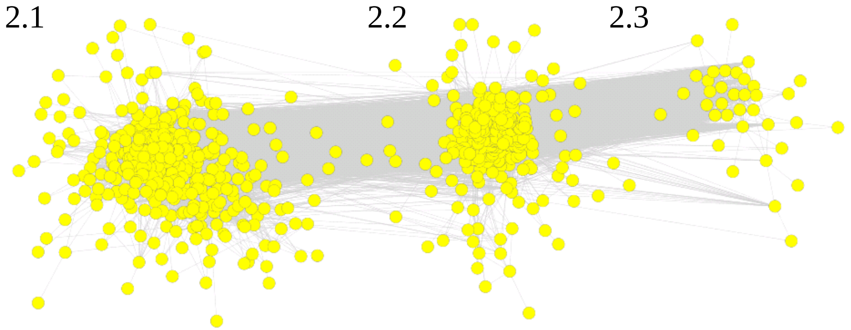
Module	Nodes	Edges	Functional Enrichment
1.1	636	26018	integral to membrane (79), ubiquitin-dependent protein catabolic process (22), gluconeogenesis (5), vacuolar transport (18), SNAP receptor activity (6), NAD(P)-binding domain (20), cellular response to heat (8), protein catabolic process(25), alcohol / Glucose/ribitol dehydrogenase(6), Aldo/keto reductase (11), Major facilitator superfamily MFS-1 (3)
1.2	344	19142	ribosome biogenesis (130), tRNA methylation (5), RNA polymerase complex (8), RNA degradation (6), snoRNA binding (17), RNA modification (14), Helicase-associated region (4), nuclear export (10), pyrimidine metabolism (6), Armadillo-like helical (7), aldolase / oxidoreductase (3), ncRNA processing (9), rrna processing, (5)
1.3	31	41	golgi apparatus (6), RNA splicing (3), hydrolase (6), DNA binding (4)
1.4	26	36	nucleotide binding (7), zinc ion binding (3), membrane (6)

Figure 5. 2 Sub-modular structure of module 1, with accompanying functional analysis. Red text represents an FDR ≤ 0.05, green text represents FDR ≤ 0.1, and black text represents non significant enrichment.

5.3.1.2 Module 2 – Protein biosynthesis and energy production pathways

Module 2 forms three sub-modules after a second round of modularisation (Figure 5.3). Sub-module 2.1 is significantly enriched in genes encoding protein biosynthesis, nitrogen binding, nitrogen compound biosynthesis, and energy metabolism processes suggesting a strong link between energy production and protein synthesis, as expected [138]. Sub-module 2.2 contains cytosolic RP biosynthesis, protein targeting and ribonucleoprotein complex. Located within this module are also four glycolysis genes *GPM1*, *FBA1*, *HXK2* and *GPD3*, of which *FBA1* and *GPM1* were identified as first neighbours of cytosolic RPs.

Sub-module 2.3 is significantly enriched only in translation initiation activity (FDR 3.37×10^{-4}).



Module	Nodes	Edges	Functional Enrichment
2.1	359	8976	protein biosynthesis (26), nucleotide-binding (49), nitrogen biosynthesis (82), IMP metabolic process (6), chaperonin-containing T-complex (4), NADP / electron carrier activity (8), ribosome biogenesis (8), oxidoreductase (28), amino acid biosynthesis (30), mitochondrion (44), pyruvate dehydrogenase complex (5), , NAD or NADH binding (8) , sulfur metabolic process (7), Flavoprotein (6), ligase (14), sulfur cluster binding (3),
2.2	300	18697	ribonucleoprotein complex / cytosolic ribosome (59), protein biosynthesis (17), protein targeting to membrane (8), nucleolus (26), Ribosomal protein L7A/RS6 family (3), Chaperonin Cpn60/TCP-1 (5), ribosome biogenesis (19), glycolysis (4), RNAPII (6)
2.3	37	188	translation initiation factor activity (6), ncRNA metabolism (4) r

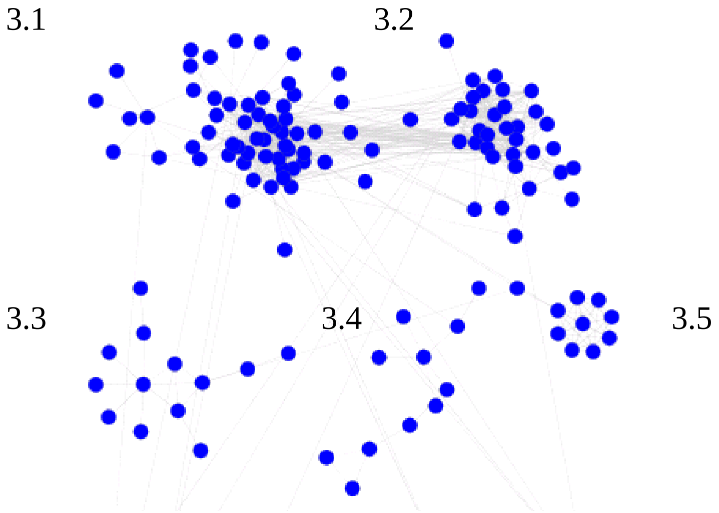
Figure 5. 3 Sub-modular structure of module 2, with accompanying functional analysis. Red text represents an FDR ≤ 0.05, green text represents FDR ≤ 0.1, and black text represents non significant enrichment.

5.3.1.3 Module 3 – The *S. pombe* stress response

Module 3 is significantly smaller than previous modules, containing only 153 nodes and 756 edges (Table 5.2). A second level of modularisation identified five interconnected sub-modules (Figure 5.4). Functional analysis of each sub-module revealed the association between stress response, protein folding and heat shock protein 70 (HSP70) (sub-module

3.1), proteasome components (sub-module 3.2), organic acid transport (sub-module 3.3) endoplasmic reticulum and zinc (sub-module 3.5) and finally transposable elements (sub-module 3.5).

A noteworthy observation is that sub-modules 3.1, 3.2 and 3.5 are significantly enriched in stress response (a feature also found in my *S. cerevisiae* expression network). In response to cellular stress misfolded proteins are targeted for degradation via the proteasomal degradation pathways [175] (sub-module 3.2). Heat shock proteins (HSP) such as HSP27 HSP70 and HSP90 identify misfolded proteins and through various mechanisms target them for proteasomal degradation [175] [176]. For example, HSP70, found in sub-module 3.1, binds hydrophobic patches which have been exposed due to protein misfolding and recruits CHIP (carboxyl terminus of HSP70-interacting protein), a co-chaperone / ubiquitin ligase to tag the protein for proteasomal degradation [175]. HSP27 and HSP90 utilise non-direct mechanism by which they act as chaperones and increase the activity of ubiquitin-proteasome degradation [176]. Transposition (sub-module 3.5) is closely correlated to heat shock response, a feature that has been reported in *Drosophila,* in which transposable elements target heat-shock promoters [177]. There are a total of 13 transposable elements (TEs) within the *S. pombe* genome, all of which are classified as Tf2 type [178]. The data suggests that module 3 has captured the general stress response in *S. pombe.* In *S. cerevisiae,* I showed that TEs are functionally represented in their own module, in *S. pombe* TEs group are located in a single sub-module with the stress response module. A key difference between *S. pombe* and *S. cerevisiae,* is that in *S. cerevisiae* there are 90 genes encoding transposable elements [58], this difference in the number of TEs may explain the lack of a TE dedicated network in *S. pombe*.

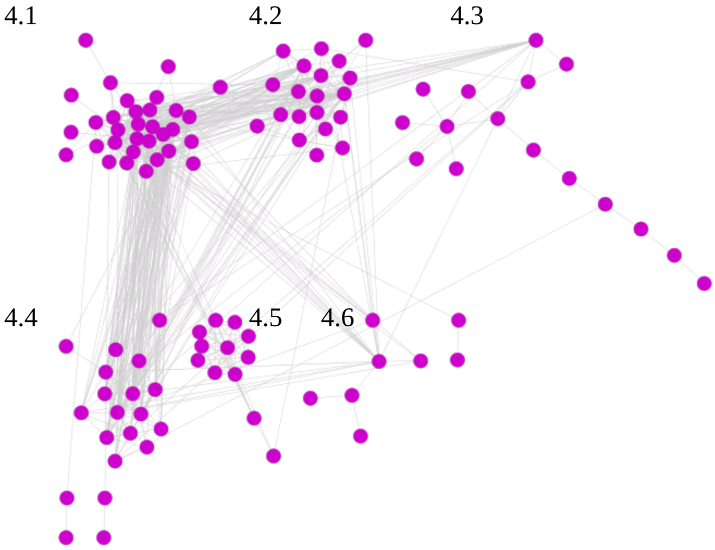


Module	Nodes	Edges	Functional Enrichment
3.1	59	355	protein folding (10), HSP 70(3), oxidation reduction (8), , unfolded protein binding / stress response (6), hexose catabolic process (3),
3.2	35	229	Proteasome (15), endopeptidase activity (13), proteasome beta-subunit complex (3), proteasome regulatory particle (7)
3.3.	13	15	organic acid transport (5), mitochondrial (3)
3.4	11	11	Zinc (3), endoplasmic reticulum (3)
3.5	10	37	transposable element (6)

Figure 5. 4 Sub-modular structure of module 3, with accompanying functional analysis
Red text represents an FDR ≤ 0.05, green text represents FDR ≤ 0.1, and black text represents non significant enrichment.

5.3.1.4 Module 4: The transcriptional coupling of cell cycle and cytokinesis genes.

Module 4 maps onto a distinct area of the parent network primarily isolated from other modules (Figure 5.1). Five smaller interconnected sub-modules, identified after an additional round of modularisation, represented the fine structure of this module (Figure 5.5). Functional analysis of the components within this module revealed a common functional association across all sub-modules. Sub-modules were either enriched in cytokinesis (sub-modules 4.1 and 4.4) or cell cycle / chromatin remodelling (sub-modules 4.2, 4.3, and 4.5). Genes encoding mitotic regulators, contractile ring cytokinesis proteins, mitotic spindle, DNA metabolism proteins and transcriptional regulators are all known periodically peak together during mitosis [179], consistent with the functional enrichment in module 4.

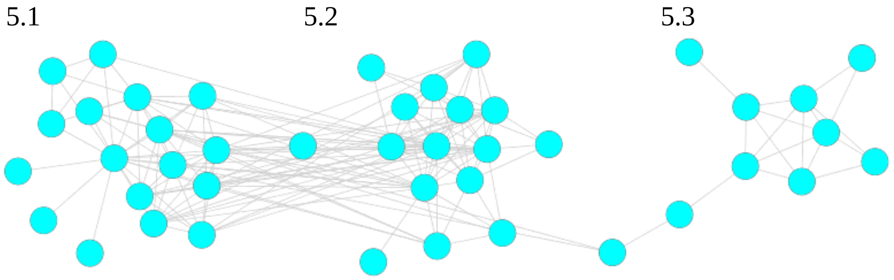


Module	Nodes	Edges	Functional Enrichment
4.1	32	227	cytokinetic process / cell division (11), cell surface (8), golgi apparatus (3)
4.2	20	72	mitotic cell cycle (10), mitotic cohesin complex (3), DNA metabolic process (6), protein-DNA complex assembly (3)
4.3	16	17	cell cycle, gpi-anchor biosynthesis (3), microtubule (3)
4.4	16	51	Cytokinesis (11), plasma membrane (6)
4.5	12	45	nucleosome core / Histone core (4)

Figure 5. 5 Sub-modular structure of module 4, with accompanying functional analysis
Red text represents an FDR ≤ 0.05, green text represents FDR ≤ 0.1, and black text represents non significant enrichment.

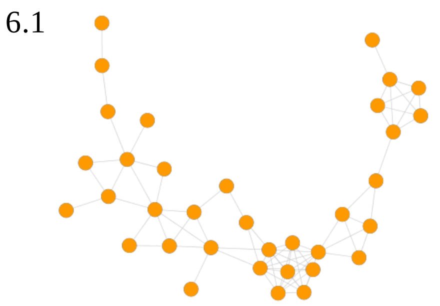
5.3.1.5 Module 5: Transcriptional coupling of mitochondrial energy production

A second level of modularisation identified three interconnected sub-modules (Figure 5.6). Functional analysis of these sub-modules revealed significant association between the expression of mitochondrial inner membrane and mitochondrial ATP synthesis (sub-module 5.1), mitochondrial membrane and mitochondrial respiratory chain (sub-module 5.2) and TCA cycle and electron carrier activity (sub-module 5.3). Every function identified within the sub-modules is significantly enriched, suggesting, with high confidence, that module 5 encapsulates energy production processes within the mitochondria. The localisation of mitochondrial energy generation to its own module was not found in the *S. cerevisiae* fitness and expression networks.



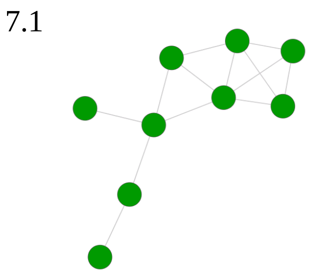
Module	Nodes	Edges	Functional Enrichment
5.1	18	53	mitochondrial inner membrane (15), oxidative phosphorylation (11), mitochondrial ATP synthesis coupled proton transport (5)
5.2	15	56	hydrogen ion transmembrane transporter / mitochondrial membrane (12), mitochondrial respiratory chain (8), oxidative phosphorylation (8), respiratory electron transport chain (5), mitochondrial proton-transporting ATP synthase complex coupling factor F(o) (4), transmembrane transport (6)
5.3	10	18	cofactor metabolic process (8), tricarboxylic acid cycle (6), electron carrier activity (4), mitochondrial lumen (3)

Figure 5. 6 Sub-modular structure of module 5, with accompanying functional analysis. Red text represents an FDR ≤ 0.05, green text represents FDR ≤ 0.1, and black text represents non significant enrichment.



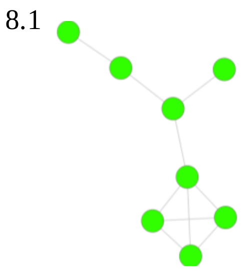
Module	Nodes	Edges	Functional Enrichment
6.1	35	61	WTF protein (5), wd repeat (3), metal ion binding (6)

Figure 5. 7 Structure of module 6, with accompanying functional analysis.
Red text represents an FDR ≤ 0.05, green text represents FDR ≤ 0.1, and black text represents non significant enrichment.



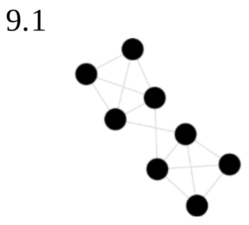
Module	Nodes	Edges	Functional Enrichment
7.1	9	13	endoplasmic reticulum (6), metal ion binding (3)

Figure 5. 8 Structure of module 7, with accompanying functional analysis.
Red text represents an FDR ≤ 0.05, green text represents FDR ≤ 0.1, and black text represents non significant enrichment.



Module	Nodes	Edges	Functional Enrichment
8.1	8	10	thiamin biosynthetic process (3)

Figure 5. 9 Structure of module 8, with accompanying functional analysis.
Red text represents an FDR ≤ 0.05, green text represents FDR ≤ 0.1, and black text represents non significant enrichment.



Module	Nodes	Edges	Functional Enrichment
9.1	8	14	cell surface (7), transmsmbrane protein (4) Schizosaccharomyces pombe

Figure 5. 10 Structure of module 9, with accompanying functional analysis.
Red text represents an FDR ≤ 0.05, green text represents FDR ≤ 0.1, and black text represents non significant enrichment.

5.3.2 Investigating the link between RPs and energy metabolism using a network based approach

Having demonstrated that cytosolic ribosomal factors were correlated to energy metabolism pathways in *S. cerevisiae*, I applied the same analysis to *S. pombe*. The aim was to identify if key genes known to be involved in glycolysis and riboneogenesis were correlated to ribosomal proteins in *S. pombe*, the results suggest that this is true. The analysis was conducted in three stages, identifying the first neighbours of cytosolic RPs (Figure 5.11), identifying the first neighbours of ribosome biogenesis genes (Figure 5.12) and finally identifying the first neighbours of mitochondrial RPs (Figure 5.13). A key difference between this ribosomal protein analysis in this study, and those reported in Chapters 2, 3 and 4 is that we constructed a separate network for just cytosolic ribosome biogenesis genes, instead of grouping them together with cytosolic RPs.

5.3.2.1 The first neighbours of cytosolic RPs

Of the 139 cytosolic RP genes within the *S. pombe* genome, 132 mapped into the parent network. The first neighbour network of cytosolic RPs contained 800 nodes and 89995 edges (Table 5.3). The first level of modularisation revealed three network modules (Figure 5.11).

Module 1 (Figure 5.11, red nodes) contains 418 nodes, however only 275 could be detected in DAVID (Table 5.3). As the *S. pombe* database in DAVID is not as extensive as the *S. cerevisiae* database, some genes are not detected due to the fact that the gene itself is not yet present within the database or because DAVID was unable to convert the gene ID to an alternative (detectable) format. Functional analysis of module 1 revealed associations between protein biosynthesis, amino acid biosynthesis, tRNA synthetases class II, ribosome subunits and oxidoreductase processes localised to the mitochondria (TCA cycle and electron transport activity). The finer structure of Module 2 was defined by two smaller sub-modules (Figure 5.11, blue nodes). Functional analysis revealed association between energy metabolism pathways, ubiquitination and proteolysis (sub-module 2.1) and ribosome biogenesis, ribosome export from nucleus and rRNA maturation (sub-module 2.2). The final sub-module contained only a ubiquitin ligase and an adenylate cyclise (module 3).

Module 1 represents the transcriptional coupling of high energy compound production and protein synthesis [180]. Notably, module 1 contains six glycolysis / alcohol catabolism genes; fructose-bisphosphate aldolase (*FBA1*), glycerol-3-phosphate dehydrogenase (*GPD2*), phosphoglycerate kinase (*GPM1*), pyruvate dehydrogenase (*PDB1*), phosphoglycerate kinase (*PGK1*) and ribose-5-phosphate isomerase (*SPAC144.12*). Though not classified as significantly enriched, these glycolysis related genes have a total of 145 direct edges to cytosolic RPs within module 1, with an average MI of 0.272 (Table

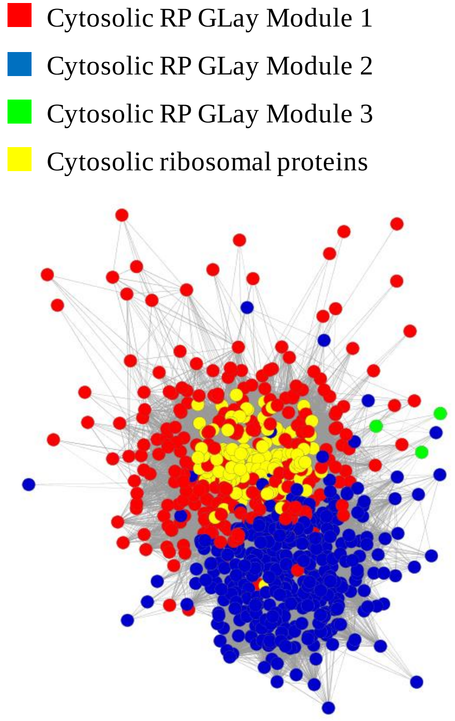
5.4) corresponding to a p-value of 10^{-84} (Table 5.2). *SPAC144.12* is connected to the most cytosolic RPs (75), consistent with reports that ribose-5-phosphate isomerase activity is required in the PPP and latter steps of riboneogenesis [92]. The association of ribose-5-phosphate isomerase to ribosomal proteins was also observed in my *S. cerevisiae* expression network (Chapter 3). FBA1 is directly connected 31 cytosolic RPs. FBA1 is essential for ribose-5-phosphate production, decreased aldolase activity has been reported to influence the production of sedoheptulose-1, 7-bisphosphate (SBP) and octulose-1-bisphosphate (OBP) [92]. These substrates are essential for the first committed step of riboneogenesis. The most significant edge between these glycolysis genes and cytosolic RPs is between *GPM1* and *RPL14* (MI score of 0.391). This was significantly higher than the average MI for *GPM1* (0.263, Table 5.4). The reason why *GPM1* is highly correlated to *RPL14* is currently unknown.

Module	Nodes	Edges	Found in DAVID
Overall	800	89995	
1	418	37067	275
2.1	192	9738	167
2.2	187	12129	170
3	3	3	2

Table 5. 3 Breakdown of nodes and edges within each sub-module from the cytosolic RP first neighbour network

Gene	Number of cytosolic RP first neighbours	Average MI
SPAC144.12	75	0.277
FBA1	31	0.273
GPM1	21	0.263
PDB1	16	0.259
PGK1	1	0.252
GDP2	1	0.292

Table 5. 4 The connectivity of the six module 1 glycolysis genes to cytosolic RPs


🟥 Cytosolic RP GLay Module 1
🟦 Cytosolic RP GLay Module 2
🟩 Cytosolic RP GLay Module 3
🟨 Cytosolic ribosomal proteins

1) protein biosynthesis (46), nitrogen compound biosynthesis (51), tRNA aminoacylation (18), translation (76), nucleotide-binding (67), aspartate family amino acid biosynthesis (13), Aminoacyl-tRNA synthetase, class II (9), carbon-oxygen lyase (7), ribosomal subunit (29), chaperonin-containing T-complex (8), intracellular protein transmembrane transport (10), magnesium ion (19), Glycine, serine and threonine metabolism (9), nucleotide biosynthesis (15), Ribosomal protein L7A/RS6 family (3), translational elongation (4), NAD (10), oxidoreductase (17), Ubiquitin (3), glycolysis / alcohol catabolism (6), acetyl-CoA biosynthesis (3), electron carrier activity (7), helicase (4), cell cycle / sexual reproduction (10)

2.1) regulation of gluconeogenesis / glucose metabolic process (4), autophagy (5), proteolysis (20), vacuole (13), phosphoinositide binding (6), phosphoric monoester hydrolase (6), asparaginase activity (3), external encapsulating structure(6), protein ubiquitination (10), ascospore formation (7), response to osmotic stress (5), cell wall (6), vacuolar transport (9), exopeptidase activity, tRNA methylation (3), protein kinase activity (8), alcohol catabolic process (3), oxidoreductase (8), ribosome biogenesis (3), mitochondrion (13), DNA repair (4)

2.2) ribosome biogenesis (107), maturation of SSU-rRNA (19), ribosome export from nucleus (11), WD40 repeat (11), tRNA metabolic process (16), ribosomal large subunit biogenesis (11), RNA-dependent ATPase activity (11), RNA modification (12), tRNA processing (8), transcription from RNAPI promoter (7), ncRNA 3'-end processing (5), RNA degradation (6), RNA recognition motif, RNP-1 (9), macromolecular complex assembly / disassembly (3), spindle (8), zinc-finger (9), mRNA metabolic process (7), integral to membrane (6), nuclear envelope (6)

3) ubiquitin protein ligase (1), Adenylate cyclase (1)

Figure 5. 11 A force directed layout of the first neighbours of *S. pombe* cytosolic RPs with GLay clusters mapped on.
Modules identified by GLay have been mapped onto the parent network. Edge length represents MI score. Cytosolic RPs are represented in yellow. Cytosolic RPs are most enriched in sub-network 1. Module 2 forms two; sub-modules. 2.1 is enriched glucose metabolic processes, protein modification and response to stresss. Sub-module 2.2 is enriched in ribosome biogenesis, nucleolus and ribosome export. Red text represents an FDR $\leq$ 0.05; green text FDR $\leq$ 0.1, black text represents non significant functional enrichment.

Interestingly, the most enriched functional annotation within sub-module 2.1 was four genes involved in regulation of gluconeogenesis / glycolysis (enrichment score 1.93, FDR > 0.1). Though these genes are not significantly enriched within DAVID, the edges to and from these genes are highly significant due to the stringent threshold used to construct the network (p-value: 10^{-78}).The four genes contained within this annotation cluster were identified as three ubiquitin ligases (*SPAC12B10.13*, *SPBC29A3.03C*, *SPBC106.13*) and transcription factor *RST2*. Sub-module 2.2 was significantly enriched entirely in ribosome biogenesis and RNA modification functions.

5.3.2.2 The first neighbours of mitochondrial RPs

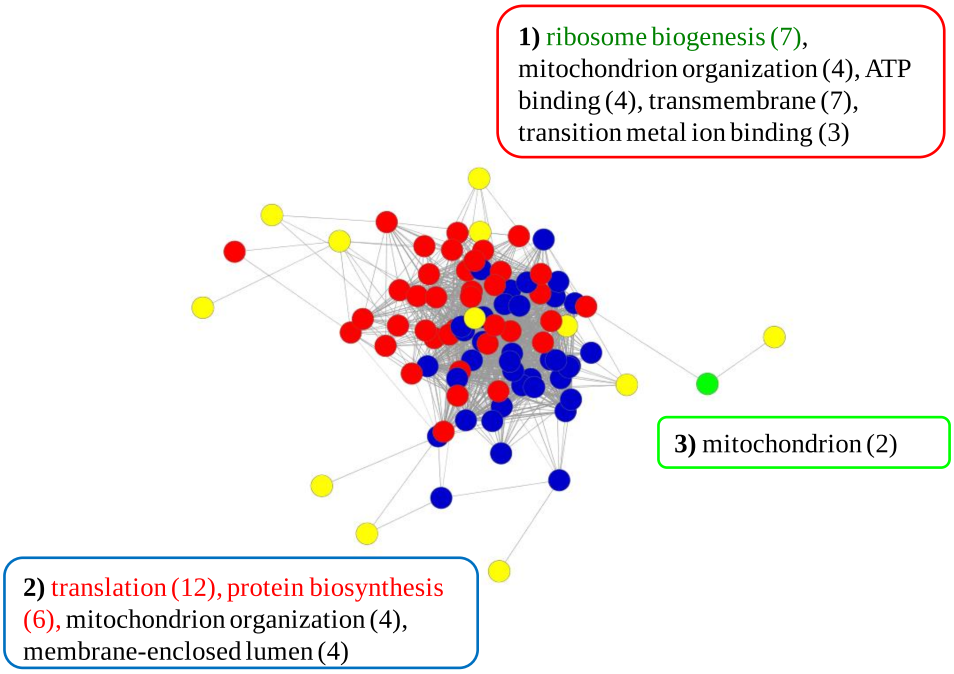
Of the 70 genes encoding mitochondrial RPs in the *S. pombe* genome, only 12 could be mapped into the parent network, the loss of 58 is likely due to the high statistical threshold used to construct the network. The first neighbour network contained 88 nodes and 1373 edges (Table 5.5). The first level of modularisation revealed three interconnected modules (Figure 5.12). Due to the size of the modules, only a single level of modularisation was required.

Module 1 (Figure 5.12, red nodes) reveals a functional association between ribosome biogenesis genes and mitochondrial organisation genes. Module 2 (Figure 5.12, blue nodes) reveals association between genes encoding translation protein biosynthesis and mitochondrial organisation. Module 3 (Figure 5.12, green nodes) is enriched in two mitochondrial genes (*ATP7*, *MRPL15*). The localisation of mitochondrial RPs and ribosome biogenesis genes within module 1 is consistent with reports that mitochondrial RPs and RNA polymerase (mRNAP) are encoded by nuclear genes, and are synthesised on cytosolic ribosomes as precursors proteins before being imported into the mitochondria

[181]. RNA processing reactions within the mitochondria are also catalysed by enzymes localised within the cytoplasm that are imported into the mitochondria when required [181] [152].

Module	Nodes	Edges	Found in DAVID
Overall	88	1373	
1.1	44	466	29
2.1	42	549	20
3.1	2	2	2

Table 5. 5 Breakdown of nodes and edges within each module from the mitochondrial RP first neighbour network



1) ribosome biogenesis (7), mitochondrion organization (4), ATP binding (4), transmembrane (7), transition metal ion binding (3)

3) mitochondrion (2)

2) translation (12), protein biosynthesis (6), mitochondrion organization (4), membrane-enclosed lumen (4)

Figure 5. 12 A force directed layout of the first neighbours of *S. pombe* mitochondrial RPs, with the Glay defined modules mapped on.
GLay modularisation identifies three modules. Edge length represents MI score. Mitochondtial RPs are represented in yellow. Red text represents an FDR ≤ 0.05; green text FDR ≤ 0.1, black text represents non significant functional enrichment.

5.3.2.3 The first neighbours of ribosome biogenesis proteins

Of the 36 genes annotated as ribosome biogenesis, 32 map onto the network. Identification of the first neighbours created a network containing 902 nodes and 107202 edges (Table 5.6). The first level of modularisation identified three modules; modules 2 and 3 did not undergo a further level of modularisation due to their size (Figure 5.13).

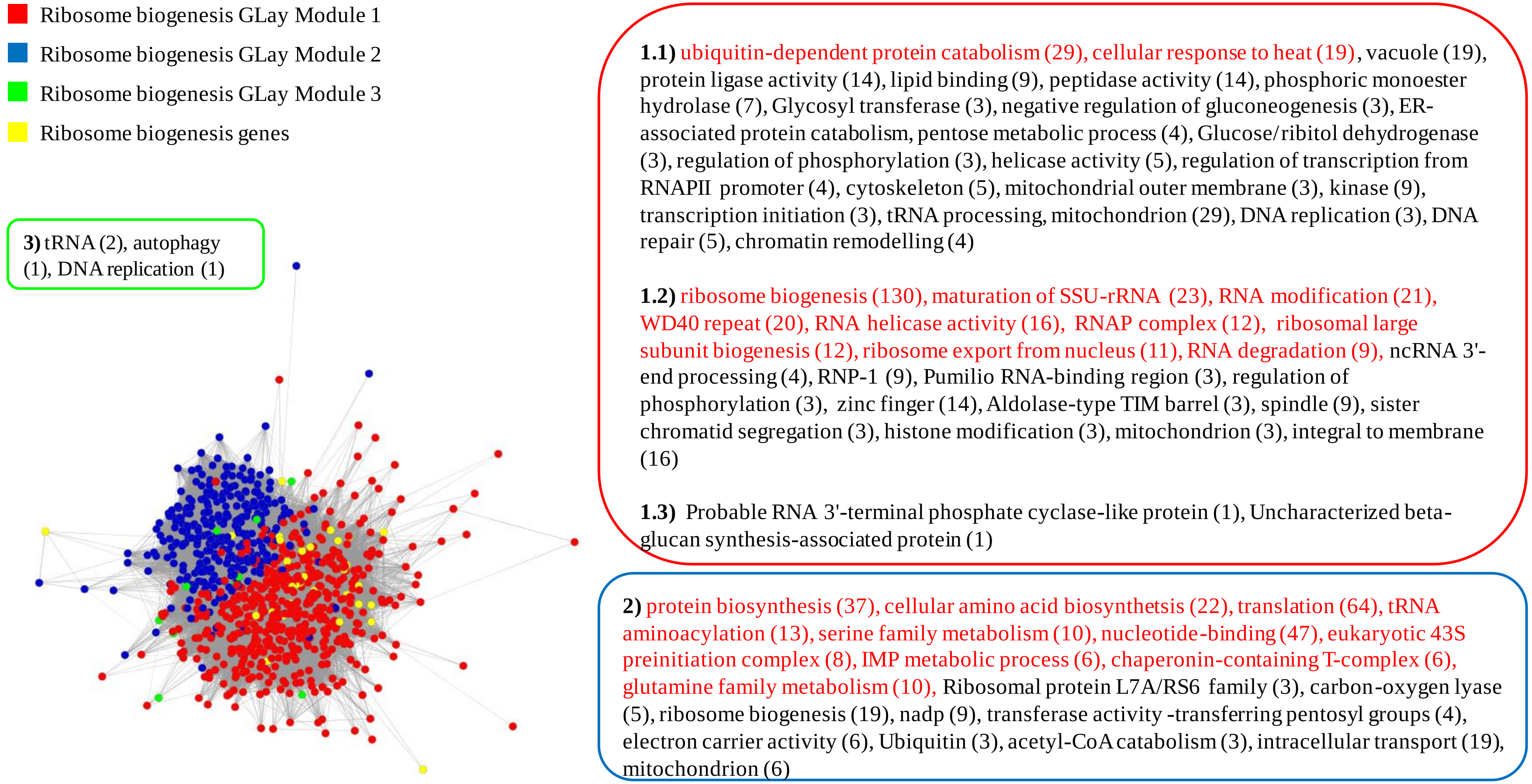
Module	Nodes	Edges	Found in DAVID
All	902	107202	
1.1	286	18730	255
1.2	273	16328	249
1.3	6	6	2
2	327	33052	200
3	10	14	7

Table 5. 6 Breakdown of nodes and edges within each module from the ribosome biogenesis first neighbour network

Module 1 (Figure 5.13, red nodes), formed three sub-modules, functional analysis identified significant associations between genes encoding ubiquitin-dependent protein catabolism and heat response (module 1.1), ribosome biogenesis, rRNA maturation and ribosome export from the nucleus (module 1.2), and a probable cyclase like protein and an uncharacterised beta synthesis protein (module 1.3). Module 2 (Figure 5.13, blue nodes), revealed associations between protein biosynthesis, amino acid biosynthesis and other processes involved in translation progression. Module 3 (Figure 5.13, green nodes), was the smallest of all modules, containing only four nodes involved in tRNA, autophagy and DNA replication.

Noteworthy, is the enrichment of ubiquitin-dependent protein catabolism and ribosome biogenesis genes in module 1. This functional association was also observed in my *S. cerevisiae* integrated network, suggesting that in response to the up-regulation of heat shock response genes, the expression of ribosome biogenesis genes are repressed [161]. Interestingly, there is enrichment of aldolase-type TIM barrel in module 1.2, which are typically found in class I aldolases, class I DAHP syntheases and class II fructose-bisphosphate aldolases [182] [183] [184]. Fructose bisphosphate aldolase (FBA1) is involved in riboneogenesis, suggesting that there is a link between ribosome biogenesis and energy metabolism pathways, a relationship link further fortified by the fact the studies in *S. cerevisiae.*

Module 2 is significantly enriched in protein synthesis / amino acid biosynthesis genes in addition to electron carrier and acetyl Co-A catabolic processes (pathways specific to mitochondria) indicative of the supply and demand of energy compounds required for protein synthesis.



🟥 Ribosome biogenesis GLay Module 1

🟦 Ribosome biogenesis GLay Module 2

🟩 Ribosome biogenesis GLay Module 3

🟨 Ribosome biogenesis genes

3) tRNA (2), autophagy (1), DNA replication (1)

1.1) ubiquitin-dependent protein catabolism (29), cellular response to heat (19), vacuole (19), protein ligase activity (14), lipid binding (9), peptidase activity (14), phosphoric monoester hydrolase (7), Glycosyl transferase (3), negative regulation of gluconeogenesis (3), ER-associated protein catabolism, pentose metabolic process (4), Glucose/ribitol dehydrogenase (3), regulation of phosphorylation (3), helicase activity (5), regulation of transcription from RNAPII promoter (4), cytoskeleton (5), mitochondrial outer membrane (3), kinase (9), transcription initiation (3), tRNA processing, mitochondrion (29), DNA replication (3), DNA repair (5), chromatin remodelling (4)

1.2) ribosome biogenesis (130), maturation of SSU-rRNA (23), RNA modification (21), WD40 repeat (20), RNA helicase activity (16), RNAP complex (12), ribosomal large subunit biogenesis (12), ribosome export from nucleus (11), RNA degradation (9), ncRNA 3'-end processing (4), RNP-1 (9), Pumilio RNA-binding region (3), regulation of phosphorylation (3), zinc finger (14), Aldolase-type TIM barrel (3), spindle (9), sister chromatid segregation (3), histone modification (3), mitochondrion (3), integral to membrane (16)

1.3) Probable RNA 3'-terminal phosphate cyclase-like protein (1), Uncharacterized beta-glucan synthesis-associated protein (1)

2) protein biosynthesis (37), cellular amino acid biosynthetsis (22), translation (64), tRNA aminoacylation (13), serine family metabolism (10), nucleotide-binding (47), eukaryotic 43S preinitiation complex (8), IMP metabolic process (6), chaperonin-containing T-complex (6), glutamine family metabolism (10), Ribosomal protein L7A/RS6 family (3), carbon-oxygen lyase (5), ribosome biogenesis (19), nadp (9), transferase activity -transferring pentosyl groups (4), electron carrier activity (6), Ubiquitin (3), acetyl-CoA catabolism (3), intracellular transport (19), mitochondrion (6)

Figure 5. 13 A force directed layout of the first neighbours of *S. pombe* ribosome biogenesis genes, with the GLay clusters mapped on. GLay modularisation identifies three modules. Ribosome biogenesis genes are represented in yellow. Edge length represents MI score. Module 1 underwent a further level of modularisation. Functional annotation for each module is shown in the coloured boxes. Red text represents an FDR ≤ 0.05; green text FDR ≤ 0.1, black text represents non significant functional enrichment.

5.4 Discussion

5.4.1 Does FBP1 in *S. pombe* take on the role of SHB17 from *S. cerevisiae* as the key enzyme responsible for regulating riboneogenesis?

To postulate that the riboneogenesis pathway may differ between *S. pombe* and *S. cerevisiae* comes from reports of key metabolism differences between the two species [66]. These include *S. cerevisiae* being able to utilise ethanol as a carbon source, whilst *S. pombe* cannot, instead ethanol is a waste product which may provide an advantage due to its toxicity against competing micro-organisms [66]. Also, the lack of key metabolic genes means that *S. pombe* does not have a glyoxylate cycle and glycogen biosynthesis pathway. Furthermore *S. pombe* has a lack of glycolytic paralogues, alcohol dehydrogenases, genes regulating transcriptional regulators of glucose repression and is completely missing the gluconeogenic enzyme phosphoenolpyruvate carboxykinase (*PCK1*) [66]. These reports highlight the degree of divergence in regards to carbon metabolism and energy production between *S. pombe* and *S. cerevisiae*. As there is a close relationship between glucose utilisation and riboneogenesis, the proposed hypothesis that FBP1 may substitute for SHB17 in *S. pombe* is not farfetched. This is further supported by the fact that *S. japonicus,* a yeast species within the same genus of *S. pombe* completely lacks the *FBP1* gene (Supplementary material [66]). This shows that even species within the same genus have major differences in how they utilise and metabolise carbon sources. The results presented in this chapter, together with existing knowledge of riboneogenesis and the structure of FBP1, suggests that FBP1 may have dual functionality within *S. pombe*. This is supported by other existing studies on FBP1 in yeast [58], photosynthetic bacteria and plants [185], which report that FBP1 has the potential to accept SBP as a substrate. Below I discuss the how FBP1 may act as the key riboneogenesis regulator in *S. pombe*.

5.4.2 Is the dual role of FBP1 in gluconeogenesis and riboneogenesis dependent on the structure of its active site?

Eukaryotic gluconeogenesis protein FBP1 is reported to have two binding sites; an active site located at the carboxyl terminal (C-terminal) and an allosteric binding site located at the amino terminal (N-terminal). The allosteric is typically bound by adenosine monophosphate (AMP) [186] [187], and in mammalian cells, the allosteric site has a conserved lysine located around residue 140 [188]. AMP binds the allosteric site and acts a modulator of FBP1 activity. In the cytosolic RP first neighbour network, I identified three ubiquitin ligases (*SPAC12B10.13*, *SPBC29A3.03C*, *SPBC106.13*) and transcription factor *RST2* as negative regulators of gluconeogenesis (Figure 5.11, module 2.1). This is significant, as repressing gluconeogenesis maintains flow through glycolysis which allows the glycolytic intermediates to be incorporated into the riboneogenesis pathway. The ubiquitin ligase SPBC106.13 is predicted to inactivate fructose bisphosphatase 1 (FBP1) [189].

Ubiquitination is a form of post-translational modification and is an essential mechanism for cellular control [190]. It is the process by which ubiquitin, a small 76 amino acid, 8.5 kDa protein [191] attaches to the target protein. Ubiquitin binds the target protein by either associating with the lysine residues on the protein substrate via an isopeptide bond, or alternatively it forms a peptide bond between the N-terminal of the substrate protein and the glycine 76 residue of the ubiquitin molecule [192] [190]. This is particularly important as the allosteric site of FBP1 is also located at the N-terminal and it is known to contain a conserved lysine residue [186] [187] [188]. In light of this evidence, it is possible that ubiquitin may act as an allosteric regulator, controlling the activity of FBP1 by affecting the hypervariable loop regions that are found at the N-terminal of eukaryotic FBP1 [187]. The association of ubiquitin to the allosteric site may cause a conformational change in the active site. One consequence of ubiquitin modification is that it can directly impact the conformation and activity of a target protein [191]. This subtle change in the structure of the active site may be the event which determines whether FBP1 has a higher affinity for FBP or SBP, and therefore determines whether cells enter riboneogenesis or gluconeogenesis.

Studies in *S. cerevisiae* have shown that FBP binds to the active site of FBP1 in a thermodynamically favoured cyclic β-furanose form [193]; SBP has also been reported to bind SHB17 in a higher similar manner, adopting a cyclic furanose form [92]. FBP has been reported to bind SHB17 in an extended linear form, just as SBP has been shown to bind FBP1 in an extended linear form, this highlights that both FBP1 and SHB17 have the potential to accept each other's substrates. However it is the extended linear form of the substrate that decreases the affinity of binding, hence why FBP1 preferentially binds FBP, and SHB17 preferentially binds SBP1 [92]. Therefore in *S. pombe*, I hypothesise that the ability of FBP1 to accept FBP or SBP in its cyclic form is the key determinant to whether FBP1 has a role in gluconeogenesis or riboneogenesis (Figures 5.14 and 5.15).

The addition of a ubiquitin molecule to the allosteric site of FBP1 may alter the structure of the active site, decreasing the Michaelis constant (K_m) and increasing V_{max} for SBP. Therefore the ability of FBP or SBP to bind in their cyclic form is dependent on the presence of an allosteric modulator binding to FBP1. The conformation of the sugar molecule (SBP or FBP) may account for the increased affinity with FBP1 due to additional hydrogen bonds being made between the sugar and active site of FBP1

Therefore under conditions when ribose demand is high, ubiquitin may bind the allosteric site of FBP1 with the aid of ubiquitin ligases, causing a conformational change in the active site. As a result of the conformational change, FBP1 may gain a higher affinity for

SBP, the key substrate required to drive riboneogenesis, and by doing so, effectively shuts down gluconeogensis. The flux through riboneogenesis is maintained and cells are able to meet the demand for ribose-5-phosphate, and growth continues with no limitations (Figure 5.14). During stress response such as glucose starvation, when the demand for glucose outweighs the demand for ribose-5-phosphate and cell growth, FBP1 keeps its canonical activity, catalysing FBP within the gluconeogenesis pathway. This maintains flux through the gluconeogenesis pathway and shuts down the riboneogenesis pathway (Figure 5.15).

Ubiquitination is a diverse cellular process, more importantly, it is also reversible [194] [191]. Deubiquitination may mean that when ribose demand decreases, ubiquitin disassociates from FBP1, returning FBP1 back to its original gluconeogenic function. By doing so, this creates an elaborate mechanism which that carefully regulates the rate of ribose-5-phosphate production whilst maintaining the pool of ubiquitin for use in other essential cellular functions. However, as previously stated, causality cannot be established due to the nature of the network inference method used, and this is a limitation of the study. Therefore experimental validation of this hypothesis is required (see section 5.4.3 for potential future experiments). A counter hypothesis could be that the correlation between FBP1 and ubiquitin ligases is a result of FBP1 becoming deformed as a consequence of stress exposure during the studies by Bähler’s Lab. It could be that the ubiquitination machinery is correlated to the expression of FBP1 in order to target it for accelerated degradation [195]. However, as ubiquitination involves a cascade of reactions including activation, conjugation and ligation [190] it is uncertain as to why only ubiquitin ligases are correlated to FBP1 and not other components of the ubiquitin dependent proteasome. The network analysis has provided evidence that there is a significant link between these FBP1 and ubiquitin ligases, however the only way to verify the true biological relationship between FBP1 and ubiquitin ligases would be through wet lab experiments.

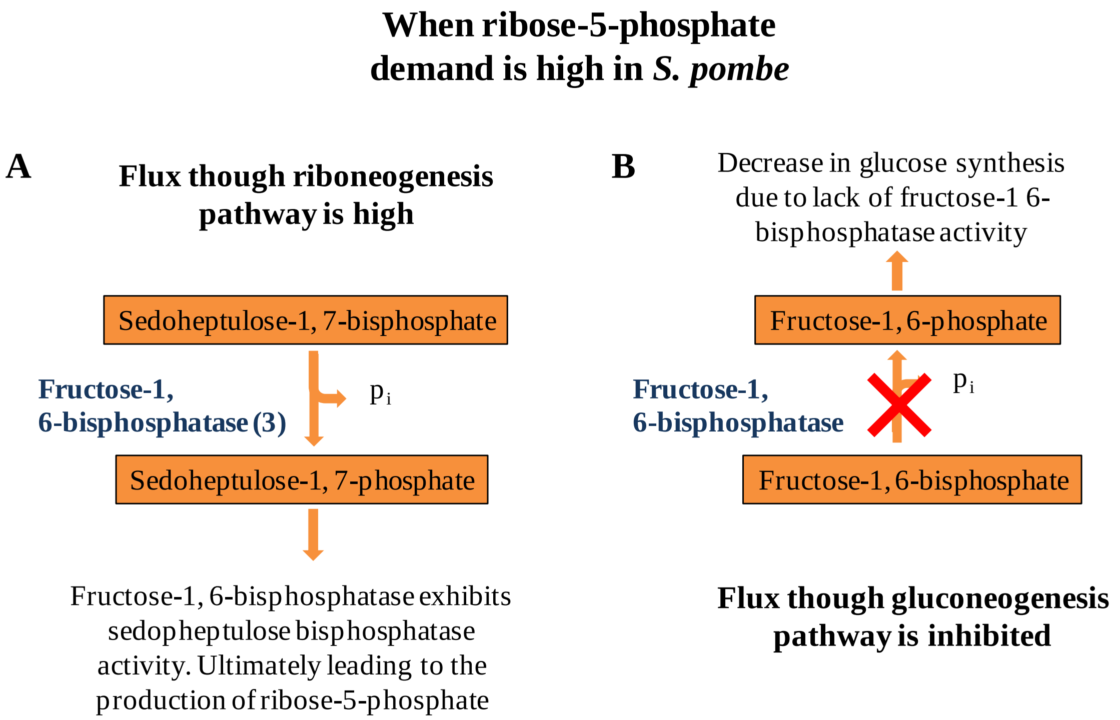

When ribose-5-phosphate demand is high in *S. pombe*
A
Flux though riboneogenesis pathway is high
Sedoheptulose-1, 7-bisphosphate
Fructose-1, 6-bisphosphatase (3)
p_i
Sedoheptulose-1, 7-phosphate
Fructose-1, 6-bisphosphatase exhibits sedopheptulose bisphosphatase activity. Ultimately leading to the production of ribose-5-phosphate
B
Decrease in glucose synthesis due to lack of fructose-1 6-bisphosphatase activity
Fructose-1, 6-phosphate
Fructose-1, 6-bisphosphatase
p_i
Fructose-1, 6-bisphosphate
Flux though gluconeogenesis pathway is inhibited

Figure 5.14 Flow chart showing the proposed dual role of fructose-1, 6-bisphosphatase (FBP1) in catalysing key reactions in riboneogenesis and gluconeogenesis in *S. pombe* when ribose-5-phosphate demand is high

Panel A. When ribose-5-phosphate demand is high, such as during rapid cell growth, FBP1 exhibits sedoheptulose-1, 7-bisphosphatase activity, thereby committing cells to riboneogenesis. The change in enzymatic activity may be due to ubiquitin binding to the allosteric site of FBP1 and altering the active site, allowing SBP to bind in its preferred cyclic conformation. The reaction catalysed by FBP1 is labelled (3) in association to the 3rd reaction catalysed by SHB17 in *S. cerevisiae* during riboneogenesis (Figure 1.2 in Chapter 1). Panel B. When ribose demand is high, FBP1 is unable to catalyse the rate limiting step in gluconeogenesis. The canonical enzymatic activity of FBP1 is restricted due to a change in the structure of the active site, due to ubiquitin binding to the allosteric site. FBP is unable to bind in its cyclic form, ultimately shutting off the gluconeogenesis pathway. Ultimately, FBP1 catalyses the thermodynamically driven step in riboneogenesis when ribose-5-phosphate demand is high.

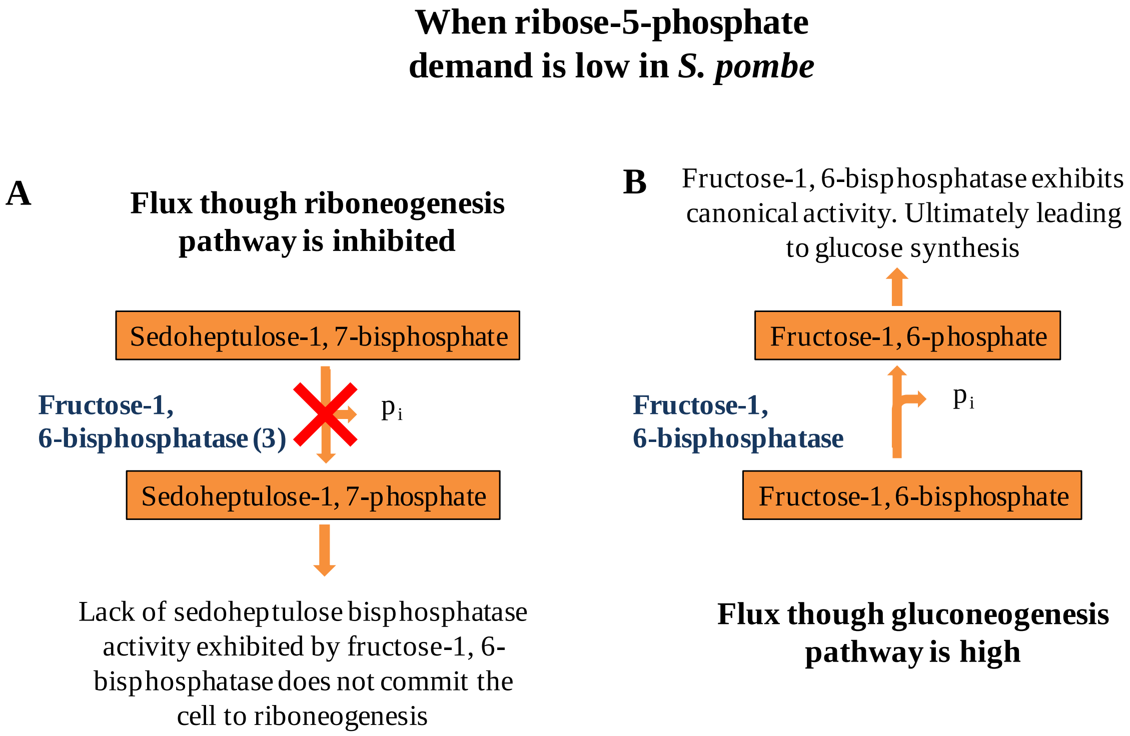

When ribose-5-phosphate demand is low in *S. pombe*
A
Flux though riboneogenesis pathway is inhibited
Sedoheptulose-1, 7-bisphosphate
Fructose-1, 6-bisphosphatase (3)
P_i
Sedoheptulose-1, 7-phosphate
Lack of sedoheptulose bisphosphatase activity exhibited by fructose-1, 6-bisphosphatase does not commit the cell to riboneogenesis
B
Fructose-1, 6-bisphosphatase exhibits canonical activity. Ultimately leading to glucose synthesis
Fructose-1, 6-phosphate
Fructose-1, 6-bisphosphatase
P_i
Fructose-1, 6-bisphosphate
Flux though gluconeogenesis pathway is high

Figure 5.15 Flow chart showing the proposed role of fructose-1, 6-bisphosphate in catalysing key reactions in riboneogenesis and gluconeogenesis in *S. pombe* when ribose-5-phosphate demand is low

Panel A. When ribose-5-phosphate demand is low, such as during times of glucose starvation, FBP1 exhibits its canonical fructose-1, 6-bisphosphatase activity. SBP is unable to bind to the active site in its cyclic form, instead only being able to bind in a linear extended form, as such the affinity for FBP is higher than that for SBP. The lack of sedoheptulose-1, 7-bisphosphatase activity stops cells from committing to riboneogenesis, thereby shutting down the riboneogenesis pathway. Panel B. When ribose-5-phosphate demand is low, FBP1 catalyses the rate limiting step in gluconeogenesis as it can accept FBP1 in its cyclic form, increasing flux through gluconeogenesis and increasing the rate of glucose formation. Ultimately, FBP1 catalyses the rate limiting step in gluconeogenesis when ribose-5-phosphate demand is low.

In addition, there were several connections between *RST2*, a zinc finger protein, reported to regulate the expression of *FBP1* [196] and genes encoding RPs. I investigated if there was transcriptional correlation between *RST2* and *FBP1* by identifying the first neighbours of *RST2* within the parent network. No direct connections between *RST2* and *FBP1* existed. Noteworthy however is that *RST2* was significantly correlated to protein kinase 1 (*PKA1*), which is known to negatively regulate both sexual development and gluconeogenesis by suppressing transcription of *FBP1* [196]. During glucose starvation, *FBP1* expression increases via the reduction of PKA1 activity, leading to the production of glucose via gluconeogenesis [196].

A bioinformatic means of supporting *S. pombe FBP1* as an orthologue of *S. cerevisiae SHB17* could be garnered using reciprocal BLAST searches. However, multiple reciprocal BLAST searches using both the protein (and nucleotide) sequence of *S. pombe FBP1* against the *S. cerevisiae* genome (and vice versa) did not yield any evidence that *S. pombe FBP1* may be an orthologue of *S. cerevisiae SHB17*. The results of BLASTP are shown in the appendix (Table A5.1 and Table A5.2). Furthermore, no evidence of *FBP1* being an orthologue of *SHB17* was obtained when using web-based reciprocal BLAST algorithms HomoloGene [197] and Inparanoid [198] (data not shown). A possible explanation for this result is that several classes of genes evolve more rapidly within fission yeasts than budding yeasts; these include genes involved in glycolysis and respiration. Conversely, genes involved in ribosome assembly have been reported to evolve slower in fission yeast than in budding yeast (Supplementary Material of [66]). Reciprocal BLAST does not entirely take into account the history where gene duplication has occurred. This is noteworthy as the evolution of *S. cerevisiae* to utilise ethanol as a carbon source occurred after a whole-genome duplication event [66]. Furthermore, the highly divergent nature of genes involved in carbon metabolism within the *Schizosaccharomyces* species itself (such as the lack of *FBP1* in *S. japonicus*) may explain why *S. pombe FBP1* was not identified as an orthologue of *SHB17*. The most appropriate way to prove or disprove the hypotheses derived within this chapter would be through experimental validation.

5.4.3 Identifying candidates for experimental validation

The work in this chapter has cumulated in a hypothesis that FBP1 has a dual role in *S. pombe*. This hypothesis is built upon the bioinformatic methods and congruent literature, what it lacks is experimental validation. It has however, identified candidate genes which could be used to test and validate the findings reported in this study.

Potential future experiments include deleting *FBP1,* then increasing / decreasing ribose-5-phosphate demand and measuring the accumulation of metabolites by using a labelled carbon source. If FBP1 does have dual functionality then it would be expected that when ribose-5-phosphate demand is high, there would be an accumulation of labelled SBP (as the deletion of *FBP1* would mean SBP would not be catalysed), conversely when ribose-5-phosphate is low, then there would be an accumulation of labelled FBP (as the deletion of *FBP1* would mean FBP would not be catalysed). Another experiment could involve deleting *TAL1* and *NQM1*, two transaldolase genes that are involved in catalysing reactions in the non-oxidative arm of the PPP. By inhibiting the non-oxidative PPP, the cell effectively loses one of its means of producing ribose-5-phosphate, therefore flux through riboneogenesis should increase in order to fulfil the demand of ribose-5-phosphate. This means that FBP1 would be catalysing the thermodynamically driven reaction within riboneogenesis at a higher rate than usual during times of high ribose demand. The flux through FBP1 could be quantified by using labelled glucose and measuring the accumulation of sedoheptulose-7-phosphate (the product of FBP1 catalysing SBP). Finally, I hypothesised that ubiquitin may act as an allosteric regulator of FBP1 by binding to its allosteric site. In order to test this, yeast strains containing a knockout of the the ubiquitin ligase *SPBC106.13* could be grown under conditions when ribose-5-phosphate demand was high (the demand could be increased by depleting the ribonucleotide pool) and then measuring metabolite flux through FBP1. If the binding of ubiquitin to FBP1 is truly the event which changes the affinity of FBP1 for SBP, then there would be an accumulation of FBP and a depletion of SBP.

5.4.4 There is evidence of riboneogenesis conservation across yeast species

To date, there have been no in-depth investigations into riboneogenesis in *S. pombe*. Given that there exist many key differences in carbon metabolism between *S. cerevisiae* and *S. pombe*, the results so far suggest that riboneogenesis is conserved within *S. pombe,* however the key difference appears to be the enzyme that thermodynamically drives riboneogenesis

The analysis of the *S. pombe* expression network ascertained the degree of overlap with the *S. cerevisiae* expression network. These included the separation of ribosome biogenesis and ribosomal proteins to separate modules (*S. pombe*: Figures 5.2 and 5.3 respectively, *S. cerevisiae*: Figures 3.2 and 3.3 respectively), the over representation of a module enriched in stress response genes (*S. pombe*: Figure 5.4, *S. cerevisiae*: Figure 3.4) and the co-expression of RPs, cell cycle and cytosolic energy metabolism genes (*S.* pombe:Figure 5.3, *S. cerevisiae*: Figure 3.2). The most significant result was the apparent transcriptional coupling of genes involved in riboneogenesis to ribosomal proteins. I identified that many genes involved in glycolysis and PPP were significantly coupled with ribosomal proteins, including *FBA1*, *GPM1*, *PGK1*, *TPI1*, *ENO101* and *SPAC144.12* (ribulose-5-phosphate isomerase) were all first neighbours of cytosolic ribosomal proteins. This is particularly important as the same (or homologous) genes were also identified as first neighbours of cytosolic RPs in my *S. cerevisiae* expression network analysis, including *FBA1*, *PGK1*, enolase enzymes (*ENO1*, *ENO2*), triose phosphate dehydrogenase enzymes (*TDH1*), and *PDB1*. These results are consistent with reports that riboneogenesis is heavily dependent on glycolytic intermediates [92]. Furthermore, my network analysis in *S. pombe* suggests that FBP1 substitutes for SHB17 in providing the sedoheptulose-1, 7-bisphosphatase required for riboneogenesis and that the mechanism involved in switching FBP1 activity is dependent on the cellular demands of the cell.

5.5 Concluding remarks

This chapter reported the construction and interrogation of a *S. pombe* gene expression network, using one of the largest and most comprehensive expression compendiums available. This chapter focused on characterising an *S. pombe* gene expression network and elucidating the degree of conservation of riboneogenesis in *S. pombe.* The constructed network identified multiple connections between genes encoding ribosomal proteins and energy metabolism enzymes, consistent with current literature [92]. The work reported in this chapter is the first time a global gene expression network in *S. pombe* has been constructed with the purpose of tackling the degree of conservation of riboneogenesis in *S. pombe*. Furthermore, the network analysis suggested FBP1 may substitute for SHB17 in *S. pombe* having functions in both gluconeogenesis and riboneogenesis. Though *FBP1* did not appear as an orthologue to *SHB17* during reciprocal BLAST searches, there is however multiple sources of evidence from both this study and existing literature that FBP1 has the potential to accept both SBP and FBP as substrates. Research into the *Schizosaccharomyces* genus showed how carbon utilisation and energy production differs compared to the *Saccharomyces* genus suggesting that the potential dual nature role of FBP1 may be possible. These results however are limited by the lack of experimental evidence demonstrating the role of FBP1 in riboneogenesis. Experimental work focused on elucidating the dual role of FBP1 is essential in order to validate the hypotheses developed in this chapter.

Similarly to the networks constructed in prior chapters, a genome-wide network inference approach was used. This means that although the focus was on understanding riboneogenesis in *S. pombe,* the network can in fact be interrogated and queried in regards to any other biological process. This is particularly important, as comprehensive compendia such those used in this study are lacking in *S. pombe,* therefore this network provides a foundation upon which to build hypotheses and aid in the identification of potentially novel gene relationships.

The next chapter reports the genome wide binding of a group of representative 60S cytosolic RPs in *S. pombe*. The aim is to identify the genome-wide binding pattern of these RPs, and possibly elucidate the mechanisms by which *FBP1* expression is regulated in response to cellular demand.

CHAPTER 6: THE GENOME-WIDE ASSOCIATION OF RIBOSOMAL PROTEINS IN *Schizosaccharomyces pombe*

6.1 Abstract

In Chapter 2, I reported that a network inference approach analysing fitness data identifies highly connected modules that are representative of key cellular functions. An outcome of this analysis was the phenotypic linkage between ribosomal factors and cytoplasmic energy metabolism processes. The expression network analysis described in Chapter 3 showed that one of the most conserved relationships is that between glycolysis and ribosome biogenesis in *S. cerevisiae*. Co-fitness and co-expression between these genes were also found in the integrated network (Chapter 4). In Chapter 5, I performed a similar network analysis in *S. pombe*. The results indicated that fructose-1, 6-bisphosphatase (FBP1) may be required for riboneogenesis in *S. pombe*. This chapter reports the RNA-dependent interactions between RPs and many transcription sites which suggest that RPs bind as components of a preassembled multi-protein complex. It also reports that RPs associate with a wide assortment of genomic loci, notably heterochromatin, tRNAs and genes encoding proteins involved in glycolysis and riboneogenesis including *FBP1* as suggested by my *S. pombe* network analysis (Chapter 5). This chapter concludes by hypothesising a possible mechanism in which ribosomal proteins regulate target genes by affecting transcription and translation.

6.2 Introduction

Ribosomal proteins (RPs) are a main component of ribosomes [199] [200], RPs are thus believed to be only present in the cytoplasm of eukaryotes. However, the unexpected finding that at least 20 RPs and tRNAs are present at transcription sites in *Drosophila*, suggested that ribosomal subunits may actually associate to nascent mRNAs [201]. Previous studies have reported that RPs bind to non-coding RNA genes in *S. cerevisiae*, suggesting that RPs association to nascent mRNAs may involve free RPs that are not part of the ribosome, indicating that association is independent of gene translation [202]. The possibility that RPs have extra ribosomal functions is not novel, several RPs have been reported to have extra ribosomal functions. Some RPs are able to regulate their own expression by binding their own mRNA or promoters and affecting transcription splicing or translation [100] [98] [101] [99]. RPs have also been reported to bind transcription factors at the promoters of other genes. Examples include ribosomal protein S3 (RPS3) in human, which regulates a subunit of the NF-κB DNA-binding complex involved in chromatin binding and transcription regulation [203]. Ribosomal protein L22 (RPL22) and other RPs, bind histone protein H1 and suppress transcription in *Drosophila* [204].

Although consensus is that RPs have specific functions at specific genes, it is unclear why multiple RPs are found together at the same transcription sites of unrelated genes. If binding of each RP occurs individually, then association to genomic loci would be dependent on its RNA binding or protein binding affinity, and if this is truly the case then it wouldn’t explain why several RPs are found together at the same sites. However if we assume that the RPs associated to chromatin as part of a non-functional silent complex, then the observation that RPs are often associated to the same genomic loci would then make sense. Conversely the presence of RPs at transcription sites may not be functionally significant, the association to chromatin may be due to excess RPs that are not incorporated into ribosomes simply interacting non-specifically with nucleic acids [205]. RPs are basic (pl > 10) [50] so at higher concentrations it is possible that they may associate to chromatin, however studies have shown that mechanisms exist that degrade excess RPs in order to maintain low cellular concentrations of RPs [200] [199] [205].

Here, we investigate the genome-wide association of three representative 60S RPs in *S. pombe*. ChIP assays, ChIP-chip and subsequent wet lab experiments were performed by Sandip De, a collaborator and PhD student at The University of Birmingham. Here we report a bioinformatic analysis of the genome wide association of RPL7, RPL11 and RPL25. Notably, we identified that these RPs have a common set of at least 178 transcriptional loci including 74 protein coding, 36 non-coding and 64 coding tRNAs. We also demonstrate RPs bind the centromeric regions of all three *S. pombe* chromosomes. The similar binding profiles of the three RPs suggest that they are bound as components of complexes consisting of multiple proteins. Further analysis revealed that seven glycolysis genes and gluconeogenesis gene *FBP1* are significantly (p-value: 10^{-4}) associated to all three RPs, furthermore a subset of these genes were identified as first neighbours of ribosomal factors in the *S. pombe* expression network (Chapter 5). The direct binding of RPs to glycolysis genes suggests a regulatory mechanism in which RPs control their own synthesis by limiting the availability of glycolytic intermediates.

This analysis was conducted in collaboration with Sandip De, who performed all lab experiments, whilst I performed all bioinformatic analysis. The genome-wide association of RPs to *S. pombe* chromosomes is published in De *et al.* 2011 [50]. In this chapter, I report on the bioinformatic analysis of the ChIP-chip data required for the paper, before focussing on the verification of the linkage RPs to energy metabolism genes.

6.3 Methods

6.3.1 Experimental analysis

S. pombe transformation, imaging, RNA analysis, synchronisation of fission yeast cells, ChIP and ChIP-on-chip experiments were performed by Sandip De. Experimental details can be found in our paper [50].

6.3.2 Processing and visualisation of ChIP-chip data

We used the Model-based Analysis of Tiling Arrays (MAT) software for analysis of the Affymetrix hybridization data [52] together with a custom made 2011 BPMAP file. MAT software is specifically developed for the analysis of ChIP-chip data produced using tiling arrays [52]. MAT identifies genomic regions significantly bound by proteins on Affymetrix Tiling Arrays. ChIP input DNA was used as the control for the analysis and was compared against the RP data. A p-value of 10^{-4} was used; remaining MAT parameters remained as default. Results of MAT were visualised in Affymetrix's Integrated Genome Browser (IGB) [206].

6.3.3 Identification of enriched regions and calculation of enrichment scores

MAT only detects regions of the genome that are significantly enriched, therefore identification of genomic features such as coding regions, introns and repeat regions had to be detected using bioinformatic methods. Genomic features were considered significantly enriched if 50% or more of the feature was bound by the RP protein in question. To calculate the average enrichment score per feature, the probe by probe enrichment scores calculated by MAT were cross-referenced with genomic feature positions using an up to date *S. pombe* genome coordinates file and an average enrichment score was calculated between the start and end coordinates of each enriched genomic feature. Thereby giving each enriched genomic feature a score based on fold enrichment. Identification of significantly bound genomic features and enrichment score calculation was done using the statistical computing language R [11]. Functional annotation of the enriched features was done using the Database for Annotation, Visualization and Integrated Discovery (DAVID) [21]. Consistent with previous chapters, functional annotations provided by DAVID were classified into three groups, those with an FDR $\leq$ 0.05, those with an FDR $\leq$ 0.1 and those with no statistically significant functional enrichment.

6.4 Results

6.4.1 RPs associate both with coding and non coding genes

Chromatin-immunoprecipitated DNA was hybridised to *S. pombe* genomic tiling arrays (see Materials and Methods and De *et al.* 2011 [50]). We analysed three yeast strains expression, HA-tagged RPL7, RPL11 and RPL25, with two independent biological replicas of each using chromatin samples prepared at independent times from independent cultures. Significant binding sites for each RP were identified using MAT software [52]. Using Pearson correlation we identified a high probe-by-probe signal correlation between RPs and their corresponding replicas ($\geq$ 0.76), demonstrating highly similar binding between replicas [50]. The analysis revealed that the three RPs associated to many loci throughout the three chromosomes (Figure 6.1).

There are limitations to the *S. pombe* genome feature file (gff) used to identify the significantly associated genomic features. Not all features are well annotated and are therefore designated as being 'unknown'. Unknown regions are regions which are uncharacterised and do not have a gene name associated to them. For this reason the analysis was split to identify annotated and unknown regions respectively (Table 6.1). The number of loci bound by RPs differs slightly between what is reported in this chapter and what we reported in 2011 [50]. This is because in the latter analysis a new up-to-date binary probe map (BPMAP) was used. The BPMAP file contains genomic probe position information for Affymetrix Tiling Arrays, including the mapping of X/Y coordinates of each probe. Probes within BPMAP files are also classified as perfect match or mismatch. An updated BPMAP was required as mapping between probe and genomic positions change slightly as the genome becomes increasingly annotated. Previously we used an older BPMAP as it was the most recent BPMAP available at the time. The construction and inclusion of an up-to-date BPMAP in this bioinformatic ChIP-chip analysis therefore led to more accurate and reliable results. Annotations of significantly associated known and unknown genomic features are shown in figure 6.2.

Yeast culture	Statistically significant genomic features bound	Annotated within gff file	Classified as 'unknown'
RPL7	355	251	104
RPL11	381	319	62
RPL25	547	458	89

Table 6. 1 Total genomic features (annotated and unknown) bound by RPL7, RPL11 and RPL25 using a MAT p-value of 10^{-4}

Analysis showed that RPL25 binds additional loci compared to RPL7 and RPL11 (Table 6.1), however visual inspection of the data suggests that the binding profile for the RPs across the chromosomes are very similar (Figure 6.1). This observation was supported by statistical validation, by calculating the Pearson correlation between the binding profiles of each RP (Table 6.2). This suggests that the reason why RPL25 associate to more genomic features is because it binds the same loci as RPL7 and RPL11, but with a higher affinity, consequentially MAT identifies more regions significantly enriched by RPL25.

Pearson correlation between ribosomal proteins		Chr I	Chr II	Chr III
RPL7	RPL11	0.84	0.89	0.87
RPl11	RPL25	0.89	0.93	0.93
RPL7	RPL25	0.85	0.86	0.87

Table 6. 2 The Pearson correlation between the RP ChIP-chip binding profiles across the three *S. pombe* chromosomes

The similar binding profile of RPL7, RPL11 and RPL25 is verified using Pearson correlation. The table shows that the correlation coefficients lie between 0.84 and 0.93 signifying a highly similar binding pattern for RPL7, RPL11 and RPL25 across all three *S. pombe* chromosomes.

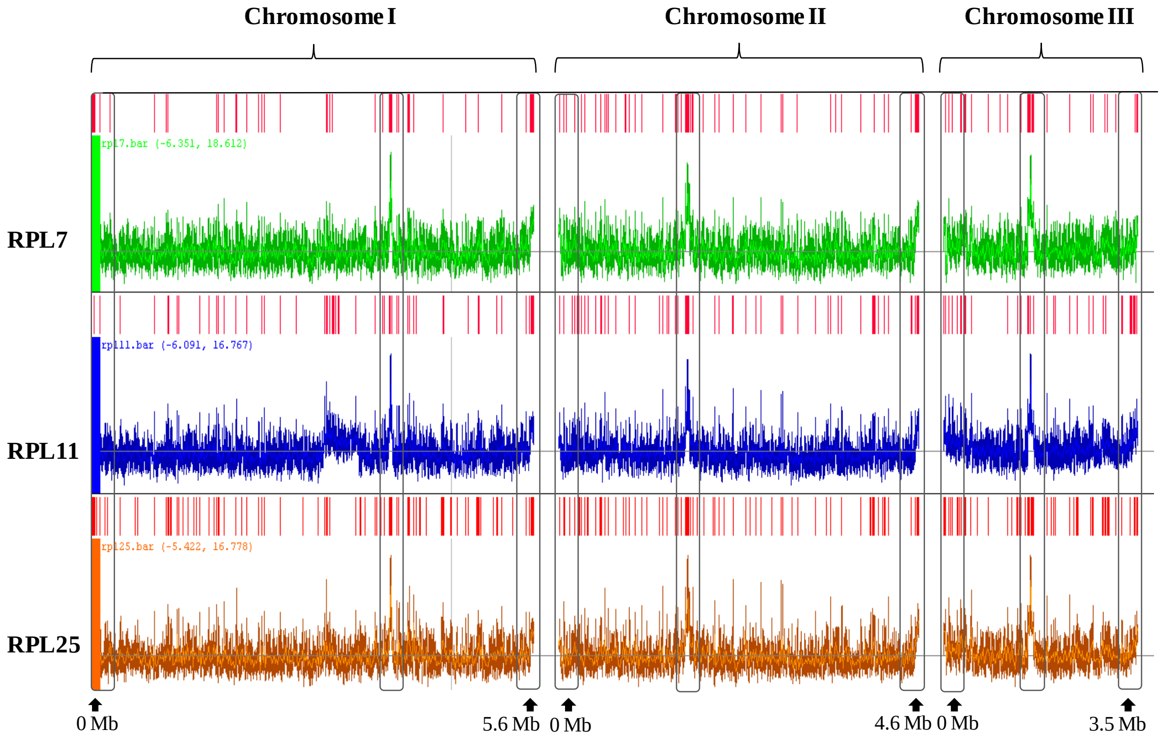


Figure 6. 1 The genome-wide association of the RPs to the *S. pombe* genome. Chromosomal binding profiles of the RPs across the three chromosomes visualised using IGB. Each RP is represented in a different colour, RPL7 (green), RPL11 (blue), RPL25 (orange). X-axis shows the distance from the left chromosome end in megabases (Mb). Chromosomes are separated as described at the top of the figure. Y-axis indicates the log2 MAT enrichment score. Regions identified as significantly bound by the RP is shown in the red boxes above each binding profile. The plot is based on two ChIP-chip biological replicas and two control arrays hybridised with input DNA, used as a standard across all yeast strains. Position of centromeres and telomeres are highlighted with the vertical boxes.

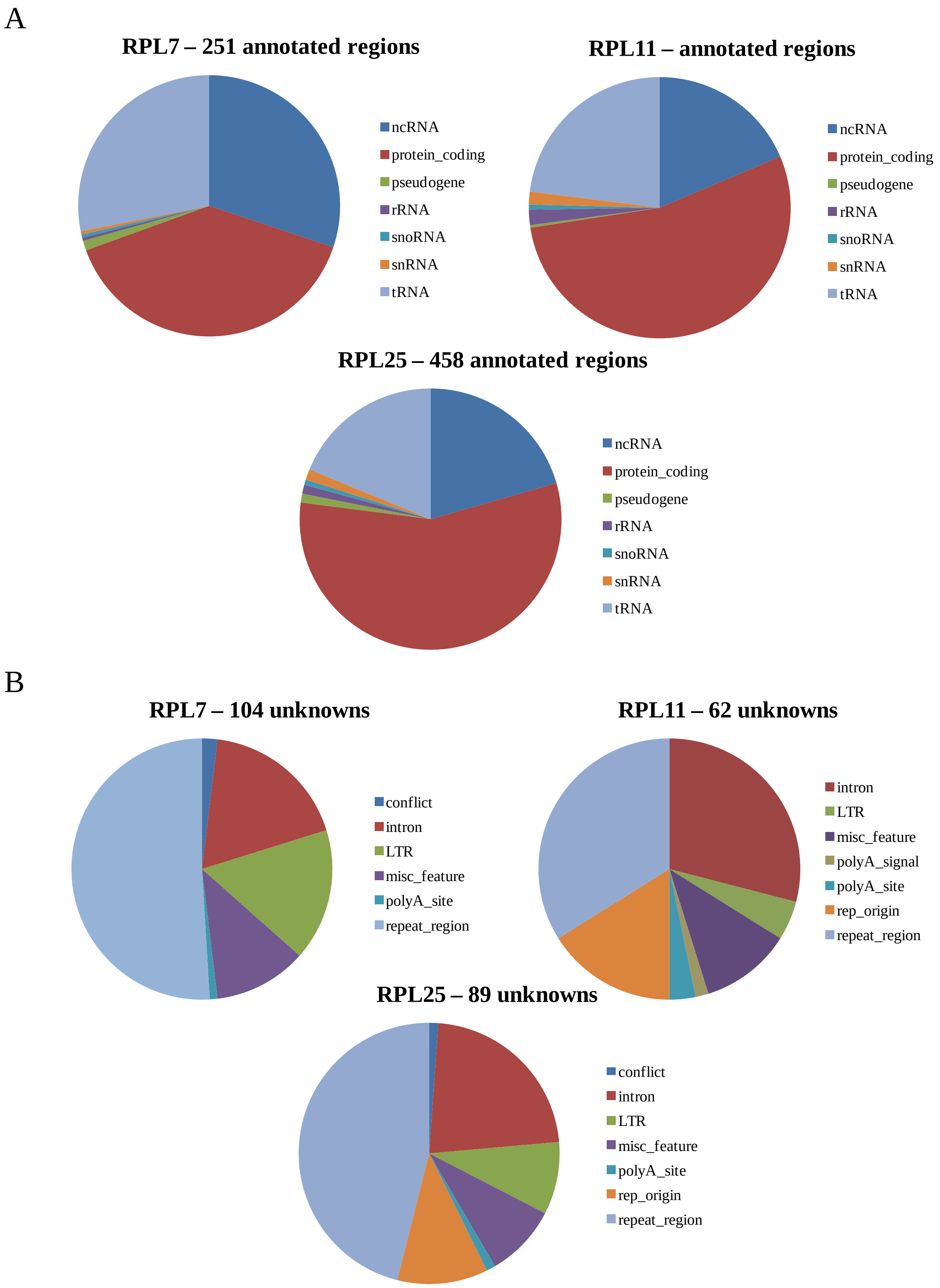


Figure 6. 2 Pie-charts showing the proportions of bound genomic regions.
Panel A. Significantly bound genomic features. Panel B significantly bound unannotated features for RPL7, RPL11, and RPL25

Results show that RPs are most associated to protein coding genes, ncRNAs and tRNAs (Figure 6.2A). A small proportion of hits corresponds to pseudogenes and other non-protein encoding RNAs (snoRNA, snRNA and rRNA) (Figure 6.2A). Analysis of enriched unannotated regions shows enrichment of specific gene features such as repeat regions, long terminal repeats (LTRs) and introns (Figure 6.2B). The data shows that RPs tend to associate to repeat regions (Figure 6.2B). Finally, RPL11 and RPL25 are clearly found at origins of replication (the association with RPL7 is visibly apparent but below significance threshold level we have used). Interestingly, all significantly bound origins of replication are located within a single dense cluster (Figure 6.3). Although a total of 401 strong DNA replication sites have been reported in *S. pombe* [207], there are a total of 16 confirmed genomic features annotated as 'origins of replication' in the 2011 *S. pombe* genome, of which ten are significantly bound by RPL11 and RPL25.

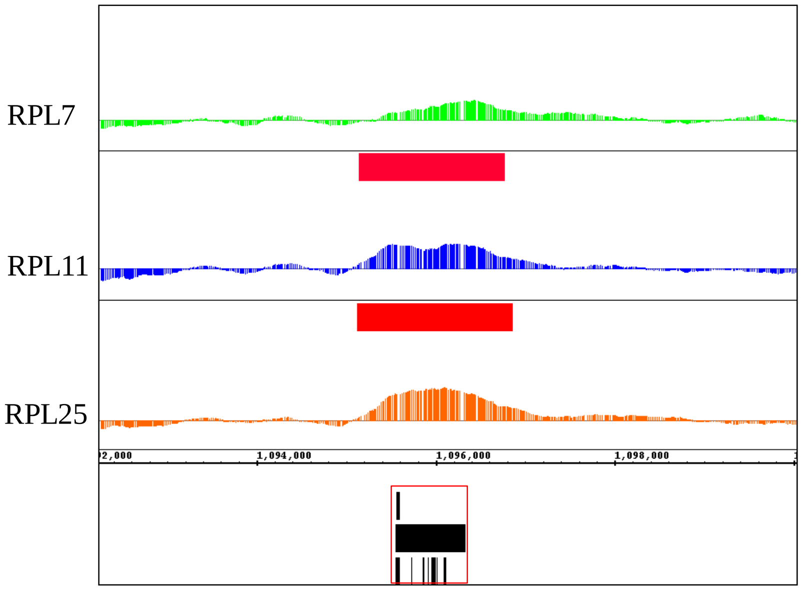


Figure 6. 3 The association of RPL7 RPL11 and RPL25 to origins of replication.
RPL7, RPL11 and RPL25 association are represented as green, blue and orange respectively. Red blocks represent significant binding as determined by MAT software using a p-value of 10^{-4}. The black blocks at the bottom of the figure represent the dense cluster of replication origins. The figure shows that only RPL11 and RPL25 have significant association to these regions, associating with all 10 replication origins within the tightly packed cluster.

6.4.2 RPs show significant association to specific regions of the centromere

RPs associate to centromeric regions (highlighted by the black boxes in Figure 6.1). We also find that enrichment is highest at tRNA genes found in dense clusters within the centromere (Figure 6.4). Fission yeast centromeres contain a central core of non-repetitive DNA (*cnt*), flanking the *cnt* region are two repeat regions termed innermost repeats (*imr*), followed by the outer repeats (*otr*); the outer repeats contain multiple copies of *dh and dg* repeats [208] [209]. The data demonstrates that RPs significantly associate to the centromeric regions of all three chromosomes, with association highest at the *cnt* and *imr* regions (Figure 6.4). Further investigation of this association using sequence specific ChIP and real time PCR confirmed the RP association to centromeres and that the association was sensitive to RNase treatment (data not shown) [50]. The *cnt* region was believed to be untranscribed [210], however a study in 2011 revealed that the cnt is transcribed but mRNAs are rapidly degraded [211], so mRNAs fail to accumulate to a measureable extent.

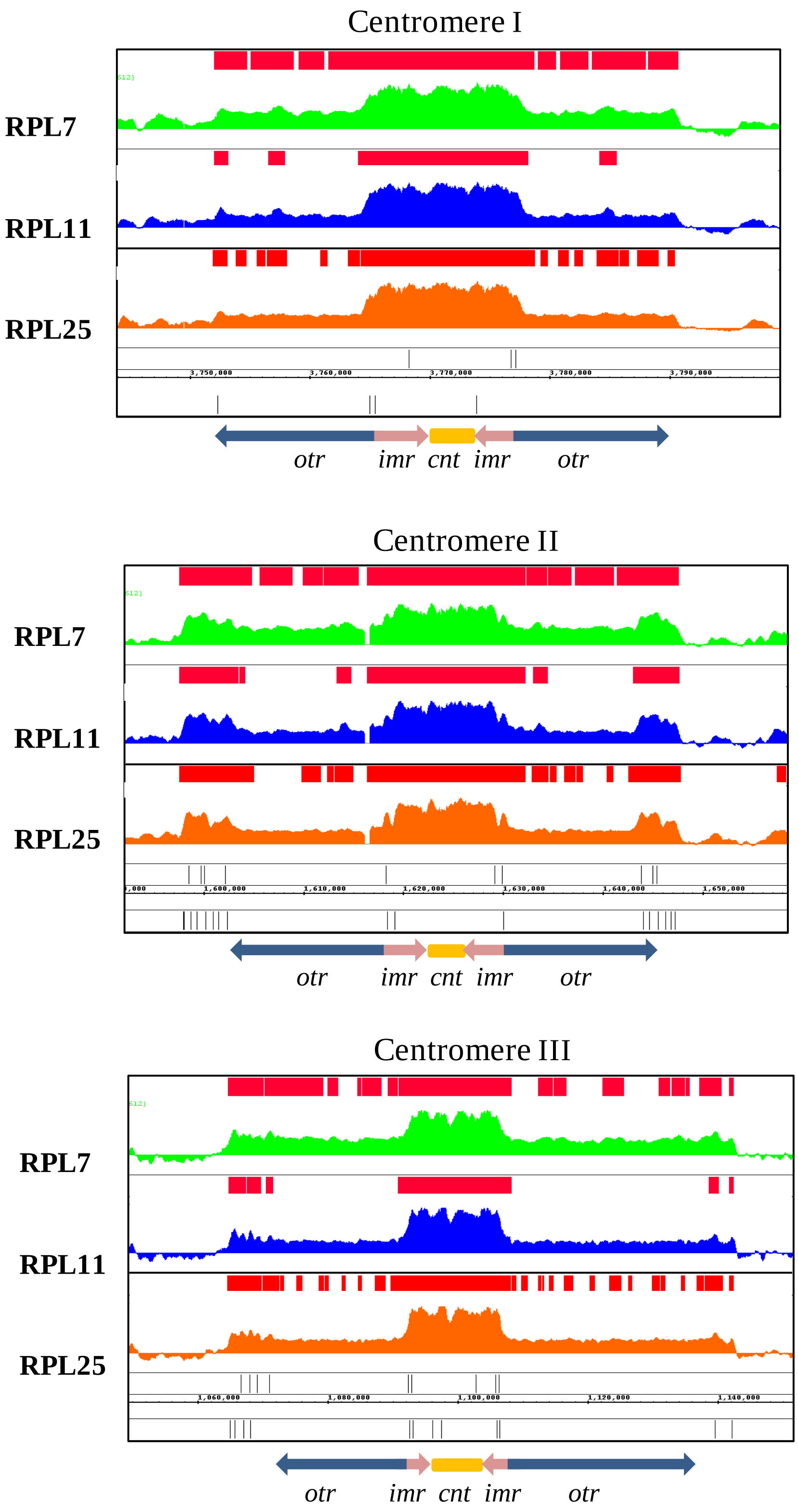


Figure 6. 4 ChIP-chip binding profile of each protein to the three *S. pombe* centromeres as visualised in IGB.

An up-to-date reanalysed version of the figure presented in De et al [50] The map below each panel shows a schematic of fission yeast centromeres, with the three major domains labelled *otr*, *imr* and *cnt*. Centromeric tRNA gene loci are indicated by black lines at the bottom of each panel. Clusters of tRNA genes can be seen flanking the *otr* regions. Regions identified by MAT as significantly bound are shown in red above the corresponding RP.

6.4.3 Functional analysis of RP binding profiles reveals links with tRNAs, energy metabolism pathways and membrane related genes

Using the annotated genes bound by the RPs (therefore excluding the regions classified as unknown), we used DAVID to determine the functional enrichment for each RP. The most significant functions and a selection of potentially interesting yet non-significant functions are shown in Table 6.2. (The raw DAVID files for the genes bound by each RP are available on the supplementary CD, in folder 'Chapter 6'). Functions with an FDR $\leq$0.05 are highlighted in red, functions with a FDR $\leq$0.1 are highlighted in green. Interestingly we identified three highly significant functional annotations that are common all three RPs. The first is triplet codon amino acid adaptor activity (average FDR: 5.82x10^{-58}) consistent with a significant association to tRNA genes (Figure 6.2A). Second, we find multiple genes encoding membrane related proteins (transmembrane, intrinsic to membrane, and cell surface) (average: FDR 10^{-4}). Finally, our data suggests that RPs bind genes involved in glycolysis (average FDR: 2.18x10^{-2}). The association of RPs to glucose metabolic processes supported the linkage with energy metabolism observed in Chapters 2, 3, 4 and 5. We also identify numerous less significant functions which RPs associate to (represented in black). Overlap analysis between the genes bound by the three RPs, to determine if all RPs significantly bind a common selection of genes. (Figure 6.5, Table 6.3). The results are reported in the next few sections

Protein	Features	Known / unknown	Detected in DAVID	Functional annotation
RPL7	355	251 / 104	168	triplet codon-amino acid adaptor activity (70), 5 transmembrane protein Schizosaccharomyces pombe (7), cell surface (12), intrinsic to membrane (30), Glycolysis / Gluconeogenesis (6), gpi-anchor (4), oxidoreductase (4), cation homeostasis (3), cytoplasm (24), specific RNAPII transcription factor activity (3)
RPL11	381	319 / 62	230	triplet codon-amino acid adaptor activity (73), cell surface (17), 5 transmembrane protein - Schizosaccharomyces pombe (6), Glycolysis / Gluconeogenesis (6), intrinsic to membrane (38), external side of plasma membrane (7), hexose metabolic process (7), oxidoreductase / NAD (11), gpi-anchor (3), cell tip (10), Zinc finger- C2H2-like (6), metal-binding (20), Transcription RNAPIII promoter (3), rrna processing (3), ubl conjugation pathway (3), regulation of mitotic cell cycle (4)
RPL25	547	458 / 89	314	triplet codon-amino acid adaptor activity (86), signal (31), intrinsic to membrane (66), anchored to membrane (8), fungal-type cell wall (9), integral to plasma membrane (8), Glycolysis / Gluconeogenesis (7), phosphate transmembrane transporter (3), regulation of glucan biosynthesis (3), Thioredoxin-like (3), magnesium (7), hexose metabolism (7,fatty acid biosynthesis (5), cell tip (13), RNA recognition motif - RNP-1 (6), Glycoside hydrolase catalytic core (3), Transcription RNAPII promoter (3), amino acid glycosylation (3), metal-binding (24), cytosolic ribosome (8), heterocycle biosynthesis (3), Protease (4), Ubiquitin mediated proteolysis (3), purine nucleotide biosynthesis (3), regulation of cell cycle (10), kinase (8), protein amino acid phosphorylation (5), meiosis (10), dna repair (4), nucleosome organization (3), mitochondrial envelope (5),

Table 6. 3 Functional analysis of genes associated to RPL7, RPL11 and RPL25
Red text indicates an FDR of ≤ 0.05, green text indicates an FDR of ≤ 0.1, and black text represents non-significant enrichment.

Overlap	Features	DAVID	Functional annotation
RPL7, RPL11, RPL25	178	134	translational elongation (65), signal (11), cell surface (10), Glycolysis / Gluconeogenesis (6), intrinsic to membrane (21), fungal-type cell wall (3), plasma membrane (3), magnesium (5), oxidation reduction (3), cation homeostasis (3), RNAPII transcription factor activity (3), cytoplasm (19)
RPL7, RPL25	34	17	intrinsic to membrane(8), base pairing with mRNA (3)
RPL11, RPL25	89	59	plasma membrane (13), , site of polarized growth (6), cell wall (5), oxidation reduction /FAD (3), triplet codon molecular adaptor activity (5), endoplasmic reticulum (10), regulation of mitotic cell cycle / cell division (4), carboxylic acid biosynthetic process (4), proteolysis / ubl conjugation pathway (3), iron ion binding (3), purine nucleotide binding (8), organelle lumen (3)
RPL7, RPL11	2	2	Isomerase (2)
RPL7 specific	37	15	Methyltransferase type 12 (3)
RPL11 specific	50	35	base pairing with mRNA (3), nucleolus (4), endoplasmic reticulum (5), metal-binding (3)
RPL25 specific	157	104	molecular adaptor activity (13), oxidoreductase (10), intrinsic to membrane (25), RNA recognition motif-RNP-1 (4), vitamin biosynthetic process (3), anchored to membrane (3), establishment or maintenance of actin cytoskeleton polarity (3), fatty acid biosynthetic process (3), anion transport (3), cytosolic ribosome (7), positive regulation of transcription from RNAPII promoter (3), cytoplasmic vesicle (3), gtp-binding (3), meiosis (7), iron (7), endoplasmic reticulum (11), cytoskeleton organization (6), kinase / phosphorylation (5), negative regulation of nitrogen compound metabolic process (3), macromolecular complex assembly (4), metal-binding (7), proteolysis (4), mitochondrial envelope (3), cellular protein localization (5), nucleolus (3)

Table 6. 4 Functional analysis of RP overlap of annotated genes.
Red text indicates a FDR of $\leq$ 0.05, green text indicates an FDR $\leq$ 0.1, and black text represents non-significant enrichment.

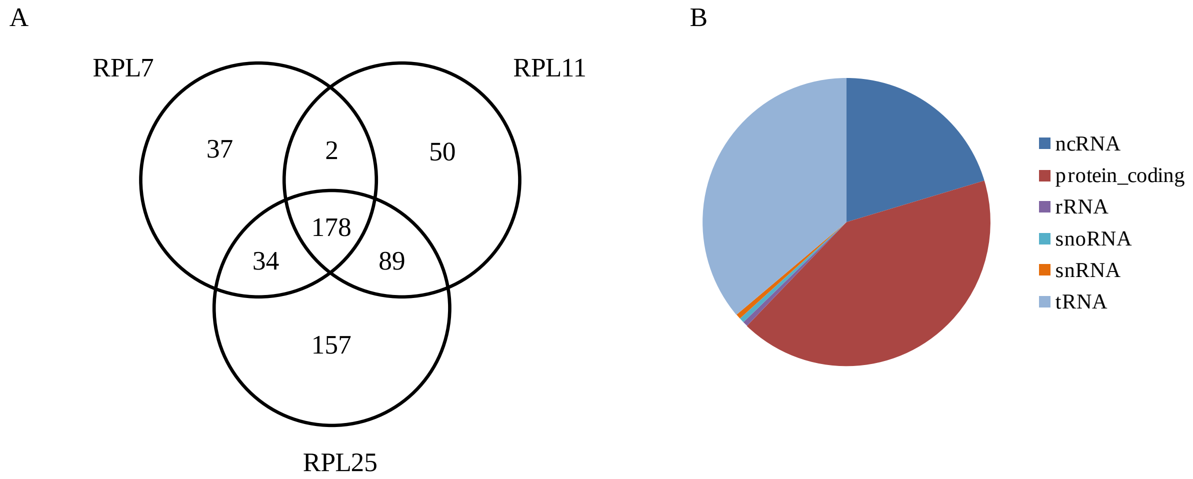
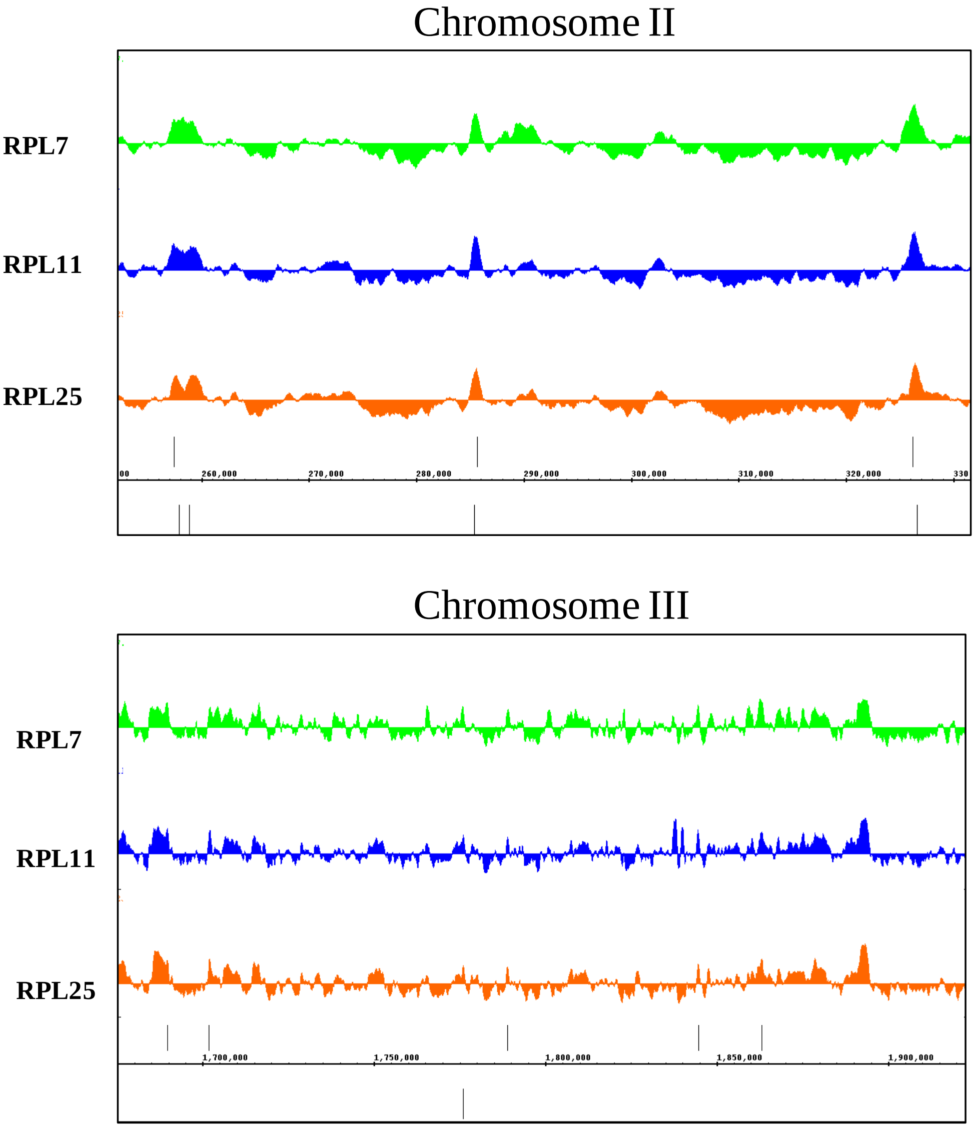


Figure 6. 5 The overlap between RPL7, RPL11 and RPL25
Panel A. Overlap of genes bound by RPs. Panel B. Piechart of the 178 overlapping genes

6.4.3.1 RPs are associated to tRNA genes

We determined that RPs associated to tRNAs, this is particularly apparent at the dense tRNA clusters within the centromeres (Figure 6.4). Clusters of tRNA genes are known to be present in *S. pomb*e centromeres [57]. This is clearly apparent within chromosome II, where there are two clear peaks indicating increased enrichment with a cluster of tRNA genes (Figure 6.4 centromere II). Bioinformatic analysis revealed a total of 27 tRNAs genes located within the centromeric regions of *S. pombe*, our overlap analysis identified that all 27 centromeric tRNAs are bound by the three RPs. In fact, of the 178 overlapping genomic features, 65 are tRNAs (Table 6.3) indicating that RPs have a stronger association to genes encoding tRNAs (Figure 6.5B). The 65 shared tRNA genes had an average fold enrichment of 9.74 with all three RPs. The remaining 38 enriched tRNAs, were dispersed in different chromosomal regions and exhibit highly specific enrichment, in which RP binding encapsulates the entire tRNA gene only, without spreading into neighbouring genomic regions (Figure 6.6). There are a total of 171 annotated tRNA genes in the tiling array, we classified them all regardless of whether they were significantly enriched or not (Figure 6.7 shows a newer and more accurate revision of the figure published in De *et al.* 2011 [50]). The association between RPs and tRNA genes were verified by Sandip De. Data not shown, but please refer to De *et al.* 2011 [50].


Chromosome II

Chromosome III

Figure 6. 6 Example of RPs association at noncentromeric tRNA genes from chromosomes II and III.

A more up-to-date and revised figure of what was reported in De *et al* 2011 [50]. tRNA genes are represented as the black vertical lines at the bottom of each plot. The upper and lower set of tRNA genes indicate the upper and lower DNA strands respectively. We show that RP association to tRNA genes are highly specific, with peaks localised only to regions where tRNA genes are located.

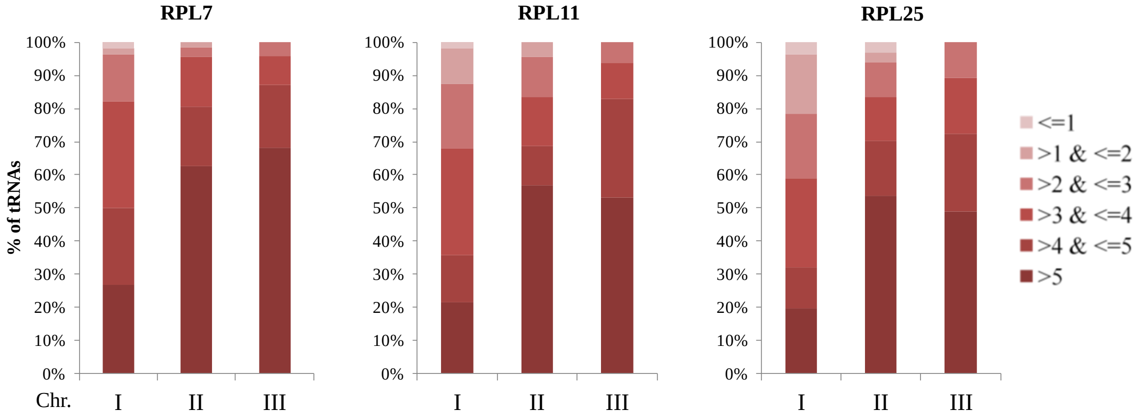


Figure 6. 7 Stacked bar charts representing the association of the RPs with all known 171 tRNA genes.

tRNA genes were classified into six classes based on their enrichment score (see colour legend). The heights of the cars represent the total percentages of the tRNAs encoded by each chromosome

6.4.4 RPs associate to genomic loci encoding proteins involved in the glycolysis and gluconeogenesis pathways

Studies in previous chapters revealed a strong linkage between genes encoding RPs and energy metabolism enzymes. This section addresses the issue of whether RPs significantly associate with glycolysis genes. To do this, every cytoplasmic energy metabolism gene that was significantly bound by RPL7, RPL11 and RPL25 were identified. A total of eight genes were met the criteria (Table 6.4) Visualisation of these genes in IGB revealed that RP binding covered the entirety of the gene (Figure 6.8), and in all cases, RPL25 had the most coverage and highest enrichment scores (Table 6.4).

Gene	Systematic name	Chr	Product	RPL7	RPL11	RPL25
FBA1	*SPBC19C2.07*	II	fructose-bisphosphate aldolase Fba1	6.98	7.93	9.70
GPD3	*SPBC354.12*	II	Glyceraldehydes-3-phosphate dehydrogenase	8.04	8.44	9.82
GPM1	*SPAC26F1.06*	I	glycerate phosphomutase	7.22	7.27	7.30
PGK1	*SPBC14F5.04c*	II	phosphoglycerate kinase Pgk1 (predicted)	5.83	6.34	8.53
SPAC1F8.07c	*SPAC1F8.07c*	I	pyruvate decarboxylase (predicted)	6.45	5.95	7.94
PYK1	*SPAC4H3.10c*	I	pyruvate kinase (predicted)	5.97	6.47	8.20
TDH1 / *GPD1*	*SPBC32F12.11*	III	glyceraldehyde-3-phosphate dehydrogenase	8.11	7.97	9.54
FBP1	*SPBC1198.14c*	II	fructose-1,6-bisphosphatase / sedoheptulose-1, 7-bisphosphatase	7.53	5.99	6.18

Table 6. 5 Information on the cytoplasmic energy metabolism genes bound by RPs, with their corresponding enrichment scores

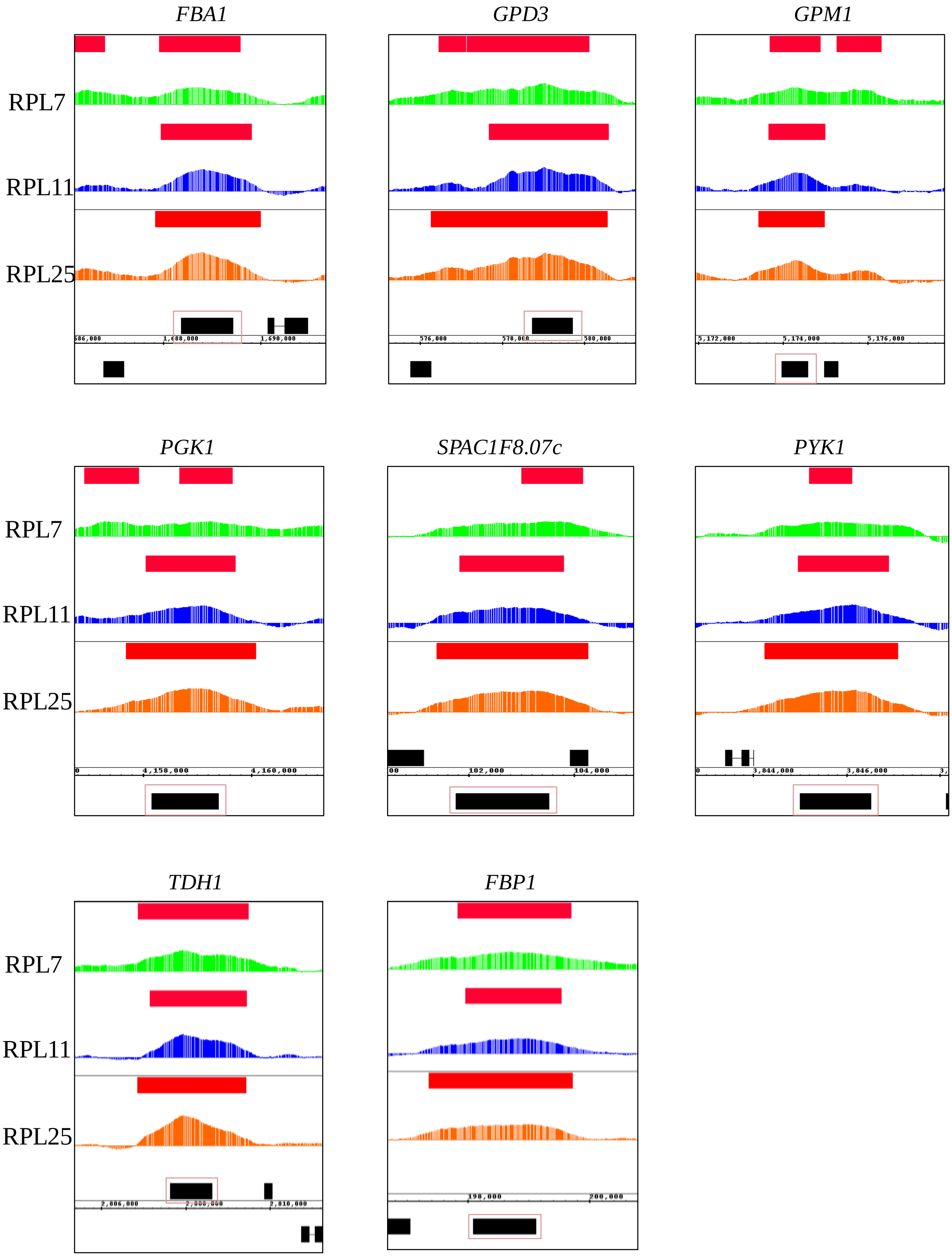


Figure 6. 8 Binding profiles for the seven glycolysis genes and gluconeogenesis gene *FBP1* bound by RPs.

Green represents RPL7, blue represents RPL11, orange represents RPL25. Red boxes above the binding profile represent the region identified to be significantly bound by the RP using MAT software. The black boxes at the bottom of each panel indicate genes as annotated in the current genome file for IGB. Genes encapsulated in the red box highlight the gene in question. The results show that in some cases, RP binding is highly specific, binding only to the gene, such as *FBP1*, *GPM1*, *TDH1* & *PYK1*. In other cases, RP binding is quite widespread, encapsulating numerous genes, including glycolysis genes (*SPAC1F8.07c*)

As described in the introduction chapter, riboneogenesis is a metabolic pathway that joins glycolysis to the non-oxidative branch of the PPP leading ultimately to the production of ribose-5-phosphate. Ribose-5-phosphate is an essential precursor of nucleotides, and is therefore essential for rRNA transcription. rRNA molecules are essential for ribosome biogenesis as they make up the core of the ribosome [212].

The discovery of riboneogenesis proved that there was a close relationship between flux through the glycolytic pathway and the rate of ribosome biogenesis [92]. The results of RP ChIP-chip binding suggested that RPs may have a role in regulating the expression of glycolysis genes. The ChIP-chip data shows that the *FBP1* gene is bound by the three RPs (Figure 6.8). This association is significant as FBP1 in *S. pombe* has sedoheptulose-1, 7-bisphosphatase activity [58] and this enzymatic activity is reported to be essential in driving riboneogenesis in *S. cerevisiae* [92]. Furthermore the glycolytic intermediates required for riboneogenesis are fructose-6-phosphate, dihydroxyacetone-phosphate and glyceraldehyde-3-phosphate (Chapter 1, Figure 1.2).

The ChIP-chip analysis identifies that both *FBA1* and *TPI1* associate with RPs, both enzymes are responsible for the entry of glycolytic intermediates the into riboneogenesis pathway [92]. Additionally, RPL25 significantly binds two aldo-keto reductases; *SPAC750.01* and *SPAC977.14c*, a class of enzyme involved in glucose metabolism [173]. The ChIP-chip results provided supporting evidence that riboneogenesis is conserved in *S. pombe* and validates our results from Chapter 5. The direct binding of RPs to genes encoding glycolytic enzymes suggests that RPs may regulate the expression of these enzymes and thus riboneogenesis. Currently known mechanisms of RP gene regulation include RPs acting as inhibitors by binding the promoter regions, coding regions and intron exon junctions and affecting the recruitment of transcription and translation machinery [100]. Inhibition of the glycolytic pathway would be expected when ribose-5-

phosphate demand is low. The excess RPs may bind glycolytic genes as a partially assembled but functionally silent complex, effectively repressing glycolysis. However, when ribose demand is high, such as during rapid cell growth, the bound RPs may dissociate from the glycolytic genes. Genes encoding glycolytic enzymes can then be expressed, thereby increasing the availability of glycolytic intermediates and thus increasing flux through the riboneogenesis pathway.

This interpretation of the results is supported by the phases in the metacycle. During the reductive - charging phase there is a massive up-regulation in glycolysis and genes involved in carbohydrate breakdown [106] [94]. Other than to increase the concentration of NADPH and acetyl-CoA, the reductive charging phase may also increase the concentration of glycolytic intermediates. This means that upon entry to the oxidative phase, flux through riboneogenesis would be rapid. These ChIP-chip binding data verify the links between energy metabolism and ribosome biogenesis that were discovered in my *S. pombe* expression analysis.

6.5 Discussion

6.5.1 RPs are present at many genomic loci

The results presented in this chapter suggest that RPL7, RPL11 and RPL25 associated to both coding and non-coding loci, consistent with reports in *S. cerevisiae* [202]. The association of RPs to chromosomes has also been observed in Drosophila [201] [213], this suggests that binding of RPs to gene loci is generally conserved in eukaryotes. Treatment with RNase eliminated or significantly decreased ChIP signal for the three RPs, confirming that association is to genomic loci is RNA dependent [50]. Suggesting that RPs associated with RNAs at protein coding and RNA coding loci [50]. The observation that the three RPs have a similar global binding pattern and have a high level of overlap between significantly bound regions indicates that the RPs may be recruited together as part of a preassembled ribosomal subunit, with a role in nuclear translation. However, further lab studies did not find convincing evidence of translation at these sites (data not shown); therefore the issue of whether translation can occur in the nucleus at the sites identified in this analysis is an area for further research. Currently, it is understood that at any given time there are a pool of free RPs that are not assembled into ribosomal subunits and that these non-assembled RPs are free to perform additional functions within the nucleus [98], therefore a possible hypothesis as to why RPs associate with chromatin may be due to a non-ribosomal function such as regulating specific groups genes, a hypothesis that has been reported in numerous eukaryotic organisms. This may account for why they associate to the same genomic loci in our ChIP-chip experiments.

6.5.2 RPs and their association to the centromere and tRNAs

RPs show a remarkably strong association to centromeres, with association being more apparent at the clusters of tRNA genes interspersed throughout the centromeric region.

tRNA association isn't isolated solely to centromeric regions though, tRNA genes located throughout the chromosomes also associate with RPs. tRNA genes make up ~ 0.1% of the *S. pombe* genome, yet represent $\geq$ 36% of the binding sites shared by all the RPs [50]. . Visual inspection and identification of the enrichment profiles strongly indicate that the three RPs associate with the same centromeric loci and tRNA genes, with only RPL25 binding additional tRNA genes. This supports the view that RPs are recruited to chromatin together as part of a preassembled but possibly silent complex.

The data suggests that the association of RP complexes to centromeres may be required in order to fulfil a particular function, possibly the transcriptional regulation of tRNA biogenesis, which has already been proven in *S. cerevisiae* [214]. The role of RPs in transcription regulation is not limited to yeast, in mammalian cells, RPL11 has been reported to repress RNAPIII transcription [215]. Our *S. cerevisiae* fitness network also hinted at a similar interaction, as genes enriched as TFIID were found to overlap spatially with RPs when using a force directed layout (Chapter 2).

6.5.3 RPs and their association to energy metabolism genes

A key result of the ChIP-chip experiments described in this chapter was that RPs appears to directly associate to a subset of genes involved in glycolysis, PPP and riboneogenesis. Here I combine the results obtained from my *S. pombe* expression network (Chapter 5) to further evaluate the functional significance of these interactions. I assessed whether genes directly bound by RPs are significantly correlated to the cytosolic RPs, mitochondrial RPs and ribosome biogenesis genes identified in Chapter 5 (Table 6.5).

I identify two main groups. The first group is classified by genes bound in the ChIP-chip data and are first neighbours in my *S. pombe* first neighbour ribosomal networks. The second group is classified by genes that are bound by RPs, but are not first neighbours in our ribosome networks. The first group includes five energy metabolism genes which are bound by RPL7, RPL11 and RPL25, and also first neighbours to cytosolic RPs in the *S. pombe* expression network. These are *FBA1*, *TPI1*, *PGK1*, *GPM1* and *ENO101* (Table 6.5). The hypothesis is that the expression of these genes is co-regulated by their direct physical interaction with RPs. The glycolytic intermediate glyceraldehyde-3-phosphate (GAP) [92] is a product of the reactions catalysed by both FBA1 and TPI1, the data demonstrate these genes are bound by and transcriptionally correlated to the expression of cytosolic RPs (Table 6.5). The flux through riboneogenesis is dependent on numerous factors including cell cycle stage, growth rate, redox stress and nutrient availability [92]. If cells are undergoing rapid growth, the demand for ribose is high; therefore the production of glycolytic intermediates such as fructose-1, 6-phosphate (F6P) and GAP would also have to increase to maintain flux through riboneogenesis. Furthermore *FBA1* and enolase enzymes were found to be first neighbours of ribosomal proteins within our *S. cerevisiae* expression network, suggesting that the enzymatic roles of key genes in riboneogenesis remain conserved across species.

The second group include *FBP1*, *GPD3*, *PYK1* and *TDH1*. These genes are bound by RPs yet do not appear as first neighbours in the *S. pombe* ribosomal networks. This observation suggests that despite having a direct physical interaction with RPs, they do not show any significant correlation to cytosolic RPs at the transcriptional level. The reason why we do not observe direct edges between *PYK1*, *GPD3* and *TDH1* may be due to the highly stringent threshold used to construct network.

These results are consistent with the three reported routes of ribose production utilising glucose [92]. The first is through the oxidative PPP, in which we identify the glucose 6-phosphate dehydrogenase enzyme, *SPAC3C7.13c*, as transcriptionally correlated to cytosolic RPs and ribosome biogenesis genes. The second is via the non-oxidative PPP (in reverse) utilising aldolase enzymes, which we identify as being significantly correlated to cytosolic RPs and ribosome biogenesis genes (*SPAC24H6.10c* and *SPAP8A3.07c*) or via riboneogenesis (*FBP1*) which is bound by RPL7, RPL11 and RPL25. Ribose production through the non-oxidative PPP and riboneogenesis requires transketolase enzymes to convert the glycolytic intermediates F6P and GAP to xylulose-5-phosphate and erythrose-4-phosphate, which are then converted to ribose-5-phosphate. Enzymes which catalyse these reactions are all identified as either being bound by RPs or are co-expressed with cytosolic RPs and / or ribosome biogenesis genes.

6.5.4 Do ribosomal proteins control their own expression by binding to *FBP1* mRNA?

The ChIP-chip data shows that *FBP1* is bound by RPL7, RPL11 and RPL25. RPs are known to regulate gene expression through a variety of means, including inhibiting splicing by binding the intron – exon junctions of premRNA [100] and inhibiting translation by binding the 5'UTR of mature mRNAs [98]. Therefore, the association of RPs to specific genomic loci may suggest that they have a role in regulating those genes. Based on this evidence, one hypothesis is that RPs regulate their own synthesis by binding to the mRNAs of genes involved in glycolysis and riboneogenesis. For example, when the demand for ribose is high, the pool of free ribosomes is low. Therefore *FBP1* and the seven glycolysis genes are expressed, providing the glycolytic intermediates for riboneogenesis. Conversely, when ribose demand is slow, the pool of free RPs is high therefore they associate to the mRNA of genes involved in riboneogenesis as part of a preassembled silent complex, effectively shutting down the riboneogenesis pathway.

In *S. cerevisiae, SHB17* activity was shown to increase during times when ribose demand was high [92], therefore, when ribose demand is low, the cell needs a way of maintaining homeostasis and returning the concentrations of sedoheptulose bisphosphatase to normal, and this may be done by excess RPs binding *SHB17* (in *S. cerevisiae*) or *FBP1* (in *S. pombe*) directly and inhibiting expression. It is important to keep in mind however that many of the genes that RPs associate to, including tRNA and glycolysis genes, are also highly transcribed. Highly transcribed genes have more RNA polymerase II (RNAPII) molecules located at the transcription site [210], and therefore the DNA at these sites would be more accessible with plenty of nascent RNA being synthesised [216] [217]. Therefore the challenge is to determine if RP association to chromatin is a consequence of non-specific interactions between RPs and nucleic acids, or whether RP association serves a specific purpose, i.e. whether they act as a feedback mechanism to control the expression of *FBP1* and glycolysis genes. In order to validate this hypothesis, further investigation is required.

Systematic name	Official name	Pathway	Product	Network first neighbours			ChIP-chip			
				cyto	mito	bio	rpl7	rpl11	rpl25	overlap
SPAC24H6.04	*HXK1*	Glycolysis	Hexokinase	N	N	N	N	N	N	N
SPAC4F8.07c	*HXK2*	Glycolysis	Hexokinase	N	N	N	N	N	Y	N
SPBC1604.05	*PGI1*	Glycolysis	Phosphoglucose isomerase	N	N	N	N	N	N	N
SPBC16H5.02	*PFK1*	Glycolysis	phosphofructokinase	N	N	N	N	N	N	N
SPBC19C2.07	*FBA1*	Glycolysis	fructose-bisphosphate aldolase	Y	N	N	Y	Y	Y	Y
SPCC24B10.21	*TPI1*	Glycolysis	triosephosphate isomerase	Y	N	N	N	Y	Y	N
SPBC32F12.11	*TDH1*	Glycolysis	glyceraldehyde phosphate dehydrogenase	N	N	N	Y	Y	Y	Y
SPBC354.12	*GPD3*	Glycolysis	glyceraldehyde phosphate dehydrogenase	N	N	N	Y	Y	Y	Y
SPBC14F5.04c	*PGK1*	Glycolysis	phosphoglycerate kinase (transferase)	Y	N	N	Y	Y	Y	Y
SPAC26F1.06	*GPM1*	Glycolysis	phosphoglycerate mutase (mutase)	Y	N	N	Y	Y	Y	Y
SPCC1620.13	*SPCC1620.13*	Glycolysis	phosphoglycerate mutase (mutase)	N	N	N	N	N	N	N
SPAC1687.21	*SPAC1687.21*	Glycolysis	phosphoglycerate mutase (mutase)	N	N	N	N	N	N	N
SPAC222.01	*SPAC222.01*	Glycolysis	phosphoglycerate mutase (mutase)	N	N	N	N	N	N	N
SPBC1815.01	*ENO101*	Glycolysis	enolase	Y	N	N	Y	Y	Y	Y
SPBPB21E7.01c	*ENO102*	Glycolysis	enolase	N	N	N	N	N	N	N
SPAC4H3.10c	*PYK1*	Glycolysis	pyruvate kinase (transferase)	N	N	N	Y	Y	Y	Y
SPAC144.12	*SPAC144.12*	Non oxidative PPP	Ribulose 5-Phosphate Isomerase	Y	N	Y	N	N	N	N
SPAC31G5.05c	*SPAC31G5.05c*	Non oxidative PPP	Ribulose 5-Phosphate 3-Epimerase	N	N	N	N	N	N	N
SPAC750.01	*SPAC750.01*	Non oxidative PPP	aldo / keto reductase	N	N	N	N	N	Y	N
SPBC2G5.05	*SPBC2G5.05*	Non oxidative PPP	transketolase	N	N	N	N	N	N	N
SPBC1709.07	*ERG27*	Non oxidative PPP	transketolase	N	N	N	N	N	N	N
SPAC977.14c	*SPAC977.14c*	Non oxidative PPP	aldo / keto reductase	N	N	N	N	N	Y	N
SPBC215.11c	*SPBC215.11c*	Non oxidative PPP	aldo / keto reductase	N	N	N	N	N	N	N

SPBC8E4.04	*SPBC8E4.04*	Non oxidative PPP	transketolase	N	N	N	N	N	N	N
SPCC1020.06c	*TAL1*	Non oxidative PPP	transaldolase	N	N	N	N	N	N	N
SPBC1198.14c	*FBP1*	Gluconeogenesis / Riboneogenesis	Closely related to SHB17 from Calvin cycle	N	N	N	Y	Y	Y	Y
SPAC732.02c	*SPAC732.02c*	Essential Riboneogenesis	Closely related to SHB17 from Calvin cycle	N	N	N	N	N	N	N
SPAC186.08c	*SPAC186.08c*	other outcomes for pyrvate	Lactate dehydrogenase	N	N	N	N	N	N	N
SPBC30D10.13c	*PDB1*	other outcomes for pyrvate	Pyruvate dehydrogenase beta	Y	N	N	N	N	N	N
SPAC26F1.03	*PDA1*	otherl outcomes for pyrvate	Pyruvate dehydrogenase alpha	N	N	N	N	N	N	N
SPAC3A12.18	*ZWF1*	Oxidative PPP	glucose 6-phosphate dehydrogenase	N	N	N	N	N	N	N
SPAC9.01	*SPAC9.01*	Oxidative PPP	glucose 6-phosphate dehydrogenase	N	N	N	N	N	N	N
SPAC3C7.13c	*SPAC3C7.13c*	Oxidative PPP	glucose 6-phosphate dehydrogenase	Y	N	Y	N	N	N	N
SPCC794.01c	*SPCC794.01c*	Oxidative PPP	glucose 6-phosphate dehydrogenase	N	N	N	N	N	N	N
SPCC16C4.10	*SPCC16C4.10*	Oxidative PPP	6-phosphogluconolactonase	N	N	N	N	N	N	N
SPBC660.16	*SPBC660.16*	Oxidative PPP	6-phosphogluconate dehydrogenase	N	N	N	N	Y	N	N
SPAP32A8.02	SPAP32A8.02		xylose and arabinose reductase	N	N	N	N	N	N	N
SPAC2F3.05c	SPAC2F3.05c		xylose and arabinose reductase	N	N	Y	N	N	N	N
SPBC28F2.05c	SPBC28F2.05c		xylose and arabinose reductase	N	N	N	N	N	N	N
SPAC24H6.10c	SPAC24H6.10c		phospho-2-dehydro-3-deoxyheptonate aldolase	Y	N	Y	N	N	N	N
SPAP8A3.07c	SPAP8A3.07c		phospho-2-dehydro-3-deoxyheptonate aldolase	Y	N	Y	N	N	N	N

Table 6. 6 Identification of genes involved in glycolysis, PPP and riboneogenesis, and whether they bound by 60S ribosomes and / or are first neighbours in our *S. pombe* ribosome expression networks

Columns 1 and 2 show the systematic and official gene name (if available), columns 3 and 4 represent the biological pathway associated to each gene and the product synthesised respectively. Column 5 indicates whether the genes were first neighbours of cytoplasmic RPs (cyto), mitochondrial RPs (mito) or ribosome biogenesis genes (bio) from the network studies conducted in Chapter 5. (Table legend continued on next page)

(Table legend continued) Column 6 indicates whether the ChIP-chip studies using RPL7, RPL11, RPL25 or all three (overlap) bound these genes. The analysis shows that *FBA1*, *TPI1*, *PGK1*, *GPM1* and *ENO101* associated to RPL7, RPL11 and RPL25, as well as first neighbours to RPs. *TDH1*, *GPD3*, *PYK1* and *FBP1* are bound by RPL7, RPL11 and RPL25 but not first neighbours of RPs in the *S. pombe* expression network.

6.6 Concluding remarks

This work reported the genome-wide binding of a representative group of 60S RPs, and in particular two key findings; the first is that that RPs associate to specific genomic loci, including to centromeres, and genes encoding tRNAs. The second is that RPs associate to genes involved in energy metabolism. The reason why RPs associate to genes involved in energy metabolism is not clear however it may be a form of autoregulation in which they bind and block the expression of genes involved in ribose-5-phosphate production when ribose demand is low. If true, the mechanism by which they do so, whether it is inhibiting transcription, blocking splicing or inhibiting translation is a question that needs to be answered in future experimental work. The role of RPs in their own autoregulation, and the regulation of genes involved in ribosome biogenesis have been reported several times, and is a conserved across eukaryotes and prokaryotes alike. However, given how recently riboneogenesis was discovered, the genome-wide RP ChIP-chip data provided a means of determining whether RPs associate to the same genes reported to be involved in riboneogenesis. This novel result is supported by evidence previously obtained in Chapter 5, as well as reported by Clasquin *et al* [92] who briefly stated that in *S. pombe* labelled glucose appeared in the form of SBP, all of which suggest riboneogenesis is conserved in *S. pombe*. The work done in Chapter 5 and complemented in this chapter suggests that FBP1 may replace SHB17 as the key regulator of riboneogenesis.

CHAPTER 7: AN INVESTIGATION INTO THE NON – CANONICAL NUCLEAR FUNCTIONS OF NMD PROTEIN UPF1 IN *Schizosaccharomyces pombe*

7.1 Abstract

Nonsense-mediated mRNA decay (NMD) is a mechanism that stimulates destruction of mRNAs containing a premature termination codon (PTC). Up-frameshift 1 (UPF1) has a fundamental role in NMD, and is conserved in all eukaryotes. UPF1 is known to localise to the cytoplasm, but there is evidence that it can accumulate in the nucleus upon the blockage of protein nuclear export. Though the canonical function of UPF1 is associated to NMD, studies in human cells have suggested that UPF1 has additional nuclear roles such as involvement in telomere maintenance, cell cycle progression and DNA replication. In this chapter, we analyse the genome-wide association of UPF1, and in conjunction with the bioinformatic analysis on various multi-level datasets, we elucidate the possible nuclear roles of UPF1. This study presents evidence in support of UPF1 nuclear functions, demonstrating that UPF1 may facilitate replication fork progression through natural replication barriers, including repetitive sequence and tRNA genes. Finally, using the developed *S. pombe* expression network (Chapter 5), I identify that *UPF1* is co-expressed with genes encoding DNA polymerase, substantiating reports that UPF1 is likely to have a role in DNA replication.

My published review on currently known UPF1 nuclear functions, on which my introduction below is based upon, is bound to the end the thesis (with permission from Biochemical Society Transactions, Ref: PPL-EX-2014-00035).

7.2 Introduction

UPF1 is a DNA / RNA helicase [218], with an essential role in NMD [219]. mRNAs containing a PTC can potentially express toxic truncated proteins. To prevent this, NMD surveys and rapidly destroys any PTC containing mRNAs [220] [221]. In yeast, the association between UPF1, UPF2 and UPF3 make up the core machinery required for NMD [222]. These three UPF proteins are also conserved across all eukaryotes [223], however in higher organisms such as *C. elegans*, humans and *D. melanogaster*, additional proteins are required for efficient NMD to occur. These include suppressor with morphogenetic effect on genitalia (SMG) proteins [221]. In humans, failure to detect and degrade mRNAs containing a PTC has been linked to recessively inherited diseases [224].

NMD in yeast occurs during translation, therefore many studies have reported that UPF1 localisation is predominantly cytoplasmic [225] [226] [227] [228]. Despite the claims that UPF1 is strictly involved in NMD, in recent years there have been a rise in publications suggesting that UPF1 may have nuclear functions. These include DNA replication, cell cycle [229] [230], telomere maintenance [231] and the potential to associate with other nuclear proteins [222].

This chapter demonstrates that UPF1 directly binds the chromosomal loci of a variety of nuclear genes, centromeric regions and highly transcribed genomic loci such as tRNA genes. This work establishes that association to the chromatin is cell cycle specific, and that UPF2 shows very limited chromatin binding. I show that UPF1 associates to all 13 transposable elements (TEs) in the *S. pombe* genome, and knockout of *UPF1* leads to a differential upregulation of eight TEs, suggesting UPF1 may act as a regulator of these elements. In further support of UPF1s involvement in DNA replication, we demonstrate that UPF1 may act as a marker for replication fork stress due to its high degree of overlap to phosphorylated histone H2A (γH2A), another protein known to bind regions prone to DNA replication stress [232]. Finally using the *S. pombe* expression network developed in Chapter 5, I demonstrate that *UPF1* is co-expressed with genes encoding DNA replication machinery, including DNA polymerase δ (delta) and DNA polymerase ε (epsilon).

7.3 Methods

7.3.1 Experimental analysis

Fission yeast transformation, imaging, RNA analysis, synchronisation of fission yeast cells, ChIP and ChIP-chip experiments were performed by Sandip De, a research student in Saverio Brogna's lab. Fission yeast transformation was done using the same methodology stated in Chapter 6 (section 6.3.1, also refer to [50]), imaging was performed using the Eclipse Ti Nikon Microscope, RNA was extracted using the hot phenol method [233], ChIP was performed according to Abruzzi *et al* [234]. Normalisation and calculation of ChIP enriched regions was done as detailed by De *et al* [50]. Probe labelling, hybridization and scanning of the *S. pombe* Tiling 1.0FR Affymetrix Arrays were performed by Dr. John Arrand at the Affymetrix facility in the School of Cancer Sciences, University of Birmingham. ChIP-chip samples included UPF1 association to chromatin in S-phase synchronized culture, G2-phase synchronized culture, and asynchronous culture, using a two replicate design. UPF2 association to chromatin was done using an asynchronous culture, using a single replicate design.

7.3.2 Processing and visualisation of ChIP-chip data

We used the Model-based Analysis of Tiling Arrays (MAT) software to analyse the Affymetrix hybridization data [52]. ChIP input DNA was used as control and was compared against the UPF1 and UPF2 samples. A p-value of 10^{-4} was used; remaining MAT parameters remained as default. Results of MAT were visualised in Affymetrix's Integrated Genome Browser (IGB) [206].

7.3.3 Identification of enriched regions and calculation of enrichment scores

Identifying genomic regions significantly bound by UPF1 or UPF2 was done using the same pipeline as detailed in Chapter 6 (section 6.3.3). Briefly, genomic regions were defined as enriched if 50% or more of the region was significantly bound by the UPF protein in question. Enrichment scores were assigned to genomic features using the *S. pombe* genome coordinates and calculating an average enrichment between the start and end coordinates of enriched genomic regions. Thereby giving each enriched region a score based on fold enrichment. Identification of significantly bound genomic features and enrichment score calculation was done using the statistical computing language R [11] using the same scripts developed in Chapter 6. Functional annotation of the enriched regions was done using DAVID [21].

7.3.4 Identifying differentially expressed genes in a *UPF1* knockout

UPF1 mutant data was obtained from the study conducted by Rodríguez-Gabriel *et al* [235], in which they disrupted *UPF1* expression by substituting the ORF with the kanMX6 cassette as detailed in Steever *et al* [236]. Identification of differentially expressed genes was done using significance analysis of microarrays (SAM) at time point 0 between wildtype (WT) and *UPF1* mutant using a 1% FDR.

7.3.5 Integration of γH2A ChIP-chip analysis

To further investigate the role of UPF1 in DNA replication we used a study published by Rozenzhak *et al* published in 2010 [232] in which they reported the genome-wide ChIP-chip binding of phophorylated histone protein γH2A. γH2A is sensitive method for identifying regions of DNA replication stress, due to its role in stabilizing stalled replication forks [232]. The γH2A MAT files were obtained by emailing the corresponding author of the paper [232]. Their MAT analysis was done using a p-value of 10^{-5}; therefore their analysis was slightly more stringent than ours. To ensure consistency, identification of significantly enriched genomic features was identified using the pipeline and R scripts reported in section 7.3.3. The γH2A ChIP-chip analysis was conducted during S-phase, which meant a direct comparison could be made against our UPF1 S-phase ChIP-chip data.

7.4 Results

7.4.1 Outline of Genome-wide association of UPF1 with transcribed regions

Our ChIP-chip binding data indicates that association of UPF1 to the chromatin is cell cycle specific (Table 7.1). UPF1 associated with chromosomal loci throughout the three *S. pombe* chromosomes, regardless of cell cycle stage; however association was higher during the S-phase than G2-phase (Figure 7.1). Over 350 additional genomic features were bound by UPF1 during the S-phase compared to G2-phase (Table 7.1). Despite having a similar binding profile, asynchronous UPF1 associated to ~100 genomic features fewer than UPF1 S-phase (Table 7.1).

The genome-wide binding of UPF1 encompassed a diverse array of genomic features, including the mating-type (MT) locus, rDNA loci, tRNA loci, and all other heterochromatin regions, including the centromeres and telomeres (Figure 7.1). UPF1 association was also observed at mobile genetic elements (Tf2-type retrotransposons and wtf elements). The highest level of UPF1 enrichment was detected in the centromeres of *S. pombe* chromosomes. Conversely, UPF2 showed a much lower association to chromatin (Table 7.1), particularly at protein-coding genes (Figure 7.2D). Overlap analysis between the UPF1 samples and UPF2 identified only 22 genomic features in common (Figure S7.1), of which 15 were tRNAs, three were ncRNAs and three were coding (Table S7.1).

Yeast culture	**Number of statistically validated genomic features**	**Of which classified as 'unknown'**
UPF1 S-phase	520	100
UPF1 G2-phase	161	54
UPF1 Asynchronous	416	71
UPF2 Asynchronous	90	34

Table 7. 1 Breakdown of genomic features bound by UPFs

Genomic features were classified into two groups, based on their annotations. Those that were annotated were called ‘known’ genomic features. Genomic features that were unannotated and had no gene name were called “unknown” genomic regions. These unknown regions are typically small (<50bp) noncoding regions that are not yet well characterised. Typically they encompass long terminal repeats (LTRs), promoter regions and other repeat regions. The classification of known genomic features bound by the UPF1 and UPF2 are shown in Figure 7.2. Functional analysis on the known genomic features from each ChIP-chip experiment are shown in Table 7.2. The unknown genomic features for each ChIP-chip experiment were also classified (Figure 7.3), however, due to their lack of a gene name, they could not be used for functional analysis.

The data showed that almost 50% of the unknown regions bound by UPF1 during the S-phase (7.3A) and in asynchronous culture (7.3C) were repeat regions, also a substantial number were classified as LTRs suggesting that UPF1 is required at repeat sequences, a feature of UPF1 that has been reported in numerous publications [222] [230] [237] [238]. Unlike in asynchronous and in S-phase cultures, UPF1 showed no preferential binding to repeat regions or any other class of unknown regions (Figure 7.3B).

Surprisingly, UPF1 association in asynchronous culture bound 416 genomic features, an unexpectedly high number. The G2-phase is the longest of all cell cycle phases (~90 minutes) in *S. pombe* [239], therefore it’s expected that the majority of cells would be in the G2-phase, consequently the number of significantly bound genomic regions would be expected to be lower.

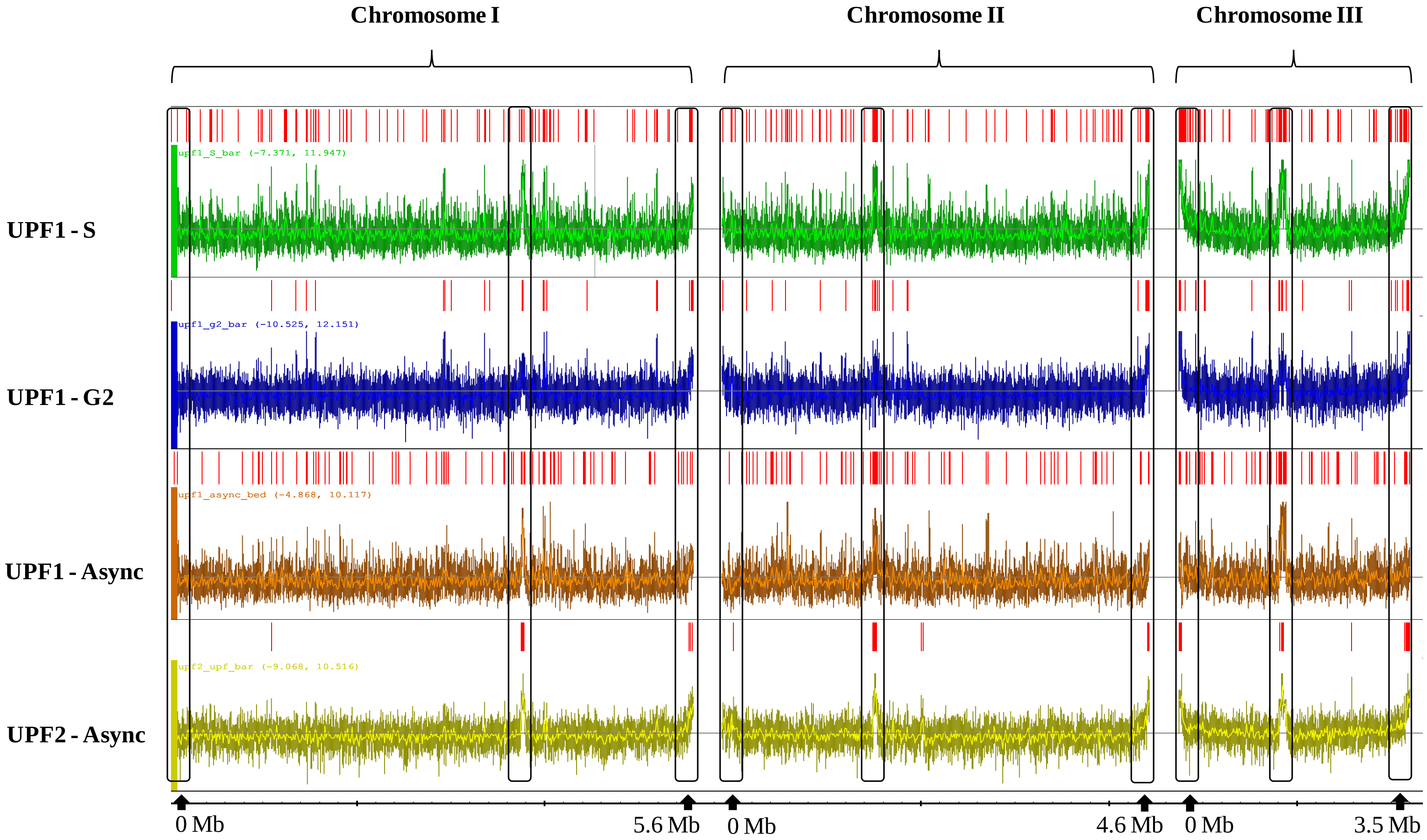


Figure 7. 1 Binding profiles of UPF1 and UPF2
In S-phase cells, (green), G2 (blue), asynchronous cells (orange) and UPF2 (yellow, asynchronous) across all three *S. pombe* chromosomes. Red bars indicate regions of the genome identified as being significantly bound using MAT. Chromosomes I – III are labelled at the top of the Figure 7.1. Vertical boxes highlight the telomeres and centromeres within each chromosome

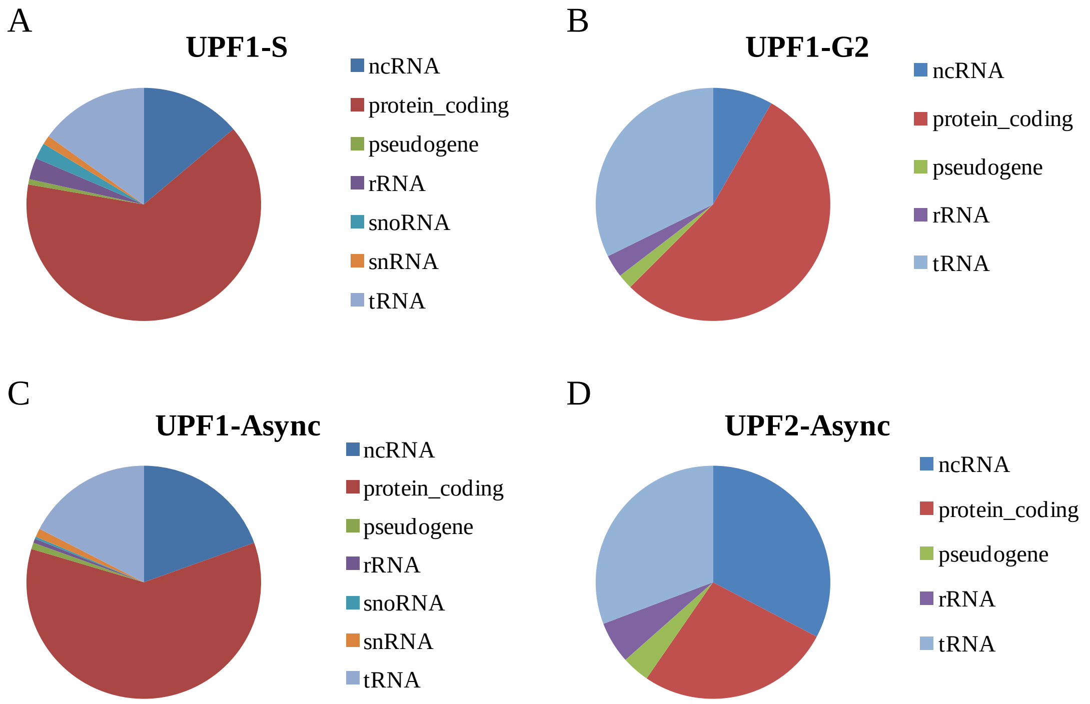


Figure 7. 2 Pie Charts showing the genomic regions associated to UPF1.
UPF1-S phase (A), UPF1-G2 phase (B), UPF1-Asynchronous culture (C) and UPF2-Asynchronous culture (D)

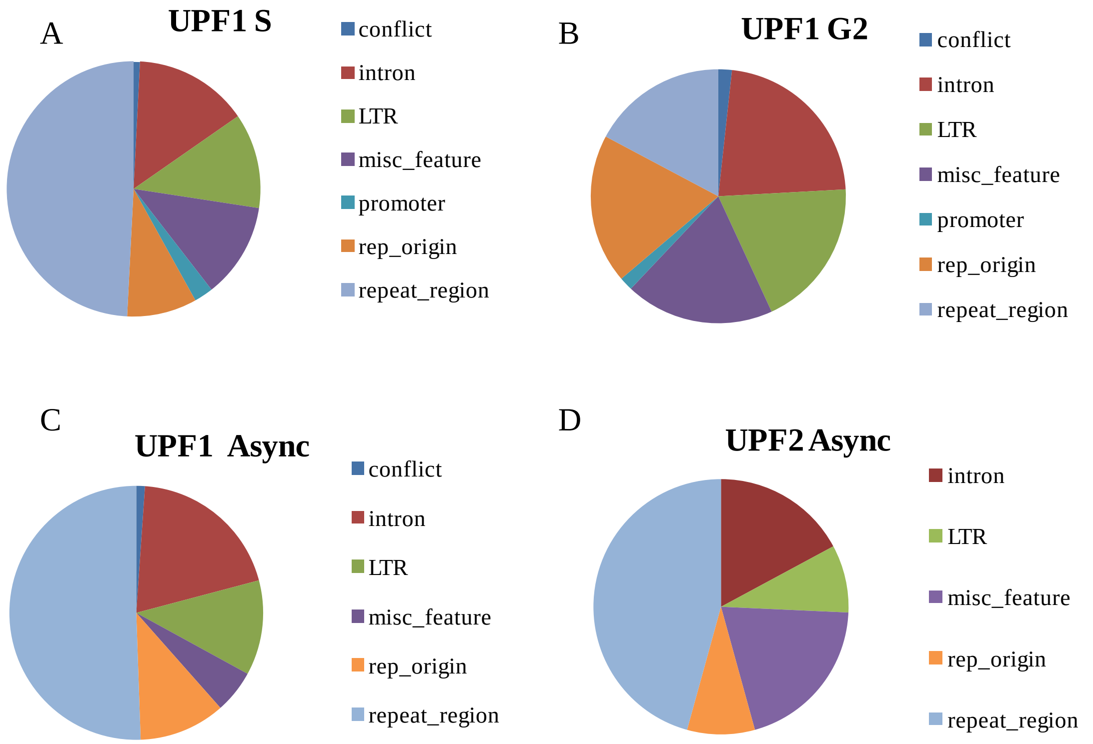


Figure 7. 3 Pie Charts showing the 'unknown' regions associated to UPF1.
UPF-S phase (A), UPF1-G2 phase (B), UPF1-Asynchronous culture (C) and UPF2-Asynchronous culture (D)

Protein	Features	Known / unknown	Detected in DAVID	Functional annotation
Upf1 S	520	420/ 100	295	triplet codon-amino acid adaptor activity (65), cell surface (28), transposable element (11), external encapsulating structure (12), 5 transmembrane protein Schizosaccharomyces pombe (7), plasma membrane (36), C4-dicarboxylate transporter/malic acid transport protein (4), cell wall, integral to plasma membrane (10), Glycolysis / Gluconeogenesis (7), NAD / oxidoreductase (16), Phosphate permease (3), Nitrogen metabolism (4), iron (9), potassium (3), polyamine transport (3), external side of cell wall (3), Glycoside hydrolase, subgroup catalytic core (4), heme (3), Cysteine and methionine metabolism (5) ,elongation factor (3), Fatty acid biosynthesis (3), lytic vacuole (7), cellular response to nutrient (3), nucleosome core (3), cytosolic ribosome (12), fungal-type cell wall biogenesis (4), ion transport (6), ATP (3), rRNA processing (5), nucleolus(14), ligase (7), GTPase activity (3), transcription (3), ligase activity (7), nucleotide biosynthetic process (3), mRNA catabolic process (3), dna repair (4), cytokinesis (4), mitochondrial matrix (4),
Upf1 G2	161	107/ 54	76	triplet codon-amino acid adaptor activity (31), Integrase - catalytic core / DNA integration (11), cell surface (13), external side of plasma membrane (5), signal (5), protein modification by small protein conjugation (3)
Upf1 Async	416	345 / 71	239	base pairing with mRNA (65), cell surface (15), gpi-anchor (5), Glycolysis / Gluconeogenesis (6), external side of plasma membrane (6), RNA recognition motif, RNP-1 (5), magnesium (5), Thioredoxin fold (3), plasma membrane (5), cell cortex part (9), Zinc finger C2H2-like (5), fatty acid biosynthesis(4), Alanine, aspartate and glutamate metabolism (3), cytosolic ribosome (10), metal-binding (25), cellular homeostasis (7), intrinsic to membrane (34), cellular response to nutrient (3), Zinc finger RING-type (3), gtp-binding (3), fungal-type vacuole (3), regulation of mitotic cell cycle (6), nucleotide biosynthetic process (3), protein modification by small protein conjugation (5), mRNA catabolic process (3), cation transport (3), macromolecular complex subunit organization (7), kinase (5), cell division (6), ribosome biogenesis (5), nucleolus (7), nuclear lumen (11), transit peptide mitochondrion (3), DNA repair (3), vesicle-mediated transport (4)
Upf2 - Async	90	56 / 34	33	translational elongation (20), endoplasmic reticulum (4)

Table 7. 2 Functional analysis of the genomic regions bound by UPF1 at different cell cycle stages and by UPF2 asynchronous culture
Text colour represents adjusted FDR as reported by DAVID. Red text represents an FDR $\leq$ 0.05; green text represents an FDR $\leq$ 0.1, black text represents FDR > 0.1..

7.4.2 The binding profile of UPF1 is cell cycle dependent

Overlap analysis demonstrated that 46 genomic regions (excluding unknowns) were shared amongst the UPF1 samples (Figure 7.4). Functional analysis of the overlap identified significant enrichment (FDR $\leq$0.1) of tRNAs and cell surface protein genes (Table 7.3). 195 features are bound only by UPF1 in S-phase synchronised cultures. Functional analysis identified genes encoding signal proteins, plasma membrane, malic acid transporters, and major facilitator superfamily MFS-1 (Table 7.3). The MFS protein superfamily is one of the two largest families of membrane transporters known, with functions encapsulating solute uniport, solute / cation symport / antiport and solute / solute antiport [240]. The 50 feature overlap between S-phase and G2-phase is significantly enriched in transposable elements (TEs) (Table 7.3) suggesting that association of UPF1 to TEs only occurs within the S and G2 phases. Only nine of the 107 features bound by UPF1 are G2-phase specific. Noteworthy are two genomic genes that are enriched in both G2 and asynchronous cells, both encode lysine tRNA. S-phase and asynchronous cultures share 129 regions that are significantly enriched in tRNAs and glycolysis genes (Table 7.3). Of the 168 regions that are specific to asynchronous cells there is no significant enrichment of specific functional groups, however there is also a minor overrepresentation of cell division, cellular homeostatis and transcription regulation genes.

The take home message is that there is a common functional overlap between genomic regions that UPF1 associates to, regardless of cell cycle stage. Notably, association of UPF1 to genes encoding TEs is a feature of only S-phase and G2 phase cultures.

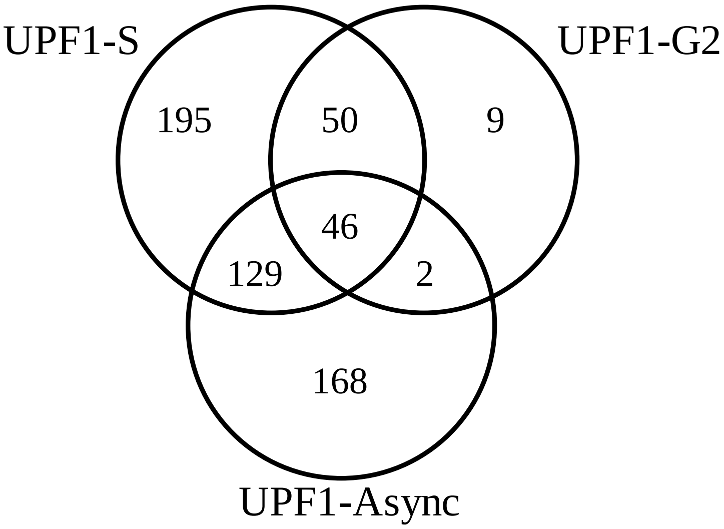


Figure 7. 4 Overlap of significantly associated genomic regions for UPF1 ChIP-chip samples.

Yeast UPF1 strain	Detected in DAVID	Functional Annotation
S-G2-Async	46	base pairing with mRNA (28), cell surface (4), protein modification by small protein conjugation (3)
S-Async	129	base pairing with mRNA (25), Glycolysis / Gluconeogenesis (6), plasma membrane (12), RNA recognition motif RNP-1 (3), magnesium (3), cellular response to oxidative stress (3), cytosolic ribosome (5), cell cortex (3), metal-binding (9), atp-binding (7), RNA polymerase II transcription factor activity (3), chromatin (3), transcription from RNAPII promoter (3)
Async-G2	2	tRNA lysine binding (2)
S-G2	50	DNA integration / Transposable element (11), cell surface (9)
S only	195	Signal (17), C4-dicarboxylate transporter/malic acid transport protein (4), translational elongation (13), iron (8), major facilitator superfamily MFS-1 (8), organic acid biosynthetic process (11), plasma membrane (17), Secreted (7), glutamine metabolic process (3), oxidoreductase (10), metal-binding (19), cellular polysaccharide metabolic process (4), NAD / NADH binding (8), response to nutrient (3), regulation of conjugation (4), cytosolic ribosome (7), ribosome biogenesis (7), ligase (5), dna-binding (5), generation of precursor metabolites and energy (4), macromolecular complex assembly (4), nuclear envelope (3), regulation of transcription from RNA polymerase II promoter (3), meiosis (5), chromatin (4)
G2 only	9	non-coding RNAs (2), carboxylases(1), lactate dehydrogenase (1), wtf protein (1), transporter chaperone (1), ankyrin repeat protein (1)
Async only	168	base pairing with mRNA (10), oxidation reduction (11), Zinc finger (3), signal (9), cellular response to nutrient (3), cell cortex (6), cell division site (6), fatty acid biosynthetic process (3),m RNA metabolic process (6), cytosolic ribosome (5), RNA polymerase II transcription factor (3), cytoskeleton organization (6), cell cycle process (11),

Table 7. 3 Functional analysis of overlap between UPF1 samples.
Red text represents FDR $\leq$ 0.05, green text represents FDR $\leq$ 0.1, black text representes FDR > 0.1

7.4.3 Cell-cycle-dependent association of UPF1 with the centromere

The genome-wide binding profile of UPF1, suggested strong association to centromeres and telomeres. Fission yeast centromeres consist of a unique central core (*cen*), flanked by inner (*imr*) and outer (*otr*) repeat regions [241]. The analysis of the ChIP-chip data revealed that UPF1 associated with the *cen*, *imr* and *otr* domains in S-phase (Figure 7.5) with highest enrichment at the *cen* and *imr* domains. During G2-phase, UPF1 association with the *otr* domains was reduced.

7.4.4 UPF1 binds tRNA genes in both S-phase and G2-phase

We observed UPF1 association to tRNA genes. UPF1 binds at 66 tRNA genes in S-phase, and 33 in G2-phase, of which 30 are shared between the two cell cycle stages (Figure 7.6A). Many tRNA genes are located in the centromeric regions in *S. pombe* and form clusters that are roughly localised at the borders between heterochromatin domains (indicated by the vertical black lines in Figure 7.5), which are thought to prevent heterochromatin spreading [242] [243]. There are a total of 27 tRNA genes located within centromeric regions, of these, 26 associate with UPF1 in S-phase, 18 in G2-phase and 17 are shared (Figure 7.6B). Ultimately, UPF1 binds all 27 tRNA genes located within the centromeric regions during S and G2-phases of the cell cycle, however when considering G2-phase alone or S-phase alone UPF1 does not associate to all 27, but rather a subset. When considering S-phase alone however, UPF1 does bind 26 of the 27 centromeric tRNAs (Figure 7.6B) therefore it is possible that UPF1 associates with the 27th tRNA, albeit at a level below the significance threshold (p-value 10^{-4}) used for this study. What is clear from this study is that UPF1 has a high affinity for genes that encode tRNAs.

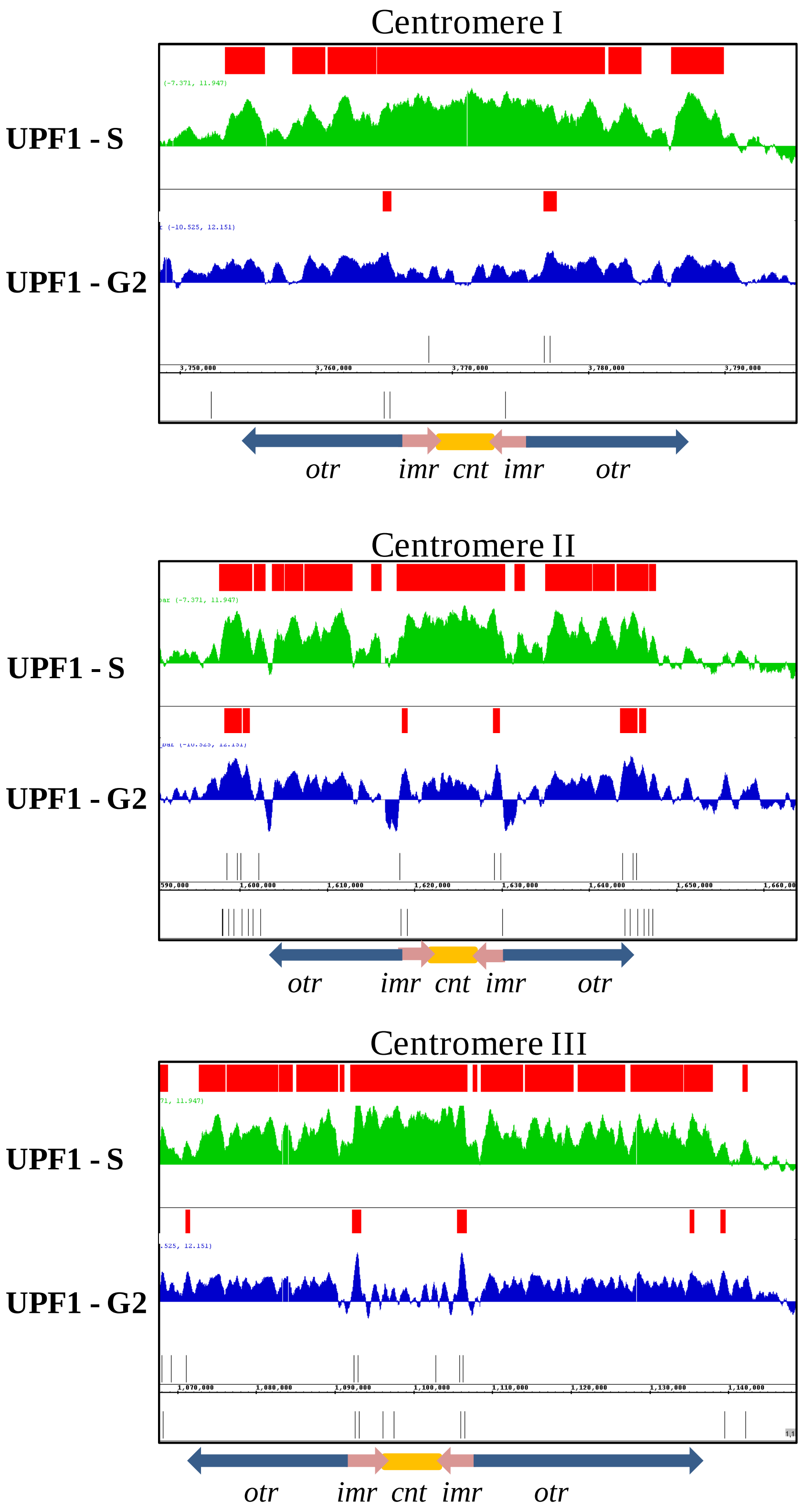


Figure 7. 5 UPF1 association to the centromere is cell cycle dependent.
tRNA genes are significantly bound by UPF1 regardless of cell cycle stage. Regions identified as significantly bound by UPF1 are shown in red tRNA gene loci are indicated by vertical black lines. During S-phase UPF1 associates to the entire centromere, however during G2 phase, UPF1 only associates to tRNA gene loci.

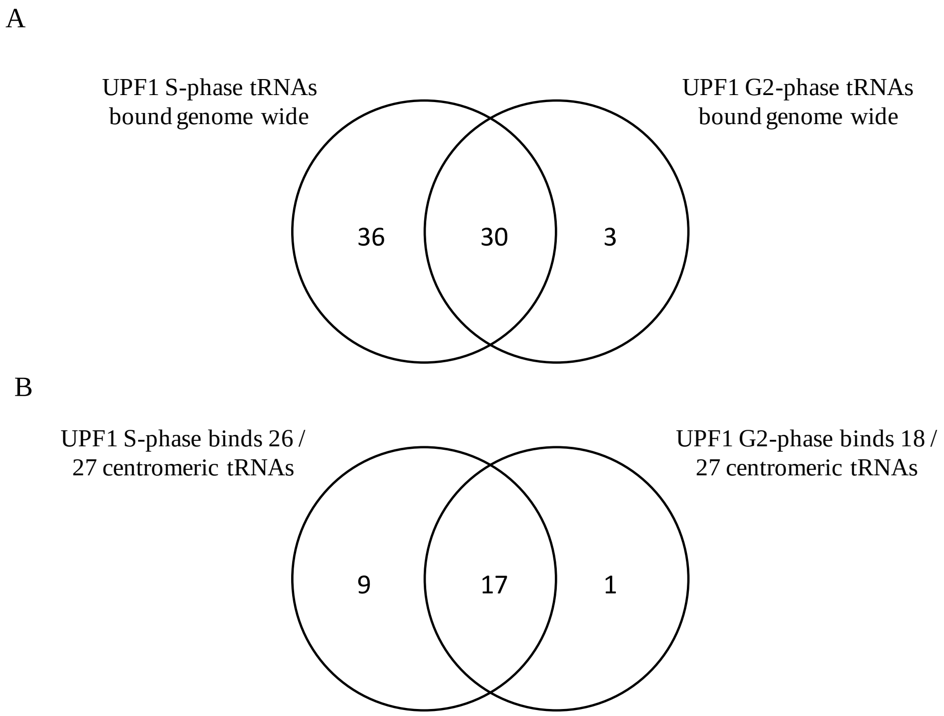

A
UPF1 S-phase tRNAs bound genome wide
UPF1 G2-phase tRNAs bound genome wide
36
30
3
B
UPF1 S-phase binds 26 / 27 centromeric tRNAs
UPF1 G2-phase binds 18 / 27 centromeric tRNAs
9
17
1

Figure 7. 6 The binding of UPF1 S-phase and G2-phase to tRNAs at the genome-wide level and centromere level.

Panel A. UPF1 binding to tRNA loci genome wide. We show that UPF1 S-phase associates to 66 tRNA genes, and UPF G2-phase associates 33 tRNA genes. 30 tRNA genes are bound during both stages of the cell cycle. Panel B. UPF1 binding to the 27 tRNA loci to centromeric regions only. We identify that UPF1 S-phase binds 26, and UPF G-phase binds 18, of which 17 are shared.

7.4.5 UPF1 association with telomeres

UPF1 binds at the telomeres (Figure 7.1). The ChIP-chip data showed a substantial enrichment of these regions in UPF1 S-phase and to a lower extent in asynchronous cell cultures (Figure 7.1). The sub-telomeric regions of chromosome III are also highly enriched; these sub-telomeric regions contain tandem arrays of rDNAs that are subject to heterochromatic silencing [244]. Telomeric regions are also bound by UPF2 (Figure 7.1).

7.4.6 UPF1 and its strong association to heterochromatic regions

It is clear that UPF1 S-phase has a strong association to heterochromatin and repeat regions of the chromosome. We observed that UPF1 S-phase significantly bound the entirety of the centromere (Figure 7.5) and the telomeres (Figure 7.1). We also determined that over 50% of unknown regions bound by UPF1 S-phase are either repeat regions or LTRs (Figure 7.3). Evidence so far suggests that UPF1 association to repetitive regions of the chromosome is not simply random, and in fact may serve an integral purpose.

7.4.7 UPF1 binds at poorly replicated chromosomal regions during S-phase

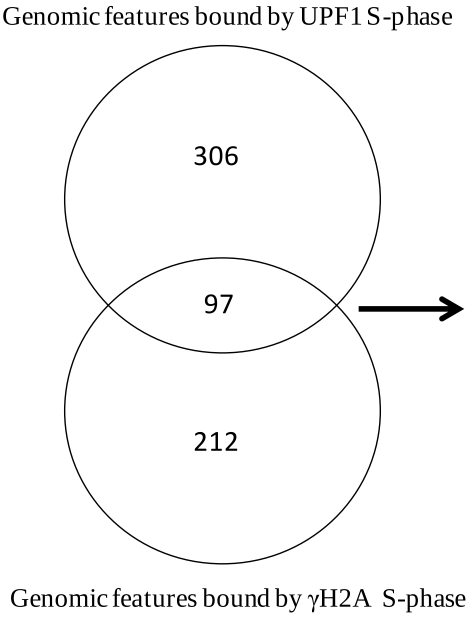
Current experimental evidence indicates that γH2A preferentially binds natural replication fork barriers, retrotransposons, heterochromatin (in both centromeres and telomeres) and rRNA repeats [232]. Therefore, to determine if UPF1 had a role in DNA replication, we did a comparative study between the genome-wide binding of UPF1 and γH2A. The γH2A ChIP-chip data was recorded during the S-phase, allowing us to make direct comparisons to our UPF1 S-phase ChIP-chip data. The γH2A data was obtained by email from the corresponding author of Rozenzhak *et al's* publication [232]. In order to maintain consistency we applied the same methodology for identifying enriched regions that we had done with the UPF1 ChIP-chip samples.

We identified 447 genomic regions significantly bound by γH2A, 309 were known genomic features and 138 were classified as unknown. A breakdown of the genomic features bound by γH2A is shown in Figure 7.8B, with functional analysis of the known regions shown in Table 7.4. γH2A binds cell surface, plasma membrane, transport and TEs with an FDR $\leq$0.05. A notable finding of this analysis is that like UPF1 S-phase, γH2A also binds TEs (Table 7.4). Less significant enrichment includes cell cycle processes such as cell division, cytoskeleton, and M-phase of cell cycle. These functionally overlap with the genomic regions bound by UPF1 S-phase (Figure 7.7). γH2A binds a significantly lower proportion of tRNAs during S-phase compared to UPF1 S-phase. However the remaining proportions for genomic features remain similar (Figure 7.8B). Analysis of the unknown regions bound by γH2A, shows that both UPF1 and γH2A bind similar proportions of repeat regions. γH2A binds far more LTR regions, and a very small number of introns (44 to 4 respectively). We also observe that γH2A does not bind the *imr* and *cnt* regions of the centromere, with enrichment only occurring within the *otr* regions (Figure 7.8C).

A comparative study between the genome-wide binding of UPF1 with γH2A, revealed a 97 region overlap (Figure 7.7). Using a random sampling method, repeated 1000 times, we calculated that the p-value corresponding to a 97 gene overlap is below 2.2^{-16} (this is the smallest integer that can be displayed in R). Functional analysis on the overlap identified tRNA binding, TEs and plasma membrane as the most significant hits (FDR $\leq$ 0.05). There were eight transposable elements, out of the genome total 13 which overlapped between UPF1 S-phase and γH2A. γH2A is reported to stabilise stalled replication forks in regions of the genome that are tough to replicate [232]. UPF1 and γH2A share similar overall binding profiles with the highest enrichment at the telomeres and centromeres (excluding the *cen* and *imr* regions for γH2A, Figures 7.8A and 7.8C), and there is a high degree of overlap between UPF1 and γH2A. Implying that UPF1 binding during S-phase may have a role in alleviating DNA replication barriers. Studies have shown that in human UPF1 depleted cells, replication fork progression and termination are affected [230]. The prominent enrichment of γH2A and UPF1 at heterochomatic and transposon loci suggests that there may be a relationship between these proteins and DNA replication, a hypothesis further supported by the fact that both ChIP-chip datasets were measured during the S-phase. Like γH2A, UPF1 may work through the natural replication barriers caused by repetitive sequence and highly transcribed genes such as tRNAs and therefore may help in the maintenance, stability, and repair of replication forks.

Features	Known / unknown	Detected in DAVID	Functional annotation
447	309 / 138	213	cell surface (36), transposable element (11), plasma membrane (32), intrinsic to membrane (60), cell wall (15), base pairing with mRNA (15), fungal-type vacuole (13), amide transporter activity (3), gpi-anchor (6), monosaccharide transport (4), cofactor binding (13), Secreted (8), serine-type peptidase activity (3), lipid binding (4), Velum formation protein (3), specific RNA polymerase II transcription factor activity (4), M phase of meiotic cell cycle (8), mitosis (3), phosphorylation (3), transit peptide (4), protein transport (4), cell division (4), kinase (3), cytoskeletal (3)

Table 7. 4 Functional annotation of γH2A binding
Text colour is representative of significance (red: FDR $\leq$ 0.05; green: FDR $\leq$ 0.1; black: FDR > 0.1) .


Genomic features bound by UPF1 S-phase
306
97
212
Genomic features bound by γH2A S-phase

Functional annotation	Count	Adjusted FDR
RNA binding / tRNA binding	23	4.50E-11
DNA integration / DNA replication / DNA polymerase activity	8	1.40E-11
cell surface	15	2.40E-14
5 transmembrane protein, Schizosaccharomyces pombe	7	8.10E-11
signal	10	5.50E-04
plasma membrane part	5	7.60E-06
integral to plasma membrane	3	1.30E-01
cytoplasm	11	1.00E+00

Figure 7. 7 The overlap of genomic features that were significantly enriched by UPF1 S-phase and γH2A S-phase.
Functional analysis on the overlap shows that tRNA, cell membrane and transposable elements are the most significant enriched functions shared between the two proteins. Text is displayed as red if the adjusted FDR $\leq$ 0.05. The 97 gene overlap was calculated as having a p-value of less than 2.2^{-16} indicating that the expectation of getting the overlap by random chance is miniscule.

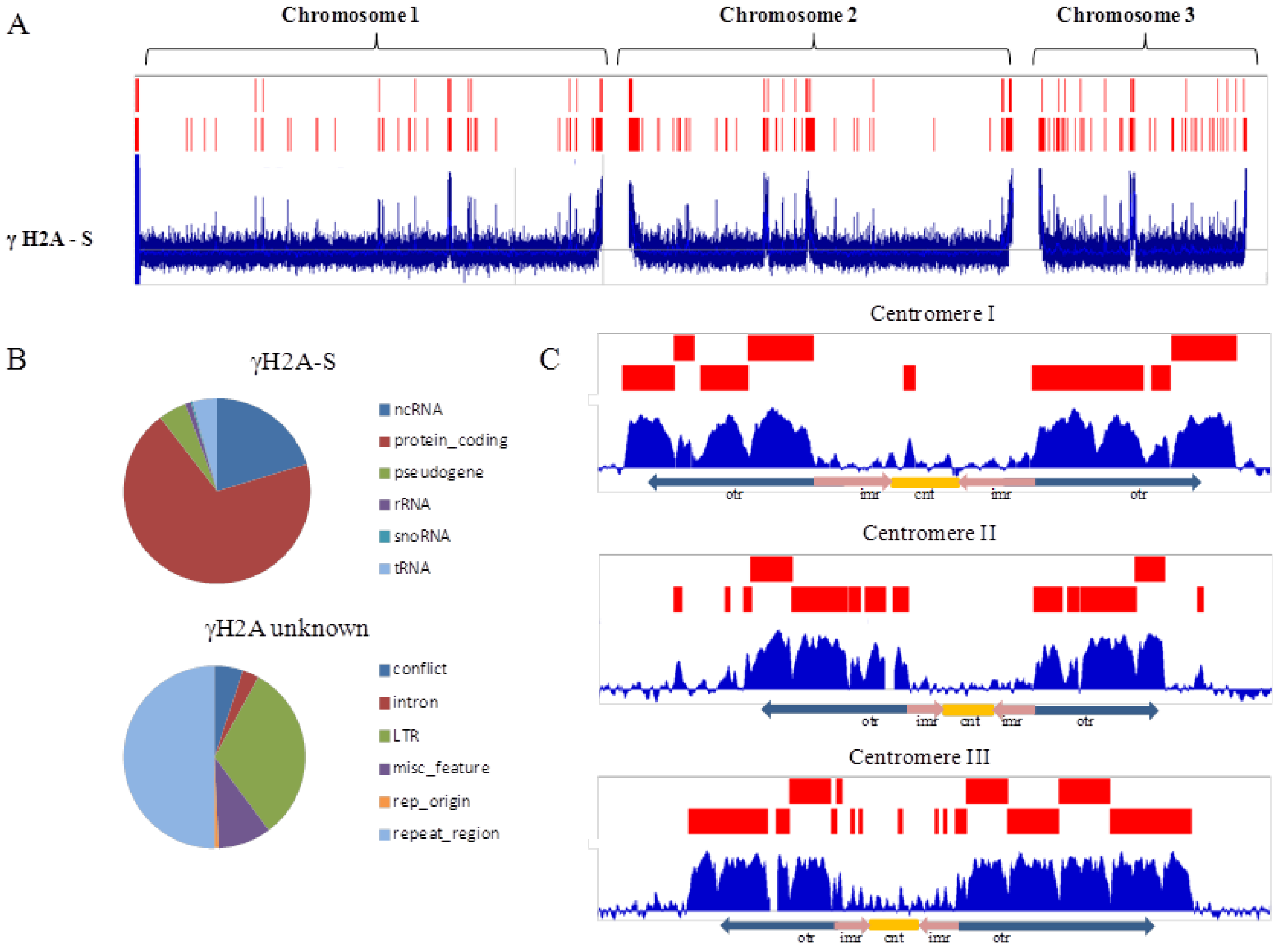


Figure 7. 8 The binding information of γH2A

Panel A Significantly high enrichment at the telomeres and centromeric regions of each chromosome. PanelB The breakdown of known and unknown genomic features bound by γH2A. Panel C γH2A does not bind the *imr* and *cnt* regions of the centromere, with enrichment only occurring within the *otr* regions

7.4.8 UPF1 during the S-phase binds to and possibly regulates TEs

The results have suggested that UPF1 S-phase binds highly repetitive sequence; this was substantiated during our comparison analysis with γH2A.We decided to further investigate this phenomenon. There are a total of 13 TEs within the *S. pombe* genome [178], of which UPF1 associates to all 13. Furthermore, the enrichment of the 13 TEs lay within the top 25% of all regions significantly bound by UPF1 S-phase, with an average score of 7.4 (Table 7.5), indicating a seven fold increase in enrichment in comparison to the control samples (ChIP input DNA was used as a control). Association of UPF1 S-phase to TEs are highly specific with binding strictly encapsulating the entirety of the transposon gene only (Figure 7.9). We also observed that the majority of TEs had a ‘three peaks’ binding pattern, in which regions of the transposon gene, specifically the terminal ends and the centre had higher enrichment than the rest of the gene (Figure 7.9). The significance of the ‘three peaks’ binding pattern is currently unknown. It is well known that TEs contain several tandem and triplicate nucleotide repeats which may lead to replicative slippage, making them tough to replicate accurately [245]. Therefore a possible hypothesis as to why we observed UPF1 binding to TEs is that UPF1 may be required to overcome the natural DNA replication fork barriers presented by repetitive DNA.

TE	Chromosome	Enrichment Score
Tf-12	chromosome3	7.919464
Tf-9	chromosome2	7.671477
Tf-6	chromosome1	7.613859
Tf-4	chromosome1	7.604309
Tf-10	chromosome2	7.571048
Tf-13	chromosome3	7.558111
Tf-1	chromosome1	7.55539
Tf-5	chromosome1	7.549103
Tf-3	chromosome1	7.473116
Tf-2	chromosome1	7.35929
Tf-11	chromosome2	6.983825
Tf-8	chromosome1	6.957233
Tf-7	chromosome1	6.46788

Table 7. 5 UPF1 enrichment of TEs, in order of fold enrichment

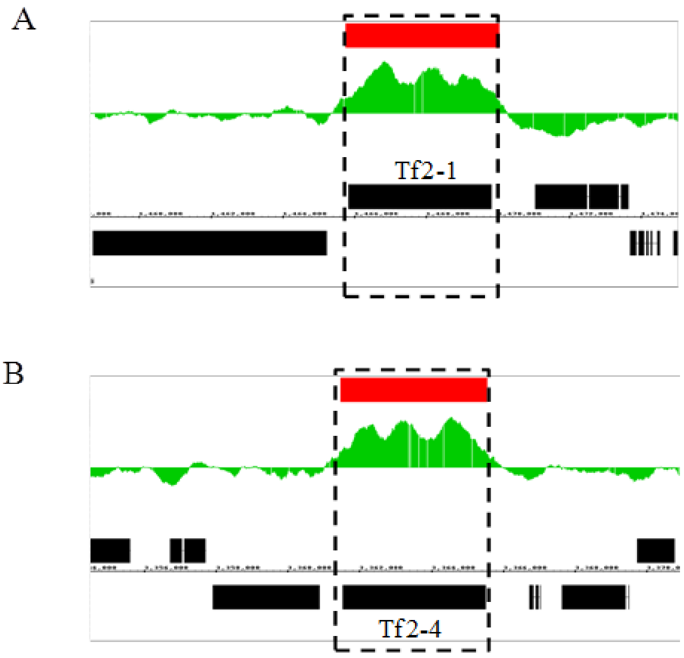
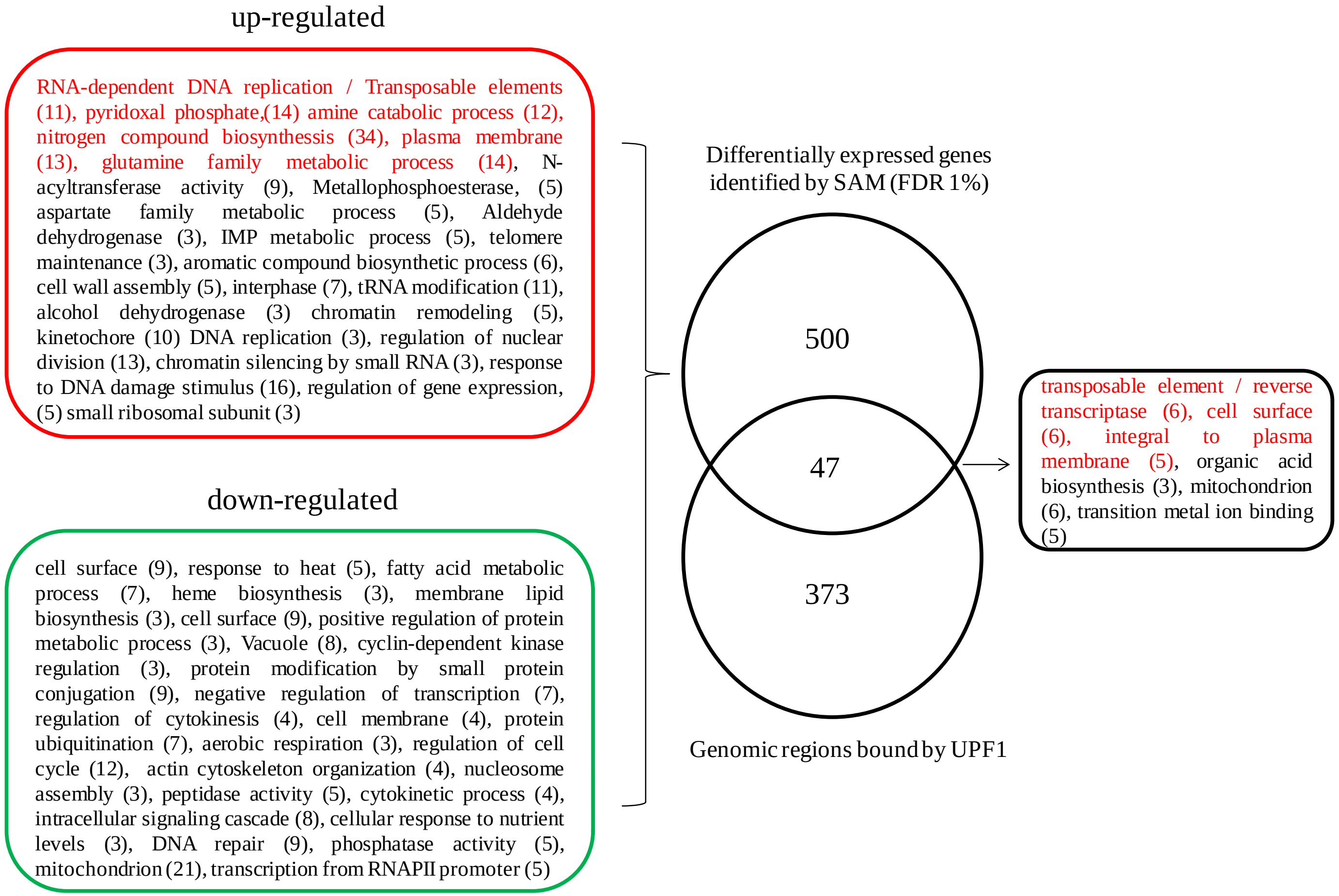


Figure 7. 9 UPF1 S-phase binding to transposable elements is highly specific
Two examples are shown, Tf2-1, and Tf2-4. This figure demonstrates the highly specific nature of UPF1 binding and the ‘three peaks’ binding pattern we observe across the TE gene.

7.4.9 UPF1 may bind to and regulate a specific subset of genes during S-phase

This analysis has highlighted that UPF1 binds to a diverse set of genomic features. The next step of investigation was to identify if the genes bound by UPF1 are also regulated by UPF1. In order to discern this, we used a *UPF1* mutant dataset published by Rodriguez – Gabriel *et al* which measured gene expression in *S. pombe* UPF mutant strains upon exposure to oxidative stress [235]. Significance analysis of microarrays (SAM) with a 1% FDR was used to identify differentially expressed genes at timepoint 0 (prior to oxidative stress) between the wildtype (WT) and *UPF1* mutant. A total of 547 genes differentially expressed genes were identified, of these, 161 were down regulated, however interestingly, more than double (386 genes) were up-regulated in response to the UPF1 knock-out. Functional analysis on the down regulated genes showed enrichment, though not significant, in cell surface, cell cycle and various lipid biosynthetic processes (Figure 7.10). Functional enrichment on the up-regulated genes however identified TEs as the most significant hit. Other significant enrichment included pyridoxal phosphate, decarboxylation, plasma membrane, and deamination reactions of amino acids [246]. Less significant hits included telomere maintenance, kinetochore, DNA damage and tRNA modification (Figure 7.10). These suggested that there was a functional overlap between genes bound by UPF1 S-phase and genes differentially expressed in response to a *UPF1* mutant. This result implied that when *UPF1* is knocked out, the expression of TEs is unregulated, leading to uncontrolled expression. This was investigated further by identifying the overlap between differentially expressed genes from the *UPF1* mutant with the regions significantly bound by UPF1 S-phase. We identified 47 regions that overlapped between the two lists (Figure 7.10). The p-value for this overlap was calculated to be 0.001 using a random sampling method looped 1000 times. This p-value indicated that the overlap was highly significant. Functional analysis of these genes showed that TEs, reverse transcriptase and cell surface / plasma membrane were the most significant hits. Interestingly, TEs, DNA replication and cell surface / plasma membrane functions are all genes which are up-regulated in response to the *UPF1* mutant. Furthermore, there is a functional overlap with genes bound by UPF1 S-phase in our ChIP-chip data. An unexpected result was that of the 13 TEs bound by UPF1, six of them are differentially up-regulated in the mutant *UPF1* yeast strain suggesting a novel idea that UPF1 influences TE expression.



up-regulated

RNA-dependent DNA replication / Transposable elements (11), pyridoxal phosphate,(14) amine catabolic process (12), nitrogen compound biosynthesis (34), plasma membrane (13), glutamine family metabolic process (14), N-acyltransferase activity (9), Metallophosphoesterase, (5) aspartate family metabolic process (5), Aldehyde dehydrogenase (3), IMP metabolic process (5), telomere maintenance (3), aromatic compound biosynthetic process (6), cell wall assembly (5), interphase (7), tRNA modification (11), alcohol dehydrogenase (3) chromatin remodeling (5), kinetochore (10) DNA replication (3), regulation of nuclear division (13), chromatin silencing by small RNA (3), response to DNA damage stimulus (16), regulation of gene expression, (5) small ribosomal subunit (3)

down-regulated

cell surface (9), response to heat (5), fatty acid metabolic process (7), heme biosynthesis (3), membrane lipid biosynthesis (3), cell surface (9), positive regulation of protein metabolic process (3), Vacuole (8), cyclin-dependent kinase regulation (3), protein modification by small protein conjugation (9), negative regulation of transcription (7), regulation of cytokinesis (4), cell membrane (4), protein ubiquitination (7), aerobic respiration (3), regulation of cell cycle (12), actin cytoskeleton organization (4), nucleosome assembly (3), peptidase activity (5), cytokinetic process (4), intracellular signaling cascade (8), cellular response to nutrient levels (3), DNA repair (9), phosphatase activity (5), mitochondrion (21), transcription from RNAPII promoter (5)

Differentially expressed genes identified by SAM (FDR 1%)
500
47
373
Genomic regions bound by UPF1

transposable element / reverse transcriptase (6), cell surface (6), integral to plasma membrane (5), organic acid biosynthesis (3), mitochondrion (6), transition metal ion binding (5)

Figure 7. 10 Overlap of differentially expressed genes in UPF mutant and regions bound by UPF

Left - Functional analysis of differentially expressed genes in the *UPF1* mutant using SAM with a 1% FDR. Middle – Venn diagram identifying the overlap between differentially expression genes in the *UPF1* mutant and genomic regions bound by UPF1. Right- Functional analysis on the 47 gene overlap. Text colour is representative of significance (red: FDR $\leq$ 0.05; green: FDR $\leq$ 0.1 ; black: FDR > 0.1). The 47 gene overlap has a p-value of 0.001.

7.5 Discussion

7.5.1 UPF1 and its' role in DNA replication

In yeast, UPF1, UPF2 and UPF3 are not essential for the viability [247], however UPF1 and UPF2 are essential in Drosophila, zebrafish and human [222]. The important issue is whether the essential requirement for UPF1 is due to these proteins having essential secondary functions unrelated to NMD or that the loss of UPF1 leads to cell mortality due to DNA damage as a consequence of lack of NMD. The analysis described in this chapter has provided evidence that UPF1 is also present at the chromosomes and suggests that that this protein has additional functions unrelated to NMD**.**

The finding that UPF1 is directly associated to replicating DNA in *S. pombe*, supports the hypothesis that UPF1 is directly involved in DNA replication, as previously reported in mammalian cells [230]. In recent years there has been an overwhelming amount of evidence linking UPF1 and UPF2 to DNA replication and cell cycle progression functions. Studies in *Drosophila* D2 cells revealed that depletion of UPF1 and UPF2 causes cell cycle arrest and the differential expression of 15 mRNA involved in cell cycle and DNA repair [229]. Azzalin *et al* have reported that UPF1 has a direct role in DNA replication in human cells, upon depletion of UPF1, cells arrested in early S-phase [230]. Cells were able to initiate DNA replication, however the lack of UPF1 lead to an arrest in replication due to the replication fork being unable to progress across the DNA [230]. The association of anti proliferating cell nuclear antigen with replication forks upon UPF1 depletion is consistent with problems with DNA replication machinery [230]. Carastro *et al* reported that UPF1 co-immunoprecipitates with the catalytic subunit of DNA polymerase δ in bovine thymus tissue [248]. The work conducted in this chapter supports existing literature. Firstly we identify that UPF1 S-phase binds heterochromatic regions of the centromere and telomere, as well as all 13 TEs which are known to contain repetitive sequence. Secondly, using available *UPF1* mutant expression data we identified that the significantly differentially expressed genes include TEs, DNA elongation / replication and telomere maintenance. Furthermore overlap analysis with the genomic regions bound by UPF1 S-phase, revealed a significant overlap with significant enrichment in TEs. Thirdly, analysis of the gene overlap with γH2A, identified hard to replicate regions such as TEs and tRNA genes as the most significantly enriched. However, how replication fork progression is co-ordinated and maintained when replicating heterochromatin regions still remains poorly understood. The helicase activity of UPF1 could have a major role in replicating centromeric DNA during the S-phase, alternatively UPF1 could function as part of a chromatin remodelling complex. Another hypothesis may be that the enrichment of UPF1 to *cen* and *imr* domains of the centromere may indicate a role of UPF1 in kinetochore formation during mitosis [249]. In order to further understand the global association of UPF1 to chromatin, additional ChIP-chip experiments will be required which take into account specific stages of the cell cycle. By doing so, the global distribution of UPF1 can be determined, which may yield further information regarding UPF1 and its involvement in DNA replication.

7.5.2 A newly constructed *S. pombe* expression network provides further evidence of UPF nuclear functions

Analysis of *UPF* knockout expression data identified that TEs, DNA replication and cell surface genes are differentially expressed in response to a *UPF1* mutant. These same functional groups and presumably same genes overlap with the genomic loci bound by UPF1 in the ChIP-chip analysis. To garner further support of UPF1 in DNA replication, the expression network developed in Chapter 5 was integrated into the analysis. For this study a slightly lower threshold of 0.2MI was used, this new MI threshold was equivalent to a p-value of $1x10^{-61}$. *UPF1* was mapped onto the network and the first neighbours identified. The resulting network contained 337 nodes and 38225 edges and was visualised using a force directed layout (Figure 7.11A, Table 7.6). A single level of modularisation using GLay [33] identified three modules (Figure 7.11B). Functional analysis of each module identified the genes *UPF1* is transcriptionally coupled to. Module 1 is significantly enriched in genes related to ribosome biogenesis and rRNA processing. It is important to note that functional enrichment, even though not statistically significant, still represent statistically significant correlations between genes. This is because the network was thresholded to leave genes which have the strongest correlations. In light of this, we observe a small number DNA replication within module 1 (Table 7.6). Specifically, these genes are DNA polymerase delta (δ) catalytic subunit, DNA polymerase epsilon (ε) subunit B and DNA replication licensing factor *MCM4*. *UPF1*s co-expression to DNA replication machinery is further substantiated in module2 in which I identify enrichment of DNA metabolic processes, which contain the genes DNA replication licensing factors *MCM3* and *MCM7*. The results suggest that *UPF1* has strong transcriptional linkage to genes encoding mini-chromosome maintenance proteins (MCM). The MCM family of proteins are reported to be essential replication initiation factors, containing six structurally related proteins, MCM2 – 7 [250], we identify that UPF1 is transcriptionally correlated to three of these. These observations support results by Azzalin *et al* which stated that UPF1 has a direct role in DNA replication in human cells by physically interacting with the catalytic subunit of polymerase δ and facilitating fork progression [230], and results by Carastro *et al* which stated that UPF1 co-immunoprecipitates with the catalytic subunit of DNA polymerase δ [248]. My expression network, served as supporting evidence showing that *UPF1* has a strong transcriptional correlation to subunits of DNA polymerase at the transcriptional level.

Together with the previous results, they suggest that the UPF1 protein physically associates to genomic loci that are tough to replicate, and aid in replication fork progression by interacting with DNA polymerase subunits and overcoming natural replication barriers caused by tRNA clusters and repetitive sequence. The results also suggest that *UPF1* is co-expressed with TEs and DNA polymerase subunits δ and ε. Overall it shows that there is both a physical and transcriptional linkage between UPF1 and DNA replication machinery and tough to replication regions.

The ChIP-chip experiments clearly showed that UPF1 association to the chromatin was cell cycle dependent with low association being observed during G2 phase, and highest during the S-phase, consistent with the involvement of UPF1 in DNA replication. These results are supported by reports that in humans, UPF1 association to chromatin is low during mitosis and early G1, increasing mid G1, before reaching highest enrichment in S-phase, before diminishing on the completion of S-phase [230]. What is interesting is that UPF1 is known interact with UPF2; however human cells depleted in UPF2 progress normally through the cell cycle even though NMD is inhibited by UPF1 depletion. Perhaps, UPF1 assembles into two specific complexes. In the first UPF1 physically interacts with UPF2 to perform NMD functions during a DNA damage response. The second involves UPF1 physically interacting with DNA polymerase to perform DNA synthesis by facilitating fork progression. Further microarray experiments with total RNA collected from cells in the presence and absence of *UPF1* might provide more detailed information on the function of UPF1 protein in transcription regulation in *S. pombe* cells.

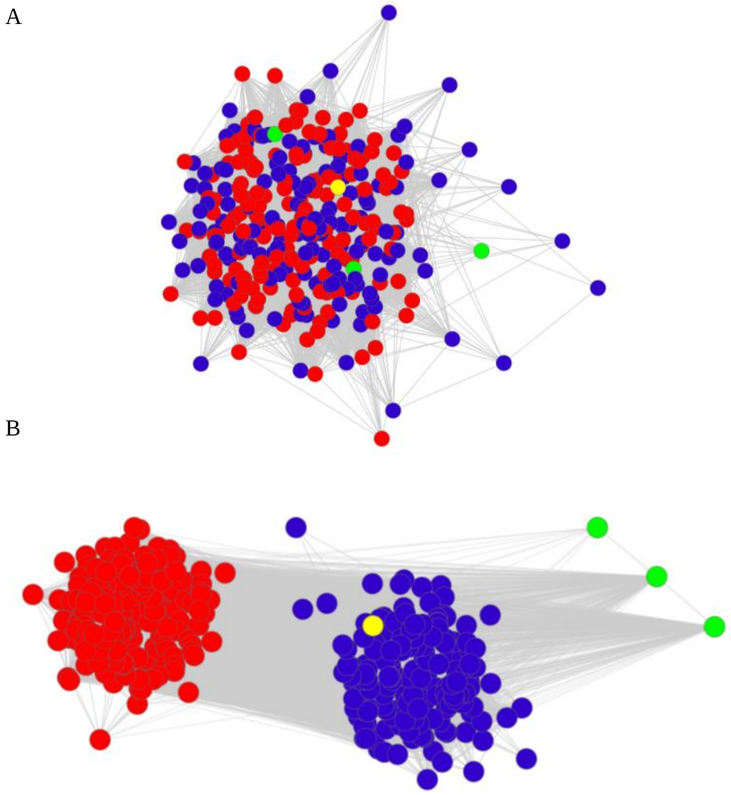


Figure 7. 11 The first neighbour network of UPF1 using a 0.2MI threshold
Panel A Force directed layout of UPF1 first neighbours. . GLay identified 4 subnetworks, 1, 2 & 3 represented as red, blue and green respectively. The yellow node represents UPF1.

Module	Nodes	Edges	Functional Enrichment
1	177	12249	ncRNA processing / ribosome biogenesis (26), RNA-dependent ATPase activity /rrna processing (15), stress response (6), RNA modification (8), snoRNA binding (7), peptidase activity (9), NAD(P)-binding domain (6), manganese ion binding (5), ubiquitin-dependent protein catabolic process (9), vacuole (8), transition metal ion binding (21), RNA recognition motif RNP-1 (4), monosaccharide catabolic process (3), Pyrimidine metabolism (3), coated vesicle (3), late endosome to vacuole transport (3), protein modification by small protein conjugation (6), glycoprotein (8), ascospore formation (4), DNA replication (3), response to DNA damage stimulus (8)
2	157	8863	protein biosynthesis (16), atp binding (27), eukaryotic 43S preinitiation complex (4), ncRNA metabolic process / ribosome biogenesis (26), ribonucleoside monophosphate metabolic process (5), ribosomal large subunit assembly (3), snoRNA binding (4), HEAT (6), 1,3-beta-glucan biosynthetic process (3), oxidoreductase / fmn (8), manganese (3), RNA-dependent ATPase activity (5), endoplasmic reticulum (27), cellular macromolecular complex disassembly (3), cell membrane (5), lipid biosynthetic process (7), RNA recognition motif RNP-1 (4), mitochondrion (16), cell membrane (5), RNA modification ((5), metal-binding (16), mRNA splicing (4), homeostasis (5), metabolic process (6), ubiquitin-dependent protein catabolic process (3), DNA metabolic process (6)
3	3	2	tRNA processing (2)

Table 7. 6 Functional analysis of UPF1 first neighbours.
Text colour is representative of significance (red: FDR ≤ 0.05; green: FDR ≤0.1 ; black: FDR > 0.1) .

7.6 Concluding remarks

The work in this chapter combines data from multiple sources to thoroughly investigate the nuclear roles of UPF1 in *S. pombe*. The incorporation of *UPF1* KO expression data, γH2A ChIP-chip data, and *UPF1* expression network data, together with UPF1 ChIP-chip data provided a wealth of information supporting a role of UPF1 in DNA replication, a result congruent with studies in mammals, specifically bovine [248] and human [230]. The discovery that UPF1 may have a role in TE regulation was novel, with the results suggesting that upon UPF1 knockout, TEs are uncontrollably expressed. The data suggested that UPF1 may silence transposons through DNA-protein interactions, consistent with mechanisms of transposon silencing specific to *S. pombe* [66]. Cross referencing UPF1 ChIP-chip binding results with γH2A binding suggested that UPF1 may bind regions where DNA replication may be hindered, particularly at repetitive sequence (transposons, centromeres, telomeres) as well as highly transcribed regions (tRNAs). The role of UPF1 in DNA replication was further supported using the *S. pombe* expression network constructed in Chapter 5, in which direct edges between UPF1 and subunits of the DNA polymerase protein, including subunit δ were identified, congruent with existing literature. In summary, this chapter provided a comprehensive study into the potential nuclear roles of UPF1 and reported evidence implicating the UPF1 protein in DNA replication in *S. pombe.*

CHAPTER 8: A NETWORK BIOLOGY APPROACH TO IDENTIFYING ADVERSE OUTCOME PATHWAYS IN METAL AND METALLOID TOXICITY

8.1 Introduction

8.1.1 Metals and metalloids are toxic

An adverse outcome pathway (AOP) can best be described as a series of events across multiple levels of biological organisation, ranging from exposure, the early molecular initiating events to the final adverse phenotypic effect [251] [252]. The advent of functional genomics technologies, particularly expression profiling in combination with advanced computational methods that allow the reconstruction of biological pathways from observational data (reverse engineering) has contributed enormously to the development of an unbiased approach to AOP inference [253].

In Chapter 2, I demonstrated the application of network inference techniques to genome-wide fitness data and shown the potential of this technique in identifying phenotypic associations, which are informative of underlying regulatory circuits. Here I utilise both the Hillenmeyer and Vulpe fitness datasets used in Chapter 2 to study mechanisms of toxicity underlying metal (zinc, cadmium and lead) and metalloid (arsenite and monomethyarsonous acid -MMA^{III}) exposure.

Arsenic is toxic to cells and is a known human carcinogen, exposure to this metalloid occurs primarily through contaminated drinking water [254]. Inorganic arsenic is known as arsenite. Humans and mammals are able to methylate arsenite to form MMA^{III}, a metabolite known to be more toxic that arsenite [254]. Arsenic proposed mechanisms include spindle disruption, formation of reactive oxygen species (ROS) and inhibition of DNA repair [254]. Cadmium is a heavy metal, due to its use within industry it is known to effect human health. Like arsenic, it has been classified as a human carcinogen [255] and is also reported to induce neurodegenerative diseases [256]. Exposure to cadmium is believed to effect cell differentiation and apoptosis [255]. Zinc ions are essential for the viability of most organisms and are known to have numerous important roles, including gene stability and expression, and in the protection of DNA [257]. High concentrations of metal ions however can lead to cell toxicity and the up-regulation of defence mechanisms including detoxification [257], ubiquitination and chaperone proteins [258]. Lead is a heavy metal ion, known to induce neurotoxicity and lead to adverse cognitive function [259].

I have shown that yeast cultures containing genes mutated in ribosome biogenesis and translation factors have increased resistance to arsenite, zinc, cadmium and lead. Interestingly, I also identified that yeast strains containing mutated mitochondrial and cytoplasmic energy metabolism genes also had increased resistance when exposed to zinc.

8.2 Methods

8.2.1 Identification of fitness modules linked to metal and metalloid exposure

The aim was to identify mutations that affect fitness in specific environmental conditions. To accomplish this task I first identified genes representing mutant strains with differential fitness between the highest and lowest dosages (cellular IC_{20} concentration vs cellular IC_{10} and IC_5 concentrations) in the zinc, lead, cadmium, arsenite and MMA^{III} exposures from Vulpe's fitness dataset (supplementary CD, folder 'chapter 8'). This was achieved by using the SAM methodology at an FDR$\leq$ 5%. Then I asked the question whether any of the sub-modules generated from the Hillenmeyer fitness network described in Chapter 2 was enriched in any of the statistically significant genes identified by SAM analysis. This was done using gene set enrichment analysis preranked (GSEAPreanked) [113]. Briefly, genes were ranked by their differential fitness as identified by the SAM analysis and the 17 network sub-modules were used as gene sets. The GSEA procedure is then used with these inputs to test whether the genes represented in each network module (gene set) were enriched at either end of the ranked list. The location of where the genes within the module map on the ranked gene list would provide information on whether when mutated, they confer resistance (increased fitness) or sensitivity (decreased fitness) when exposed to the chosen toxicant. An FDR $\leq$ 0.1 was used to identify significant hits as using a more stringent FDR cut-off may lead to overlooking potentially significant results. Due to the way the fitness score was calculated (Chapter 2), deletion strains that infer resistance have a negative score, deletion strains that infer sensitivity have a positive score.

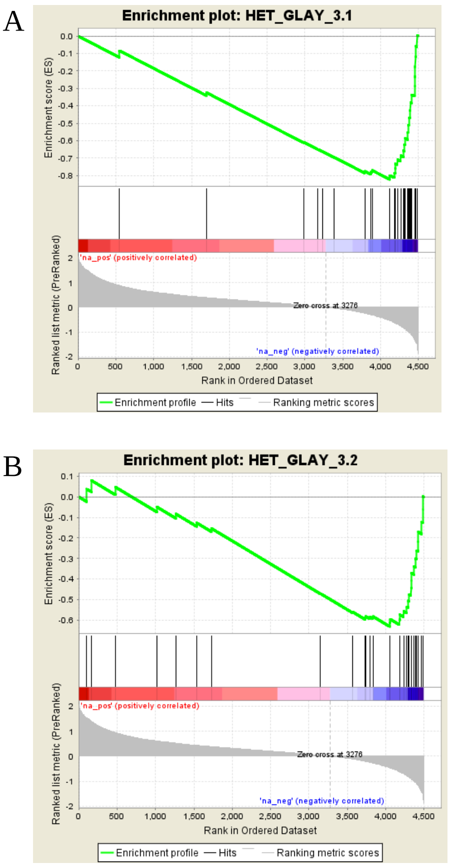
8.3 Results

8.3.1 Yeast strains mutated in ribosomal proteins and ribosomal biogenesis genes are resistant to aresenite exposure

GSEA analysis revealed that genes exhibiting differential fitness in response to arsenite exposure significantly hit two sub-modules (Table 8.1). Sub-modules 3.1 and 3.2 were both significantly negatively correlated (FDR $\leq$ 0.1), suggesting yeast strains containing this mutation have increased tolerance to arsenite. Studies have shown that arsenite interferes with protein folding which triggers the formation of toxic protein aggregates by associating the molecular chaperones [260]. A decrease in translation activity has also been reported to protect against arsenite toxicity [260], therefore the negative correlation of ribosomal biogenesis and ribosomal protein genes in response to arsenite exposure may be a defence mechanism. This is consistent with reports that mutations in ribosome biogenesis and RPs leads to increased arsenite tolerance [261] [262]. Overall yeast strains mutated in ribosomal proteins have a higher relative fitness in a high dosage metal exposure experiment than in a low dosage culture suggesting that these mutations make a yeast cell able to adapt to high concentrations of arsenite. Surprisingly there was no positive enrichment (decreased fitness) of yeast strains containing ubiquitin-proteasome pathway mutants. The ubiquitin-proteosome pathway which acts as to ease protein disaggregation and reactivate or eliminate aggregated proteins has been reported in response to arsenite exposure [263] [260], however is not observed in this data.

NAME	SIZE	NES	FDR q-val	RANK	Resistant or sensitive
HET_GLAY_3.1	30	-3.24	0	375	RES
HET_GLAY_3.2	32	-2.51	0	444	RES

Table 8. 1 Significant fitness sub-modules associated to arsenite exposure
Two sub-modules were identified as having a statistically significant negative enrichment (FDR q-val $\leq$ 0.1) to arsenite exposure. This suggests that strains containing mutations within these sub-modules confer increased fitness (resistance) when exposed to arsenite. The genes within these sub-modules are shown in Figure 8.1.



Hillenmeyer GLay sub-module 3.1 is enriched in

rRNA processing / Ribosome biogenesis (44), small ribosome subunit (26), rRNA related processes (24), translation regulation (11), ribonucleoprotein complex assembly (5), RNA polymerase I (3), RNA polymerase III (3)

Hillenmeyer GLay sub-module 3.2 is enriched in

Cytosolic large ribosomal subunit (26), translation regulation (13), RNA polymerase III (7), ribosome export (6)

Figure 8. 1 Arsenite exposure significantly hits two fitness sub-modules.
Panels A and B represent the enrichment plots for sub-modules 3.1 and 3.2 respectively. The functional annotation adjacent to each enrichment plot represents the functional enrichment of the sub-module. Red text represents a corrected FDR $\leq$ 0.05, green text represents a corrected FDR $\leq$ 0.1 and black test represents no statistically significant enrichment. Enrichment plots are separated into three key portions. The top portion with the green line represents the enrichment score for the genes within the sub-module when mapped across the ranked list of genes from the entire genome (ranked by fitness). The distinct (statistically significant) drop in enrichment score at the end (far right) of the ranked list shows that mutations in genes within the sub-module confer arsenite resistence in *S. cerevisiae*. If a statistically significant peak were present at the beginning (far left) then it would indicate that genes within the sub-module confer arsenite sensitivity to *S. cerevisiae* when mutated, however this is not the case her. The middle portion, with the vertical black lines indicates where the genes within the sub-module are located within the ranked gene list. As genes are ranked in order of fitness, the increased frequency of hits towards the end of the ranked list indicates that genes within the module predominantly confer resistence to arsenite toxicity when mutated. The bottom portion of the plot is a ranking metric indicating a gene's correlation with the phenotype (positive representing *S. cerevisiae* growth sensitivity upon arsenite exposure and negative representing *S. cerevisiae* growth resistence upon arsenite exposure). Both sub-modules have a significant negative correlation to arsenite exposure, suggesting that yeast strains containing ribosomal protein mutants increase tolerance to arsenite exposure.

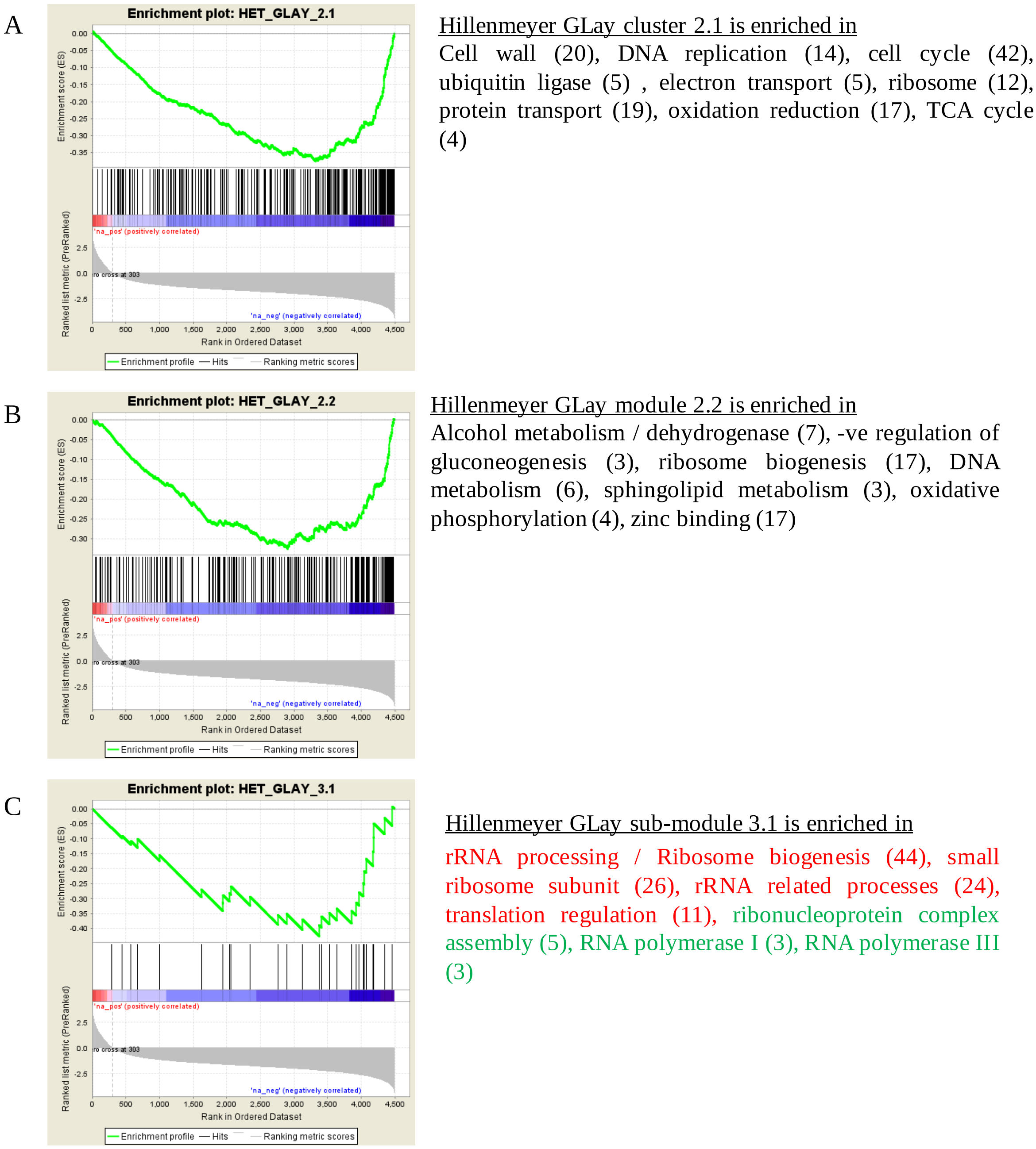
8.3.2 Fitness modules linked to zinc exposure

GSEA analysis revealed that genes exhibiting significant differential fitness (FDR ≤ 0.1) in response to zinc exposure significantly hit five sub-modules (table 8.2). Sub-module 2.1 is enriched in cell cycle and DNA replication genes (Figure 8.2A), sub-module 2.2 is enriched in ribosome biogenesis and energy metabolism pathways (Figure 8.2B). Zinc has a mechanistic role in the genetic stability and gene expression of chromatin structure, DNA replication, transcription, DNA repair and apoptosis genes [257] [264]. These results suggest that mutating genes that associate with zinc, leads to a higher tolerance when exposed to toxic levels of zinc. Similarly to arsenic, yeast strains containing ribosomal biogenesis mutants (sub-modules 3.1 and 3.2) have a higher fitness when exposed to zinc (Figure 8.2C, Figure 8.2E). The inability to form functional ribosomes inhibits translation which prevents the build up of protein aggregates. Mutating genes encoding alcohol dehydrogenase, and other metabolic enzymes (sub-modules 2.1 and 2.2) may increase resistance to zinc exposure by minimising the oxidative stress effects and production of reactive oxygen species (ROS) by decreasing activity through respiratory chains [265]. Sub-module 1.4 is enriched primarily in protein transport functions such as endoplasmic reticulum and golgi body (Figure 8.2D). Yeast strains that contain mutations in these functions may increase tolerance to zinc by minimising transport of malformed proteins caused as result of zinc exposure.

NAME	SIZE	NES	FDR q-val	RANK	Resistant or Sensitive
HET_GLAY_2.1	193	-2.04	0.003	1419	RES
HET_GLAY_2.2	30	-1.76	0.038	1031	RES
HET_GLAY_3.1	223	-1.73	0.032	1382	RES
HET_GLAY_1.4	143	-1.6	0.045	1350	RES
HET_GLAY_3.2	9	-1.55	0.078	896	RES

Table 8. 2 Significant fitness sub-modules associated to zinc exposure
(Table description continued on next page)

(Table legend continued) Five sub-modules were identified as having a statistically significant negative enrichment (FDR q-val $\leq$ 0.1) to zinc exposure. This suggests that strains containing mutations within these sub-modules confer increased fitness (resistance) when exposed to zinc. The genes within these sub-modules are shown in Figure 8.2.

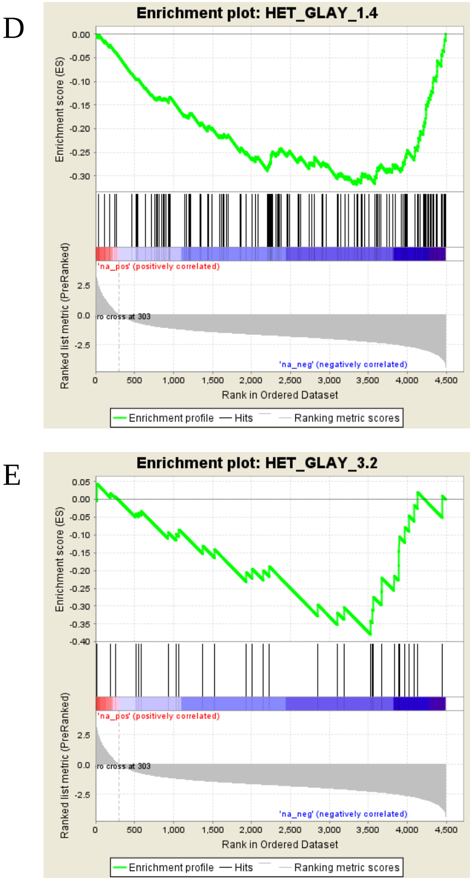


Hillenmeyer GLay cluster 2.1 is enriched in
Cell wall (20), DNA replication (14), cell cycle (42), ubiquitin ligase (5) , electron transport (5), ribosome (12), protein transport (19), oxidation reduction (17), TCA cycle (4)

Hillenmeyer GLay module 2.2 is enriched in
Alcohol metabolism / dehydrogenase (7), -ve regulation of gluconeogenesis (3), ribosome biogenesis (17), DNA metabolism (6), sphingolipid metabolism (3), oxidative phosphorylation (4), zinc binding (17)

Hillenmeyer GLay sub-module 3.1 is enriched in
rRNA processing / Ribosome biogenesis (44), small ribosome subunit (26), rRNA related processes (24), translation regulation (11), ribonucleoprotein complex assembly (5), RNA polymerase I (3), RNA polymerase III (3)

Figure 8.2 (Figure continued on next page)



Hillenmeyer GLay sub-module 1.4 is enriched in
ER (23), membrane (58), golgi membrane (5) helicase (6), translation initiation (5), mannosyltransferase (6)

Hillenmeyer GLay sub-module 3.2 is enriched in
Cytosolic large ribosomal subunit (26), translation regulation (13), RNA polymerase III (7), ribosome export (6)

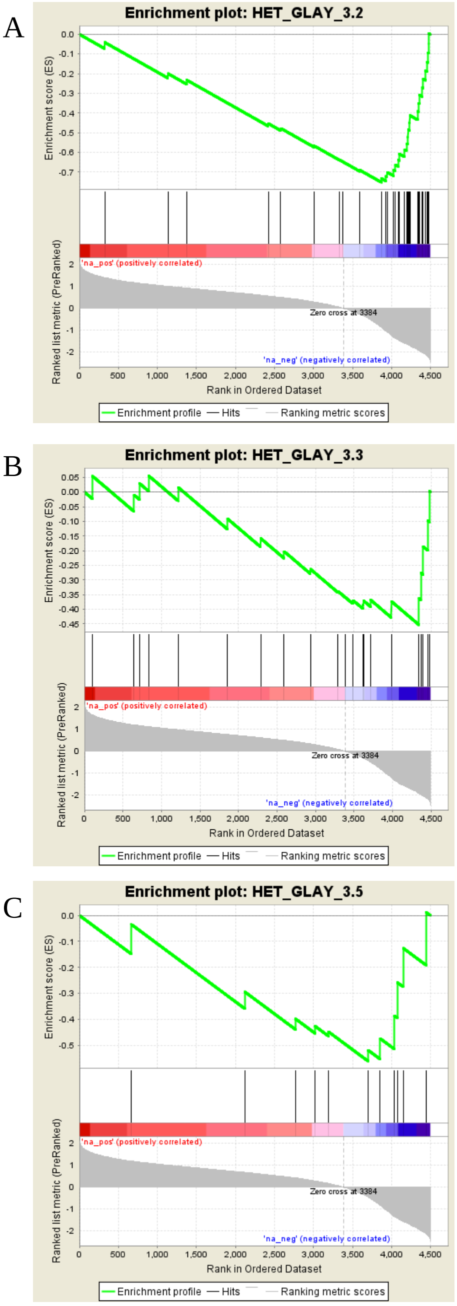
Figure 8. 2 Zinc exposure significantly hits five fitness sub-modules.
Panels A - E represent the enrichment plots for the sub-modules that are statistically significant targets of zinc exposure in *S. cerevisiae*. Panel A. *S. cerevisiae* strains containing cell wall, DNA replication and energy metabolism gene mutants are more resistant (have increased fitness) when exposed to zinc. Panel B Mutated genes involved in alcohol dehydrogenase, ribosome biogenesis and zinc binding increase resistance to zinc exposure. Panels C and E are enriched in ribosome biogenesis and ribosome proteins, suggesting lack of translation increases fitness. Panel D, mutations in protein transport and translation initiation causes increased tolerance to zinc. The functional annotation adjacent to each enrichment plot represents the functional enrichment of the sub-module. Red text represents a corrected FDR $\leq$ 0.05, green text represents a corrected FDR $\leq$0.1 and black test represents no statistically significant enrichment. All enrichment plots show the same trend. The top portion with the green line represents the enrichment score for the genes within the sub-module when mapped across the ranked list of genes from the entire genome (ranked by fitness). A distinct (statistically significant) drop in enrichment score at the end (far right) of the ranked list shows that mutations in genes within the sub-module confer zinc resistence in *S. cerevisiae*. The middle portion, with the vertical black lines indicates where the genes within the sub-module are located within the ranked gene list. As genes are ranked in order of fitness, the increased frequency of hits towards the end of the ranked list indicates that genes within the module predominantly confer resistence to zinc toxicity when mutated. The bottom portion of the plot is a ranking metric indicating a gene's correlation with the phenotype (positive representing *S. cerevisiae* growth sensitivity upon zinc exposure and negative representing *S. cerevisiae* growth resistence upon zinc exposure).

8.3.3 Yeast strains mutated in ribosome and chaperone genes exhibit tolerance to high concentrations of cadmium

Once again, there were no positively correlated sub-modules; however there were three sub-modules which represent mutant strains with increased fitness after cadmium exposure (Table 8.3). Yeast strains containing ribosomal protein and translation initiation mutants (Figure 8.3A) once again inferred tolerance against toxic metal exposure, consistent with reports that lack of translation decreases cellular toxicity by limiting the concentration of proteins that can undergo a conformational change. Surprisingly, yeast strains containing chaperone mutants, specifically the TCP-1 family have increased tolerance to cadmium exposure (Figure 8.3B). Typically, chaperones are induced upon cellular toxicity and have a key role in adapting cellular response and ensuring cell viability [266]. Sub-units of TCP -1 assist in the folding of specifically actin and tubulin proteins *in vivo* [267]. The eukaryotic cytoskeleton contains three kinds of filaments, actin filaments, intermediate filaments and microtubules and one of its functions is the intracellular transport, of vesicles and organelles [268]. Therefore, having mutations in RP and TCP-1 genes effectively shuts down protein synthesis and transport, which minimises the available protein molecules that can undergo a conformation change and form toxic aggregates. This theory can also be applied to sub-module 3, which is also enriched in protein transport and protein complexes (Figure 8.3C).

NAME	SIZE	NES	FDR q-val	RANK	Resistant or Sensitive
HET_GLAY_3.2	32	-3.18	0	628	RES
HET_GLAY_3.3	22	-1.75	0.028	150	RES
HET_GLAY_3.5	11	-1.71	0.024	798	RES

Table 8. 3 Significant fitness sub-modules associated to cadmium exposure
Three sub-modules were identified as having a statistically significant negative enrichment (FDR q-val $\leq$ 0.1) to cadmium exposure. This suggests that strains containing mutations within these modules confer increased fitness (resistance) when exposed to cadmium. The genes within these sub-modules are shown in Figure 8.3.



Hillenmeyer GLay sub-module 3.2 is enriched in
Cytosolic large ribosomal subunit (26), translation regulation (13), RNA polymerase III (7), ribosome export (6)

Hillenmeyer GLay sub-module 3.3 is enriched in
RNA polymerase II (12), chaperonin / TCP-1 (6)

Hillenmeyer GLay sub-module 3.5 is enriched in
Protein / nucleocytoplasmic transport (4), protein complex (3)

Figure 8. 3 Cadmium exposure significantly hits three sub-modules.
Panels A – C represent the enrichment plots that are statistically significant targets of cadmium exposure in *S. cerevisiae*. Panel A. Yeast strains containing ribosomal protein and translation initiation mutations had a higher tolerance to cadmium exposure. Panels B and C. Yeast strains containing mutated genes encoding protein transport or actin / microtubule chaperones also confer increased tolerance. The functional annotation adjacent to each enrichment plot represents the functional enrichment of the sub-module. Red text represents a corrected FDR $\leq$ 0.05, green text represents a corrected FDR $\leq$0.1 and black test represents no statistically significant enrichment. All enrichment plots show the same trend. The top portion with the green line represents the enrichment score for the genes within the sub-module when mapped across the ranked list of genes from the entire genome (ranked by fitness). (Figure legend continued on next page)

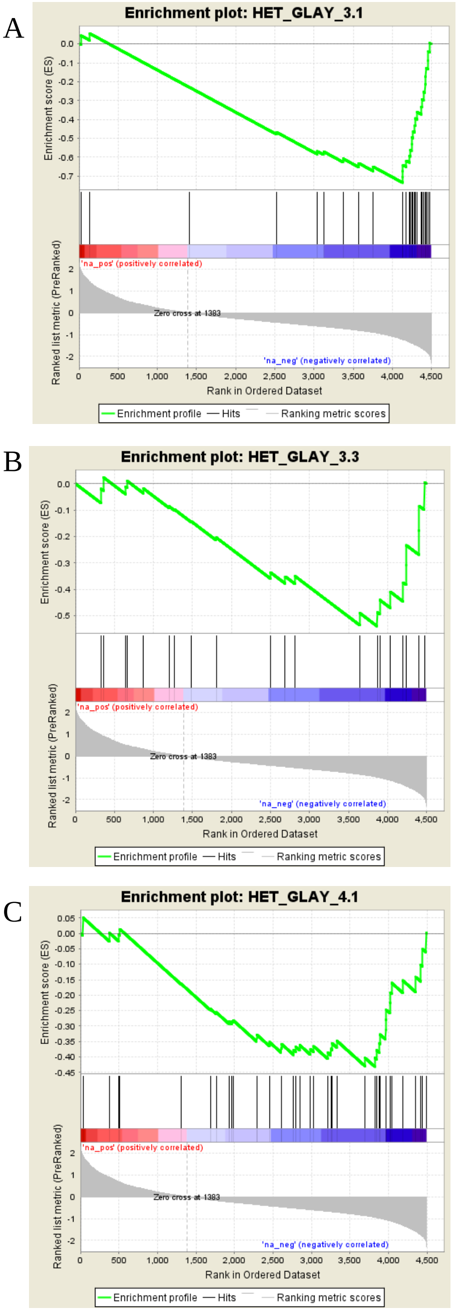
(Figure legend continued) A distinct (statistically significant) drop in enrichment score at the end (far right) of the ranked list shows that mutations in genes within the sub-module confer cadmium resistence in *S. cerevisiae*. The middle portion, with the vertical black lines indicates where the genes within the sub-module are located within the ranked gene list. As genes are ranked in order of fitness, the increased frequency of hits towards the end of the ranked list indicates that genes within the module predominantly confer resistence to cadmium toxicity when mutated. The bottom portion of the plot is a ranking metric indicating a gene's correlation with the phenotype (positive representing *S. cerevisiae* growth sensitivity upon cadmium exposure and negative representing *S. cerevisiae* growth resistence upon cadmium exposure).

8.3.4 Yeast strains mutated in ribosomal proteins and protein transport genes exhibit tolerance to high concentrations of MMA^{III}

Humans and mammals are able to methylate arsenite to form MMA^{III}, a metabolite known to be more toxic that arsenite [254]. GSEA analysis identified three sub-modules, all of which inferred resistance onto the yeast strain (Table 8.4). Once again the significant negative correlation of sub-modules 3.1 and 3.3 (Figure 8.4A, Figure 8.4B) suggests that yeast strains mutated in ribosomal proteins have a higher relative fitness to high dosage metal exposure compared to low dosage, as does sub-module 4.1, which is enriched in glucose metabolism and endoplasmic reticulum genes.

NAME	SIZE	NES	FDR q-val	RANK	Resistant or Sensitive
HET_GLAY_3.1	30	-2.42	0	366	RES
HET_GLAY_3.3	22	-1.68	0.029	630	RES
HET_GLAY_4.1	40	-1.54	0.081	673	RES

Table 8. 4 Significant fitness sub-modules associated to MMA^{III} exposure
Three sub-modules were identified as having a statistically significant negative enrichment (FDR q-val $\leq$ 0.1) to MMA^{III} exposure. This suggests that strains containing mutations within these modules confer increased fitness (resistance) when exposed to MMA^{III}. The genes within these sub-modules are shown in Figure 8.4.



Hillenmeyer GLay sub-module 3.1 is enriched in

rRNA processing / Ribosome biogenesis (44), small ribosome subunit (26), rRNA related processes (24), translation regulation (11), ribonucleoprotein complex assembly (5), RNA polymerase I (3), RNA polymerase III (3)

Hillenmeyer GLay sub-module 3.3 is enriched in

RNA polymerase II (12), chaperonin / TCP-1 (6)

Hillenmeyer GLay sub-module 4.1 is enriched in

regulation of glucose metabolism carbohydrate (3), ER (3)

Figure 8. 4 MMA^{III} exposure significantly hits three sub-modules.

Panels A – C represent the enrichment plots that are statistically significant targets of MMA^{III} exposure in *S. cerevisiae*. Panel A. *S. cerevisiae* strains containing ribosomal protein and translation initiation gene mutations had a higher tolerance when exposed to MMA^{III}. Panel B Stains lacking genes encoding protein transport or actin / microtubule chaperones also had higher tolerance, as do yeast strains containing mutated glucose metabolism and protein transport genes (Panel C). The functional annotation adjacent to each enrichment plot represents the functional enrichment of the sub-module. Red text represents a corrected FDR $\leq$ 0.05, green text represents a corrected FDR $\leq$0.1 and black test represents no statistically significant enrichment. All enrichment plots show the same trend. The top portion with the green line represents the enrichment score for the genes within the sub-module when mapped across the ranked list of genes from the entire genome (ranked by fitness). (Figure legend continued on next page)

(Figure legend continued) A distinct (statistically significant) drop in enrichment score at the end (far right) of the ranked list shows that mutations in genes within the sub-module confer MMA^{III} resistence in *S. cerevisiae*. The middle portion, with the vertical black lines indicates where the genes within the sub-module are located within the ranked gene list. As genes are ranked in order of fitness, the increased frequency of hits towards the end of the ranked list indicates that genes within the module predominantly confer resistence to MMA^{III} toxicity when mutated. The bottom portion of the plot is a ranking metric indicating a gene's correlation with the phenotype (positive representing *S. cerevisiae* growth sensitivity upon MMA^{III} exposure and negative representing *S. cerevisiae* growth resistence upon MMA^{III} exposure).

8.3.5 Exposure to lead does not cause differential fitness in any yeast mutants

Surprisingly neither SAM nor GSEAPreranked identified any yeast strains with differential fitness (Supplementary CD, folder entitled 'Chapter 8').

8.4 Discussion

Though this study focussed on only a limited set of metal and metalloid exposures, the most important finding in this study was the demonstration that fitness data can be used to identify targets of metal and metalloid toxicity. A shared result across toxicant exposure was that increased fitness was observed in strains containing mutants involved in ribosome formation, translation and protein synthesis, consistent with reports that translation repression increases cell survivability [261] [262]. This preliminary study has demonstrated how fitness data can be applied to identify adverse outcome pathways in *S. cerevisiae.*

8.4.1 Potential targets of metals and metalloids reveal the mechanisms of toxicity

Metal ions are essential for cell viability in most organisms. However, a delicate balance must be maintained as high concentrations of metal ions are known to cause toxicity in both mammals and microorganisms alike [257]. Metal ion exposure holds great significance in environmental and occupational studies, as exposure to heavy metals such cadmium and lead can lead to prenatal and developmental defects in humans [258]. Metal and metalloid toxicity is often caused by imprecise folding of affected proteins, causing a build up of abnormal proteins, leading to cell toxicity [260]. Cells with diminished translation activity have been shown to protect against arsenite toxicity, furthermore mutations in genes encoding RPs and biogenesis factors increase arsenite tolerance [261] [262]. Although RPs are essential for cell viability, many are duplicated; therefore mutating a single copy is non-lethal. Deletions of duplicated RPs have known to delay ribosome assembly and diminish the rate of translation. The metals and metalloids tested in this study suggest that toxic effects are minimised upon exposure to yeast strains lacking ribosome and translation functionality.

Complementary to the results, in normal yeast strains, toxic doses of arsenite and cadmium are believed to cause disruption to cellular protein structure [269]. Therefore, to protect against toxicity, cells must repress translation. The presence of stress granules within the cytosol is linked with stress response, and occurs when translation initiation is aborted or when ribosomes stall on mRNA [270]. Formation of stress granules induces phosphorylation of eukaryotic initiation factor α (eIF2α). eIF2α is essential for protein synthesis, and usually forms a ternary complex with GTP and the initiator tRNA Met-tRNA, however phosphorylation of eIF2α causes a reduction in the levels of the eIF2α – GTP – Met-tRNA ternary complexes, leading to a decrease in translation initiation rates [271]. This demonstrates a possible mechanism for translation repression in normal yeast strains which is essential for establishing tolerance. Hence, this is why yeast strains containing mutant RP and ribosomal biogenesis genes, show increased differential fitness, as the rate of translation is also repressed.

A very interesting observation was that chaperone proteins involved in the post-translation modification of actin and tubulin proteins may contribute to cadmium and MMA^{III} toxicity. We identified that yeast strains containing mutants of the TCP-1 complex had increased resistance to high concentrations of cadmium and MMA^{III}. This was a very surprising result, as typically the canonical function of a chaperone is to aid in the correct folding of unfolded and misfolded proteins [272]. The TCP-1 ring complex (also known as TRiC) is 900 kDa complex containing two hetero-oligometic protein rings, each ring is made of up eight homologous sub-units encoded by the essential genes *CCT1- CCT8* [262]. The TCP-1 complex is required for the correct folding of actin and tubulin proteins [262], accumulation of unfolded β-tubulin is known to be toxic [273]. Our results show that when exposed to high concentrations of cadmium and MMA^{III}, yeast strains containing mutations encoding the TCP-1 complex had a higher fitness than when exposed to lower concentrations. This suggests that at high concentrations, cadmium and MMA^{III} inhibit TCP-1 function in normal yeast cells, by repressing the ability to correctly fold actin and tubulin proteins, leading to cell toxicity. Arsenite has also been reported to target the TCP-1 complex [262]. These results fit well with my results demonstrating that strains lacking ribosome and translation related genes, have a higher fitness. Under normal conditions, cadmium and MMA^{III} may target chaperone proteins directly, which means newly synthesised proteins are unable to undergo correct folding, leading to a build-up of toxic protein aggregates. By knocking out genes encoding the TCP-1 complex, it's possible that cadmium and MMA^{III} are unable to be as effective. Combined with yeast cells natural defence mechanisms used to repress translation (as described above), it all could infer an increased fitness for yeast strains containing TCP-1 mutants.

Interestingly, none of the metals tested hit fitness sub-module 3.4 which is significantly enriched in proteasome functions (Chapter 2, section 2.3.3.3). A general trend across our results was that mutants of RPs and translation related genes are more tolerant of metal exposure due to the lack of protein synthesis [260] [261] [262]. However, proteins synthesised prior to exposure would still be a target of metals and metalloids. As such, it would be expected that yeast proteasome mutants would be hypersensitive to metal and metalloid exposure and therefore be characterised by a low fitness after exposure, however this was not observed. An explanation for this is that possibly the genes within sub-module 3.4 are not induced during a metal or metalloid stress response.

8.4.2 Shortcomings and further work

Upon completing this work, I identified two factors which may have limited this study. Firstly, an unexpected result was that no modules were enriched upon lead exposure, furthermore lead exposure did not cause differential fitness in yeast strains between the lowest and highest dosages (as detected by SAM). This suggests that the cellular mechanisms involved in defending against lead exposure do not become more apparent as lead concentration increases. This is a consideration that needs to be taken into account in future studies, possibly by incorporating a control sample (taken prior to exposure). Incorporating a control sample would capture genes that are induced upon toxicant exposure, in addition to genes that are induced as toxicant concentration increases. The current study compared fitness scores between the lowest toxicant dose and the highest toxicant doses, with the aim of identifying genes that are induced / repressed as the toxicant concentration increased. Secondly, only five toxicants were used in this study. Though the results are consistent with those reported in current literature, a greater number of metals and metalloids would be pertinent in validating the use of fitness data for adverse outcome pathway studies. This limitation is due to the lack of fitness data in *S. cerevisiae* that focuses on metal exposure response. This study made use of two of the largest fitness compendia available for *S. cerevisiae*, utilising the metal exposure fitness data from Vulpe Lab together with the fitness network constructed from Hillenmeyer's compendium. However given the recent surge in popularity of using fitness data as a means of analysing biological systems, it is only a matter of time until this data becomes available. As this study was only a preliminary look into using fitness data to identify adverse outcome pathways for metal and metalloid toxicity, there are multiple ways this study could be expanded. One possible direction would be to expand the analysis to chemical exposures rather than just metal exposure. This could be done by integrating the Hillenmeyer fitness networks constructed in Chapter 2 with molecular descriptors. Molecular descriptors, also termed physico-chemical features (PCFs) describe the 2D and 3D topological, electrostatic, geometrical and atomic properties of a chemical [274]. Integrating my fitness networks with these descriptors would require the use of a genetic algorithm, which in theory would predict PCFs that were signatures of the fitness modules identified in Chapter 2. Specifically, it would be possible to identify what properties of a chemical are responsible for affecting cell viability.

8.5 Concluding remarks

Though there were limitations to this study, including a lack of diverse toxicants and the lack of a control sample, this chapter demonstrated how fitness data can be used to identify of adverse outcome pathways in *S. cerevisiae.* It marked a novel preliminary approach in applying fitness data to identify targets of metal toxicity. The results demonstrated that in response to toxic levels of zinc, arsenic, cadmium and MMA^{III} , *S. cerevisiae* strains lacking the ability to translate proteins all had higher resistance that those that did not have the mutation. This is consistent with the down-regulation of protein machinery in response to metal toxicity in normal *S. cerevisiae* strains which aims to minimising the concentration of misfolded potentially toxic protein aggregates [260] [261] [262]. This study was a preliminary piece of work that highlights the diverse applications of fitness data. The next stage in applying fitness data would be to identify targets of chemical exposure, rather than limiting the study to only metals and arsenicals.

CHAPTER 9: GENERAL DISCUSSION

9.1 Exploring the global and local organisation of the yeast system

The ability to quantify measurements of multi-level systems such as gene expression, phenotypic growth and protein binding has led to the generation of extraordinary amount of data. A systems biology approach provides a powerful means to analyse the consequences of perturbations on a biological species. The application of reverse engineering methods with single level data is useful in inferring underlying regulatory networks without prior knowledge, and aids in the inference of novel relationships between different entities. Moreover the integration of multi-level datasets using computational methodologies has the potential to identify biological models that exhibit significantly correlated behaviour across the diverse data, which can then be tested experimentally. The aim of this study was to apply a systems biology and network inference approach to available yeast fitness and expression datasets. This thesis demonstrated how network interrogation techniques such as modularisation and functional association can be used to characterise underlying global biological networks for *S. cerevisiae* and *S. pombe*. The relationships inferred between functional modules sharing co-expression, co-fitness or both, allowed for a hypothesis driven analytical approach, identifying potentially significant new areas for study. The results provided both novel hypotheses as well as additional evidence supporting existing biological pathways. This work focused particularly on applying these methods to elucidate the relationship between energy metabolism and ribosome biogenesis.

9.2 Understanding cell cycle progression in yeast

Analysis of *S. cerevisiae* fitness data identified a potential mechanism in the control of cell cycle progression involving the non-essential cell cycle checkpoint protein BUB1 (Chapter 2) and cytosolic RPs. The highly significant co-fitness profiles between *S. cerevisiae* strains containing deletions of *BUB1* and genes encoding cytosolic RPs suggested that there may be an underlying biological connection between these genes. One possible hypothesis was that the lack of a functional ribosome may cause cell cycle arrest at the anaphase checkpoint, a response which is also observed when *BUB1* is knocked out [134]. The involvement of RPs and ribosome biogenesis proteins in maintaining cell cycle progression has been reported before. As discussed in Chapter 2, inactivation the ribosome biogenesis gene *RRB1*, is reported to cause abnormal chromosome segregation [136]. What is novel however, is that the analysis of fitness data conducted in Chapter 2 revealed that *BUB1* exhibited statistically significant co-fitness to over 45 cytosolic RPs, with the most significant correlation being to genes encoding small 40S RPs. In fact, the top 25 most significant edges between RPs and *BUB1*, 21 belonged to 40S RPs. The reason as to why *BUB1* has higher co-fitness to genes encoding 40S RPs is currently unknown, however it was also observed in the fitness and expression integrated network (Chapter 4), in which the HOPACH cluster representing chromosome segregation was a first neighbour of the HOPACH cluster representing small RPs. In addition, no edges were observed between chromosome segregation and large RPs. Furthermore, BUB1 requires other spindle checkpoint components into order to fulfil its function, including the MAD protein family and additional BUB proteins [135]. With this in mind, it is extraordinary as to why only *BUB1* shows such a significant phenotypic correlation to RPs.

It can be argued that the linkage between *BUB1* and genes encoding RPs is simply a consequence of them both contributing similar fitness to *S. cerevisiae* cells, and that there is no underlying biological relationship between them. *BUB1* is a non-essential gene which has a paralogue (*MAD3*) [130], some RPs also have paralogues which arose as a result of a whole duplication event. Therefore it could be that *BUB1* has a greater correlation to small RPs simply because deletion of both genes leaves the cell in a less fit yet still viable state. This explanation is plausible, however it is important to keep in mind that some of the small RPs significantly correlated to *BUB1* (shown in Table 2.14) do not have a paralogue and are fatal when deleted (such RPS20, RPS13, RPS3, RPS15, RPS2). The question then, is why does *BUB1* have such a strong correlation to genes encoding RPs, why is this correlation stronger with genes that encode 40S RPs rather than those that encode 60S RPs and why is non-essential *BUB1* strongly correlated to genes that encode essential RPs? This is an area of research which requires further investigation in order to identify the true underlying cause including validation using wet lab experiments. A shortcoming of this analysis is that the linkage between *BUB1* and RPs was only observed in the Hillenmeyer fitness network, this feature was not observed in the Vulpe fitness network (Chapter 2). This may be due to the limited size of the Vulpe fitness dataset compared to the extensive Hillenmeyer fitness compendium.

9.3 A new perspective in understanding the intricacies of ribosome biogenesis in *S. pombe*

The fitness networks reported in Chapter 2 took into consideration entire fitness datasets and encapsulated the global as well as local organisation of the *S. cerevisiae* system. The analytical pipeline used in this study was a novel way of applying fitness data. This is also true for the approach used to analyse the comprehensive TF knockout data for the *S.*

cerevisiae expression network as well as the Bähler compendium used to construct the *S. pombe* expression network. The methodogies applied in this thesis allowed for the genome-wide analysis of yeast fitness and expression data, as well as the integration and analysis of genome-wide multi-level data. This resulted in the identification of several potentially interesting hypotheses. The investigation into the linkage between RPs and energy metabolism genes is just a single example of how network based hypothesis driven research can reveal additional potentially interesting functional interactions. Furthermore, prior to this study, there had not been a comprehensive investigation into riboneogenesis in *S. pombe*. Below I discuss the contribution each study within this thesis made to understanding and elucidating riboneogenesis in yeast.

9.3.1 Similarities between *S. pombe* FBP1 and *S. cerevisiae* SHB17

Sedoheptulose-1, 7-bisphosphatase (SHB17) is the enzyme responsible for catalysing the committed step of riboneogenesis in *S. cerevisiae* [92], however the network analysis conducted in *S. pombe* suggested that FBP1 may instead catalyse the committed step in riboneogenesis. As discussed in Chapter 5 (section 5.4.1), *S. pombe* does not have a dedicated sedoheptulose-1, 7-bisphosphatase, rather, FBP1 has the potential to act as both a sedoheptulose-1, 7-bisphosphatase (accepting SBP as the substrate) [58] or a canonical fructose-1, 6-bisphosphatase (accepting FBP as the substrate). Unlike *S. pombe*, *S. cerevisiae* has its own fructose 1, 6-bisphosphatase and sedoheptulose-1, 7-bisphosphatase (SHB17). FBP1 in *S. cerevisiae* does not have sedoheptulose-1, 7-bisphosphatase activity suggesting that in *S. cerevisiae*, SHB17 is the enzyme dedicated to catalysing the committed step in riboneogenesis whilst FBP1 catalyses the rate limiting step in gluconeogenesis.

In *S. pombe* however, FBP1 may fulfil the role of SHB17 in addition to its canonical FBP role. Therefore, its role in gluconeogenesis and riboneogenesis regulation may be dependent on cellular demands. The dual function of FBP1 in *S. pombe* is supported by the structural similarities between FBP and SBP and studies in other organisms (discussed in section 5.4.1), suggesting that both substrates have the potential to bind FBP1. The mechanisms that dictate which substrate binds FBP1 and when, is not fully understood, but I hypothesised that the conformation of FBP and SBP (whether they bind in cyclic or extended form) is the key factor which determines the role that FBP1 has in the cell (discussed in section 5.4.2). Below I provide a possible mechanistic model based on results obtained during this project and reported in existing literature.

9.3.2 FBP1 switches enzymatic activity depending on cellular demands for ribose-5-phosphate

Due to the lack of a dedicated sedoheptulose 1, 7-bisphosphatase in *S. pombe,* I hypothesise that FBP1 in fact has a dual role, switching between acting as a fructose 1, 6-bisphosphatase or sedoheptulose 1, 7-bisphosphatase depending on cellular demands. During times of rapid growth when increased protein translation is required, glycolysis flux increases, raising the concentrations of glycolytic intermediates that can be used for riboneogenesis. During such states FBP1 acts as a sedoheptulose-1, 7-bisphosphatase, catalysing the first committed step of riboneogenesis, therefore fulfilling the cellular demands for ribose-5-phosphate. The mechanism by which FBP1 may switch enzymatic activity is not known, however one hypothesis supported by the network analysis done in Chapter 5 suggests that the binding of ubiquitin to the allosteric binding site of FBP1 may cause a conformation change to the active site which allows SBP to bind in its cyclic form **(**Figure 9.1**)**. This hypothesis is supported by the fact that the allosteric binding site of FBP1 is located at the N-terminus [186] and that mammalian FBP1 contains a conserved lysine residue within the allosteric site [188]. There have been reports that the N-terminus of the target protein has been used for ubiquitination [275] [276], as well as the canonical lysine residue [191], therefore it is plausible that ubiquitin may associate to the allosteric site of FBP1. Upon binding, a conformation change in the active site of FBP1 may occur, allowing SBP to bind in its preferred cyclic form, thereby increasing the affinity of FBP1 for SBP, rather than for FBP **(**Figure 9.1C**).** Studies in pig kidney have reported that specific regions of FBP1 have a higher level of disorder, making them susceptible to multiple conformations [186], this may impact the structure of the active site.

Conversely, when growth on a non-fermentable carbon source is required, such as during glucose starvation, there is an up-regulation of genes encoding proteins involved in gluconeogenesis [277], and a co-ordinated down regulation of genes involved in translation elongation and initiation [278]. The lack of glucose also leads to the repression of glycolysis and increased degradation of glycolytic enzymes [278]. The anti-correlated expression between gluconeogenesis and RPs may be due to the demand for glucose being higher than that of ribose-5-phosphate, and as a result FBP1 exhibits its canonical fructose-1, 6-bisphosphatase activity to increase flux through gluconeogenesis. The lack of sedoheptulose 1, 7-bisphosphatase activity exhibited by FBP1 subsequently leads to the repression of riboneogenesis (Figure 9.1 B).

This hypothesis assumes that riboneogenesis in *S. pombe* can only occur during times when glucose is readily available, consistent with reports in *S. cerevisiae* [92]. The data suggests that FBP1 is a strong candidate for the role as the key enzyme in determining whether the *S. pombe* cells proceeds through gluconeogenesis or riboneogenesis.

However, this hypothesis is limited in two ways. The first, there is currently no experimental evidence or comprehensive studies that have investigated whether FBP1 in *S. pombe* changes its role when bound by ubiquitin and secondly as discussed in section 5.4.1, reciprocal BLAST searches did not identify *S. pombe FBP1* as a potential orthologue for *S. cerevisiae SHB17*. This study, has however, identified potentially key differences in riboneogenesis between *S. pombe* and *S.cerevisiae*. Given the divergence in carbon metabolism between these two species [66], it is plausible that *S. pombe* does in fact regulate riboneogenesis differently to *S. cerevisiae.* This study has identified several candidate genes and proteins which can serve as a means of experimentally testing this hypothesis (section 5.4.3) .

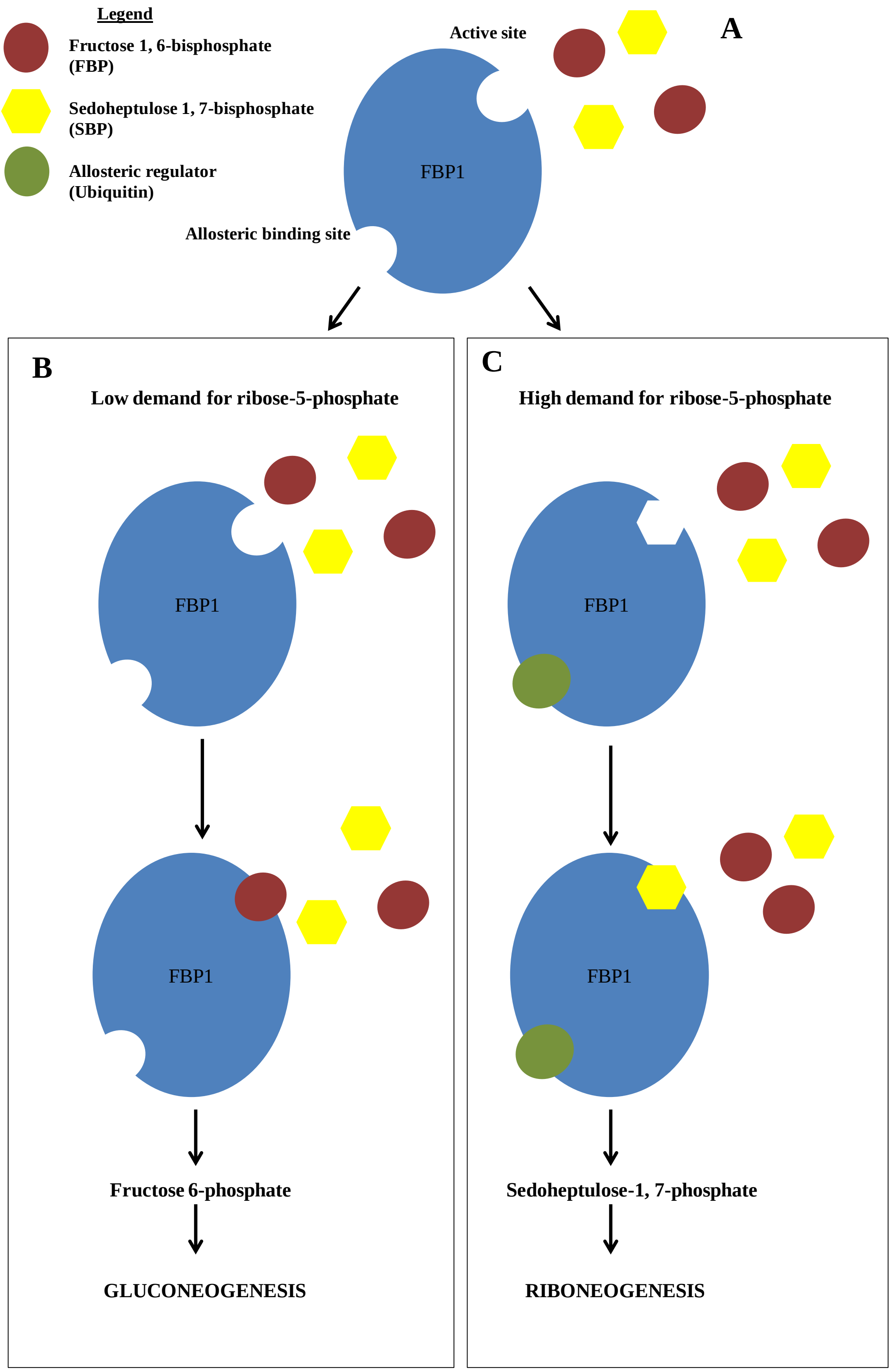

Legend
Fructose 1, 6-bisphosphate (FBP)
Sedoheptulose 1, 7-bisphosphate (SBP)
Allosteric regulator (Ubiquitin)
Active site
A
FBP1
Allosteric binding site
B
Low demand for ribose-5-phosphate
FBP1
FBP1
Fructose 6-phosphate
GLUCONEOGENESIS
C
High demand for ribose-5-phosphate
FBP1
FBP1
Sedoheptulose-1, 7-phosphate
RIBONEOGENESIS

Figure 9. 1 A flow chart representing the hypothesised dual functionality of FBP1 in *S. pombe* based on results obtained throughout this study and evidence from existing literature.

Panel A. A cartoon showing the structure of FBP1 (represented in blue) with its allosteric and active site. (Figure legend continued on next page)

(Figure legend continued) Coloured shapes represent the FBP, SBP and ubiquitin as shown in the figure legend. FBP1 can either have a role in gluconeogenesis (Panel B) or riboneogenesis (Panel C), depending on the demand for ribose-5-phosphate. Panel B. During periods when ribose-5-phosphate demand is in low demand and the demand for glucose is high (such as glucose starvation) FBP1 acts as a fructose-1, 6-bisphosphatase and catalyses the key rate limiting step in gluconeogenesis. The shape of the active site, allows FBP to bind in its cyclic form, therefore the affinity of FBP1 for FBP is higher than that of SBP. In some circumstances finite regulation of FBP1 is also dictated by the binding of AMP to the allosteric site (not shown in figure). FBP is then catalysed to fructose 6-phosphate, an intermediate product of gluconeogenesis. Panel C. During periods when ribose-5-phosphate demand is high, such as during rapid cell growth, FBP1 exhibits sedoheptulose-1,7-bisphosphatase activity. Ubiquitin binds to the allosteric site of FBP1 with the aid of a ubiquitin ligase. The binding of ubiquitin causes a change in the active site of FBP1, allowing SBP to bind the active site in its cyclic form rather than its extended linear form, thereby increasing FBP1 affinity for SBP rather than FBP. In a thermodynamically driven reaction, FBP1 catalyses SBP to sedoheptulose-1, 7-phosophate, which is a key intermediate in riboneogenesis. When ribose-5-phosphate demand decreases, the ubiquitin molecule may disassociate from FBP1 through the aid of a deubiquitinase (not shown in figure), returning FBP1 to its canonical function in gluconeogenesis.

9.3.3 Can RPs regulate their own synthesis by effecting the expression of *FBP1* and glycolysis genes?

Eukaryotic cells produce and import RPs into the nucleus; the majority are incorporated into ribosomal subunits, which are subsequently exported back into the cytoplasm to be assembled into a mature ribosome [98]. However, RPs are imported into the nucleus in excess of demand, which means that at any given time there are a pool of unassociated RPs in both the nucleus and cytoplasm which are free to perform other functions aside from their canonical role in ribosome formation [98]. Unassociated RPs have been reported to have many additional functions such as stabilising and protecting proteins via protein – protein interactions [279], as well as roles in mRNA processing, transcription, translation, DNA repair and apoptosis [100]. The role of RPs in mRNA processing is particularly interesting as in both eukaryotes and prokaryotes, RPs have been reported to regulate their own gene expression as well as the expression of other genes by DNA and RNA interactions to create structures capable of affecting gene expression. The ChIP-

chip data presented in Chapter 6 showed that RPs bind to many of the same genomic loci, suggesting that they may bind as a silent incomplete protein complex. Their presence at specific genomic loci suggests that they may have a role in regulating gene expression. The mechanisms by which RPs regulate expression of genes are numerous and varied **(**Figure 9.2**)**. RPs are able to bind their own premRNA and inhibit expression (Figure 9.2A). In eukaryotes for example, RPS14 binds to its own pre-mRNA and inhibits transcription [280]. RPs can also affect splicing, such as *S. cerevisiae* RPL30, which inhibits splicing by binding to the intron exon junctions **(**Figure 9.2B**)** within its own transcript, it can also bind other mature mRNAs and inhibit translation [97] (Figure 9.2C) Ultimately, many RPs have the potential to regulate their own expression and the expression of other genes.

In accordance with induction of riboneogenesis being heavily dependent on cellular conditions and demands, it could be that RPs themselves repress their own synthesis by binding to gluconeogenesis gene *FBP1* and glycolysis genes involved in riboneogenesis. In Chapter 6, I reported that *FBP1* is bound by RPL7, RPL11 and RPL25. The conclusion was that the similarity in binding patterns between the three RPs was likely due to excess RPs that were not incorporated into ribosome subunits, binding to genes as part of a non-functional ribosomal complex [50]. This suggests that when there are excess RPs (therefore a low demand for ribose); they associate to many genomic loci, including *FBP1* and glycolysis genes. By associating to these genes, they reduce the expression of fructose-1, 6-bisphosphatase and glycolytic proteins respectively, effectively inhibiting the riboneogenesis pathway.

The ChIP-chip analysis revealed that the following energy metabolism genes were bound by RPs; *FBA1*, *PGK1*, *PYK1*, *TDH1* and *FBP1*. Therefore it could mean that when ribose demand is low, the excess RPs bind glycolysis genes and *FBP1* directly as part of a feedback loop using one of the mechanisms shown in Figure 9.2, shutting down glycolysis and inhibiting expression of *FBP1*. When ribose demand is high, the preassembled silent ribosomal complexes may disassociate from the glycolysis genes (possibly to form functional ribosomes). As a result, the genes encoding key enzymes in the glycolytic pathway are expressed, thereby increasing flux through glycolysis (Figure 9.1B). The glycolytic intermediates can then be shunted into the non-oxidative arm of the PPP where FBP1 catalyses the first committed step in riboneogenesis. Thus explaining why mutations in RPs leads to a phenotypic effect that is not typically associated to RP function. Such as those reported in Chapter 2 in which strains containing deleted RP genes were shown to have a statistically significant correlation to strains containing deleted glycolysis and hexose metabolism genes, indicating both strains had a highly similar phenotype . To test whether the association of RPs to these genes is truly because of an underlying regulatory mechanism is an area of the study which requires exploration and experimental validation.

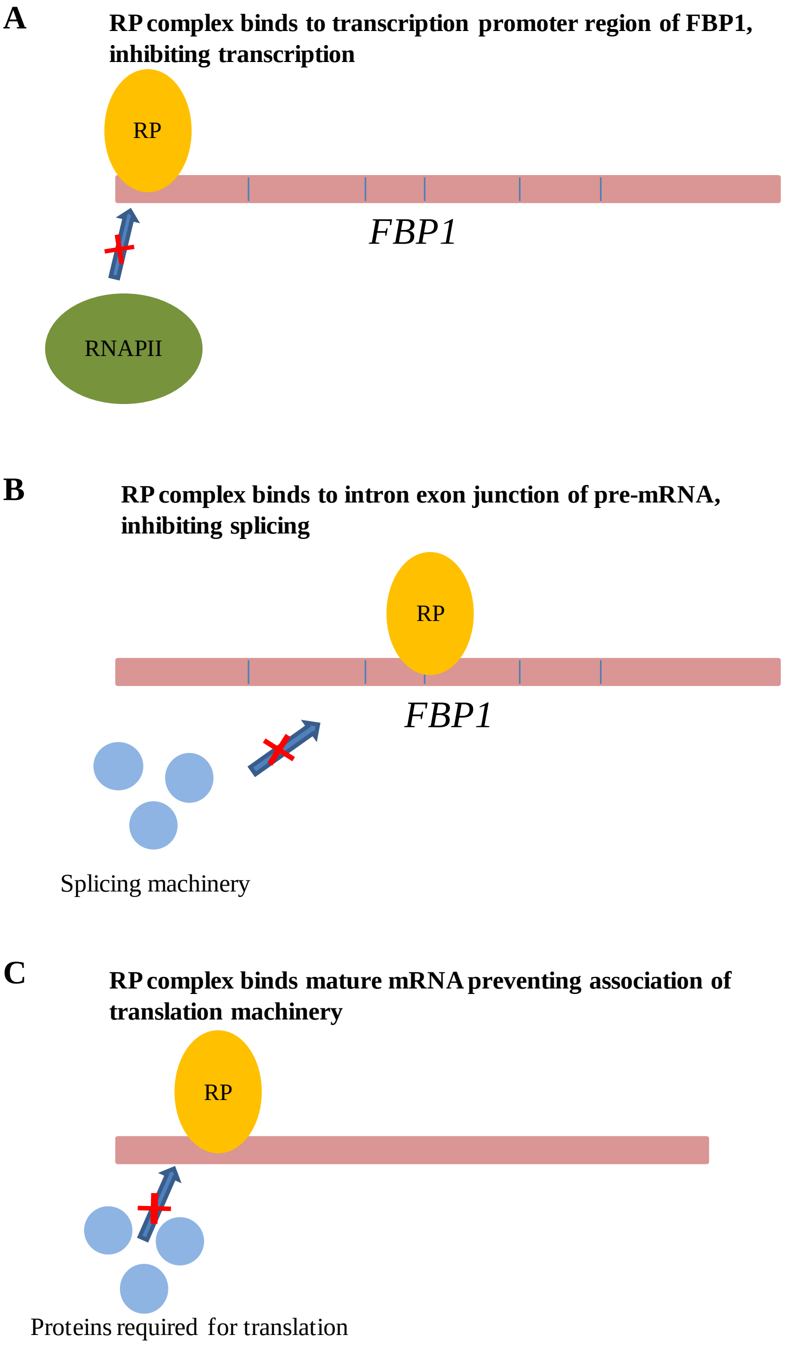

A
RP complex binds to transcription promoter region of FBP1, inhibiting transcription
RP
FBP1
RNAPII
B
RP complex binds to intron exon junction of pre-mRNA, inhibiting splicing
RP
FBP1
Splicing machinery
C
RP complex binds mature mRNA preventing association of translation machinery
RP
Proteins required for translation

Figure 9. 2 Possible mechanisms of action in which RPs inhibit *FBP1* and glycolysis gene expression when demand for ribose-5-phosphate is low.

The figure represents three mechanisms reported in literature by which RPs may regulate the expression of genes involved riboneogenesis. The pink bars represent codons, blue bars represent introns, the yellow oval represents a RP complex. Panel A. RPs bind the promoter region of *FBP1* preventing the recruitment of transcription machinery. Panel B. RPs bind the intro exon junctions of pre-mRNA inhibiting splicing from occurring, thereby preventing the manufacture of mature mRNA. Panel C. RPs bind the promoter region of the mature mRNA, inhibiting the recruitment of components required for translation, thereby inhibiting translation of the gene.

9.3.4 Understanding ribosome biogenesis: Expanding the scope to higher eukaryotes

Since their sequencing, both *S. cerevisiae* and *S. pombe* have been used as a model for analysing human disease. Of the 4824 proteins in the *S. pombe* genome, 172 are related to human disease proteins, of which 50 have statistically significantly similarity (as calculated using BLASTP) [57]. Half of these significant genes are cancer related [57]. A similar number of genes in *S. cerevisiae* (182) are identified as having similarities with human disease proteins.

Many tumour cells show up-regulation of the non-oxidative arm of the pentose phosphate pathway [281]. Numerous studies have reported and confirmed the link between tumour cell growth and ribose-5-phosphate production [282, 283]. Therefore studying the process by which glycolytic intermediates are converted to ribose-5-phosphate in yeast, may yield new directions for human cancer research. In addition to cancers, human diseases such as Wernicke–Korsakoff syndrome [284] and maturity-onset diabetes of the young (MODY) [285] are caused by dysfunctional transketolase and hexokinase enzymes respectively. These classes of proteins are involved in regulating glycolysis and riboneogenesis, and were identified as first neighbours to RPs in my analysis. By elucidating the roles of rate limiting enzymes in energy metabolism pathways, it is possible to identify drug targets [93]. Experimental work may shed some light on the possible mechanisms of diseases that are caused by defects in energy metabolism pathways. In *S. pombe*, the hypothesised dual role of FBP1 may aid in understanding how to control the rate of gluconeogenesis and tackle type 2 diabetes [286]. Type 2 diabetes is caused by the excessive glucose production via gluconeogenesis [286]. Fructose bisphosphatases catalyse a rate limiting reaction in gluconeogenesis and therefore have the potential to control the rate of glucose production; because of this they are attractive targets for the development of drugs aimed to tackle type 2 diabetes [287]. Verifying whether FBP1 does indeed accept both FBP and SBP as substrates depending on the structure of its active site may open new possibilities on how to control flux through gluconeogenesis.

9.4 Limitations and Future work

9.4.1 The lack of fitness data in *S .pombe*

In order to achieve a more comprehensive understanding of the *S. pombe* system, ideally fitness data should have also been used. However, currently there is a severe lack in *S. pombe* fitness data. One study, published in March 2010 by Kim *et al*, constructed the first deletion library for *S. pombe* and mutated 98.4% of the fission yeast genome [73] using the *KanMX* deletion cassette, the same methodology used in existing *S. cerevisiae* fitness studies [36] [37] [41]. The experiment was limited as it only used one condition, exposure to rich media. The study used a six replicate design and this lack of samples meant that the analytical pipeline used in Chapters 2 – 5 could not be applied; as MI based reverse engineering methods require more than 50 samples to be effective. Furthermore, in December 2010, the *S. pombe* deletion library previously constructed by Kim *et al*, was reported to be imprecise, as up to 30% of the DNA barcodes, which are vital for determining the fitness contribution of each gene, may have deviated from their original design. This meant the entire deletion library for *S. pombe* had to be verified using deep sequencing [74]. As a result, a comprehensive deletion library for *S. pombe* fitness experiments has only been available from 2011 onwards, and therefore the volume of fitness data for *S. pombe* is lacking compared to *S. cerevisiae*. There does exist a preliminary study published by Han *et al* which used with the newly validated *S. pombe* deletion library [74], however it only contained four conditions, including exposure to three types of genotoxin and anti-microtubule compound thiabendazole (TBZ). Again, the lack of samples means that this dataset cannot be used for network inference in this study. In the future, when fitness data for *S. pombe* is more readily available, it would be of scientific interest to apply the network inference methodology and analytical workflow used in this thesis to *S. pombe* fitness data. Doing so would allow parallels and comparisons against the *S. cerevisiae* fitness networks constructed in Chapter 2, as well as determine if the modules identified in the *S. pombe* expression network (Chapter 5) are conserved at the phenotypic level. This approach may also shed further light on the linkage between *FBP1* and riboneogenesis. Additionally, it would be possible to do an integrated analysis such as that described in Chapter 4. *S. pombe* still contains many uncharacterised genes, and an integrated network may aid in predicting gene function (as discussed is section 4.4.2).

9.4.2 The need for experimental validation

The analysis in this thesis focused on bioinformatics, and with the exceptions of the RPL ChIP-chip and UPF1 ChIP-chip data, there has been minimal experimental validation to confirm the hypotheses presented in this thesis. MI based networks are broad and robust, allowing the organisation of a biological system to be visualised, mapped and interrogated. MI based network inference methods can reveal correlations and dependencies between genes, however it cannot determine edge directionality or causality. Due to this limitation, it is essential that any hypotheses developed from MI based networks are validated experimentally. Though this is a substantial limitation to using MI based networks, one key advantage is that they have the ability to identify candidate genes for experimental validation. For example this thesis reported the novel finding that FBP1 in *S. pombe* may be the rate limiting enzyme in riboneogenesis and based on network interactions, several possible experiments were described which could be used to test the hypothesis (Chapter 5, section 5.4.3). In order to increase the robustness of the results presented in this thesis, experimental validation is required, however this thesis has demonstrated how hypothesis driven analysis using reverse engineering methods can identify novel interactions for future research. It also demonstrated that these approaches can provide supporting evidence and additional links between cellular pathways already known to exist.

APPENDIX

Rank	Ensembl ID	Gene name	Degree	Node description
1	*YOR292C*	*YOR292C*	342	Vacuolar membrane protein YOR292C
2	*YPL223C*	*GRE1*	341	Protein GRE1
3	*YDR508C*	*GNP1*	335	High-affinity glutamine permease
4	*YMR098C*	*ATP25*	332	Uncharacterized protein YMR098C
5	*YOR219C*	*ste13*	326	Dipeptidyl aminopeptidase A
6	*YKR034W*	*DAL80*	325	Nitrogen regulatory protein DAL80
7	*YDR217C*	*Rad9*	322	DNA repair protein RAD9
8	*YPL262W*	*FUM1*	321	Fumarate hydratase, mitochondrial
9	*YOR152C*	*YOR152C*	317	Uncharacterized membrane protein YOR152C
10	*YNL187W*	*SWT21*	312	Uncharacterized protein YNL187W
11	*YGR070W*	*Rom1*	310	RHO1 GDP-GTP exchange protein 1
12	*YDR157W*	*YDR157W*	309	Dubious open reading frame
13	*YDR191W*	*HST4*	307	NAD-dependent histone deacetylase HST4
14	*YPR116W*	*RRG8*	304	Uncharacterized protein YPR116W, mitochondrial
15	*YOR159C*	*sme1*	302	Small nuclear ribonucleoprotein E
16	*YMR101C*	*SRT1*	301	Putative dehydrodolichyl diphosphate synthetase
17	*YKL096W-A*	*CWP2*	295	Cell wall protein CWP2
18	*YBL083C*	*YBL083C*	294	Dubious open reading frame overlaps with ALG3
19	*YJR107W*	*YJR107W*	293	Putative lipase YJR107W
20	*YPR139C*	*VPS66*	292	Vacuolar protein sorting-associated protein 66
21	*YGL226C-A*	*OST5*	292	Dolichyl-diphosphooligosaccharide--protein
22	*YOR175C*	*ALE1*	290	Lysophospholipid acyltransferase
23	*YOR049C*	*RSB1*	290	Sphingoid long-chain base transporter RSB1
24	*YDR283C*	*gcn2*	288	Serine/threonine-protein kinase GCN2
25	*YOR378W*	*YOR378W*	287	Drug resistance protein YOR378W
26	*YHR106W*	*TRR2*	286	Thioredoxin reductase 2, mitochondrial
27	*YPL197C*	*YPL197C*	283	Dubious open reading frame, overlaps with RPB7B
28	*YLR088W*	*GAA1*	283	GPI transamidase component GAA1
29	*YDR255C*	*RMD5*	281	Sporulation protein RMD5
30	*YOR136W*	*IDH2*	281	Isocitrate dehydrogenase [NAD] subunit 2, mitochondrial
31	*YOL083W*	*YOL083W*	281	Uncharacterized protein YOL083W
32	*YMR119W*	*ASI1*	280	Protein ASI1
33	*YGL098W*	*Use1*	279	Protein transport protein USE1
34	*YLL007C*	*YLL007C*	279	Uncharacterized protein YLL007C
35	*YPL217C*	*bms1*	278	Ribosome biogenesis protein BMS1
36	*YDL094C*	*YDL094C*	276	Dubious open reading frame, overlaps with PMT5
37	*YAL035C-A*	*YAL035C-A*	276	Dubious open reading frame
38	*YDR472W*	*trs31*	276	Transport protein particle 31 kDa subunit
39	*YNL100W*	*AIM37*	275	Uncharacterized protein YNL100W
40	*YGL073W*	*hsf1*	269	Heat shock factor protein
41	*YKL139W*	*CTK1*	269	CTD kinase subunit alpha
42	*YPL157W*	*tgs1*	267	Trimethylguanosine synthase
43	*YLR124W*	*YLR124W*	266	Dubious open reading frame
44	*YOR133W*	*EFT1*	265	Elongation factor 2
45	*YFR033C*	*QCR6*	265	Cytochrome b-c1 complex subunit 6
46	*YPL073C*	*YPL073C*	265	Dubious open reading frame overlaps with UBP16
47	*YPL035C*	*YPL035C*	263	Dubious open reading frame overlaps with YPL034W
48	*YNL071W*	*LAT1*	263	Dihydrolipoyllysine-residue acetyltransferase
49	*YOL013W-A*	*YOL013W-A*	262	Uncharacterized protein YOL013W-A
50	*YGR257C*	*Mtm1*	262	Mitochondrial carrier protein MTM1

Table A2. 1 Top 50 most connected nodes (hubs) within Hillenmeyer's fitness network

NAME	Description	CNT	NES	FDR q-val	RANK AT MAX
GO:0002181	cytoplasmic translation	92	-4.0	0.000	786
GO:0005839	proteasome core complex	12	-3.5	0.000	512
GO:0003735	structural constituent of ribosome	114	-3.5	0.000	846
GO:0030529	ribonucleoprotein complex	163	-3.5	0.000	846
GO:0004175	endopeptidase activity	11	-3.4	0.000	512
GO:0010499	proteasomal ubiquitin-independent protein catabolic ...	11	-3.4	0.000	512
GO:0004298	threonine-type endopeptidase activity	11	-3.4	0.000	512
GO:0000502	proteasome complex	31	-3.3	0.000	551
GO:0022627	cytosolic small ribosomal subunit	41	-3.2	0.000	812
GO:0001056	RNA polymerase III activity	14	-3.2	0.000	661
GO:0034515	proteasome storage granule	20	-3.2	0.000	551
GO:0042797	tRNA transcription from RNA polymerase III promoter	15	-3.2	0.000	869
GO:0022625	cytosolic large ribosomal subunit	50	-3.1	0.000	882
GO:0003899	DNA-directed RNA polymerase activity	23	-3.1	0.000	869
GO:0005666	DNA-directed RNA polymerase III complex	14	-3.1	0.000	661
GO:0000467	rRNA processing	11	-3.1	0.000	672
GO:0005622	intracellular	158	-3.0	0.000	892
GO:0006412	translation	152	-3.0	0.000	989
GO:0006364	rRNA processing	109	-2.9	0.000	855
GO:0005840	ribosome	151	-2.9	0.000	826
GO:0030686	90S preribosome	46	-2.9	0.000	855
GO:0051603	proteolysis involved in cellular protein catabolic process	13	-2.8	0.000	512
GO:0071051	polyadenylation-dependent snoRNA 3'-end processing	11	-2.7	0.000	883
GO:0003968	RNA-directed RNA polymerase activity	10	-2.6	0.000	868
GO:0071035	nuclear polyadenylation-dependent rRNA ...	10	-2.6	0.000	883
GO:0071042	nuclear polyadenylation-dependent mRNA ...	10	-2.6	0.000	883
GO:0005665	DNA-directed RNA polymerase II, core complex	10	-2.6	0.000	868
GO:0030515	snoRNA binding	12	-2.6	0.000	1006
GO:0043161	proteasomal ubiquitin-	22	-2.6	0.000	524

	dependent protein ...				
GO:0071038	nuclear polyadenylation-dependent tRNA ...	10	-2.5	0.001	883
GO:0000463	maturation of LSU-rRNA from ...	14	-2.4	0.002	1016
GO:0006407	rRNA export from nucleus	10	-2.3	0.004	797
GO:0000176	nuclear exosome (RNase complex)	11	-2.2	0.005	672
GO:0070478	nuclear-transcribed mRNA catabolic process, 3'-5 ...	13	-2.2	0.008	902
GO:0005730	nucleolus	133	-2.1	0.009	490
GO:0016586	RSC complex	12	-2.1	0.009	836
GO:0000462	maturation of SSU-rRNA from tricistronic ...	40	-2.1	0.009	1066
GO:0034427	nuclear-transcribed mRNA catabolic process ...	10	-2.1	0.009	902
GO:0006337	nucleosome disassembly	12	-2.1	0.011	836
GO:0042254	ribosome biogenesis	106	-2.1	0.011	855
GO:0043044	ATP-dependent chromatin remodeling	11	-2.0	0.018	836
GO:0035091	phosphatidylinositol binding	14	-2.0	0.021	1311
GO:0070651	nonfunctional rRNA decay	11	-1.9	0.026	849
GO:0032040	small-subunit processome	35	-1.8	0.046	833
GO:0030687	preribosome, large subunit precursor	31	-1.8	0.053	379
GO:0019843	rRNA binding	21	-1.8	0.054	807
GO:0015616	DNA translocase activity	10	-1.8	0.065	836
GO:0006457	protein folding	43	-1.7	0.073	213
GO:0006368	transcription elongation from RNA ...	26	-1.7	0.095	1049

Table A2. 2 GSEAPreranked -the top 50 GO terms with negative significant enrichment using node radality

FDR q-values are coloured by significance. Red ≤ 0.05,_green ≤ 0.1.CNT represents for the number of genes contained with the GO term. NES stands for normalised enrichment score, and is a statistic used for examining gene enrichment results. FDR q-val is the estimated probability of a false positive and RANK is the location with ranked list

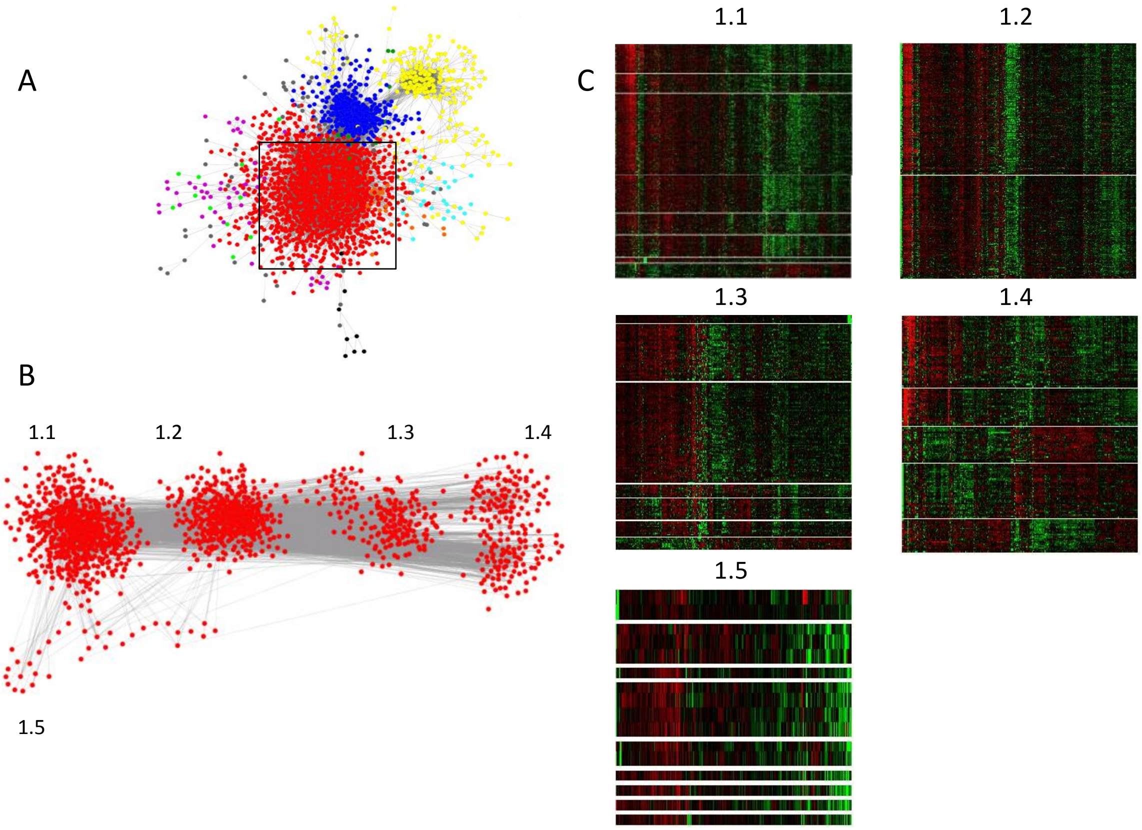


Figure A2. 1 HOPACH on Hillenmeyer Module 1

Panel A A force directed layout of the parent network with the GLay modules mapped on. Module 1 is highlighted in red, with the black box encapsulating it. Panel B A further level off modularisation, identifying five sub-modules. Panel C HOPACH clustering identifying the fitness profiles present within each Glay sub-module. White line breaks indicate HOPACH cluster.

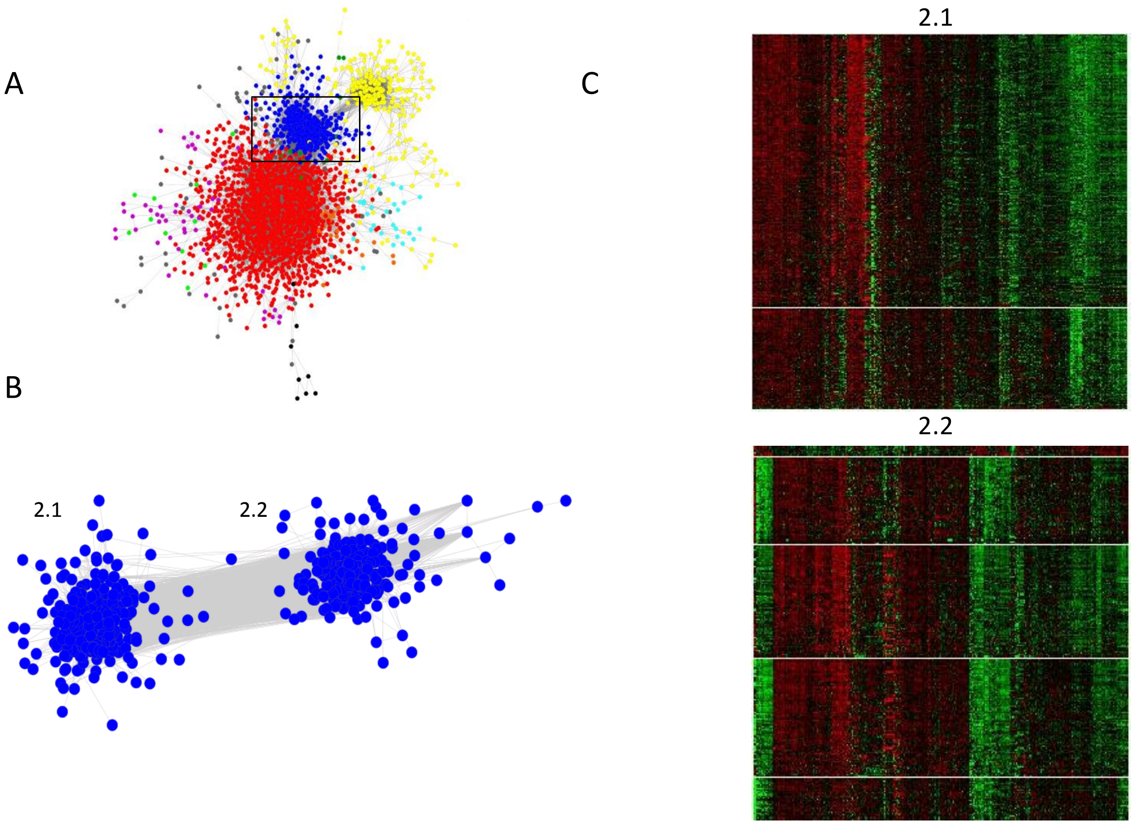


Figure A2. 2 HOPACH on Hillenmeyer Module 2

Panel A A force directed layout of the parent network with the GLay modules mapped on. Module 2 is highlighted in blue, with the black box encapsulating it. Panel B A further level off modularisation, identifying two sub-modules. Panel C HOPACH clustering identifying the fitness profiles present within each Glay sub-module. White line breaks indicate HOPACH cluster.

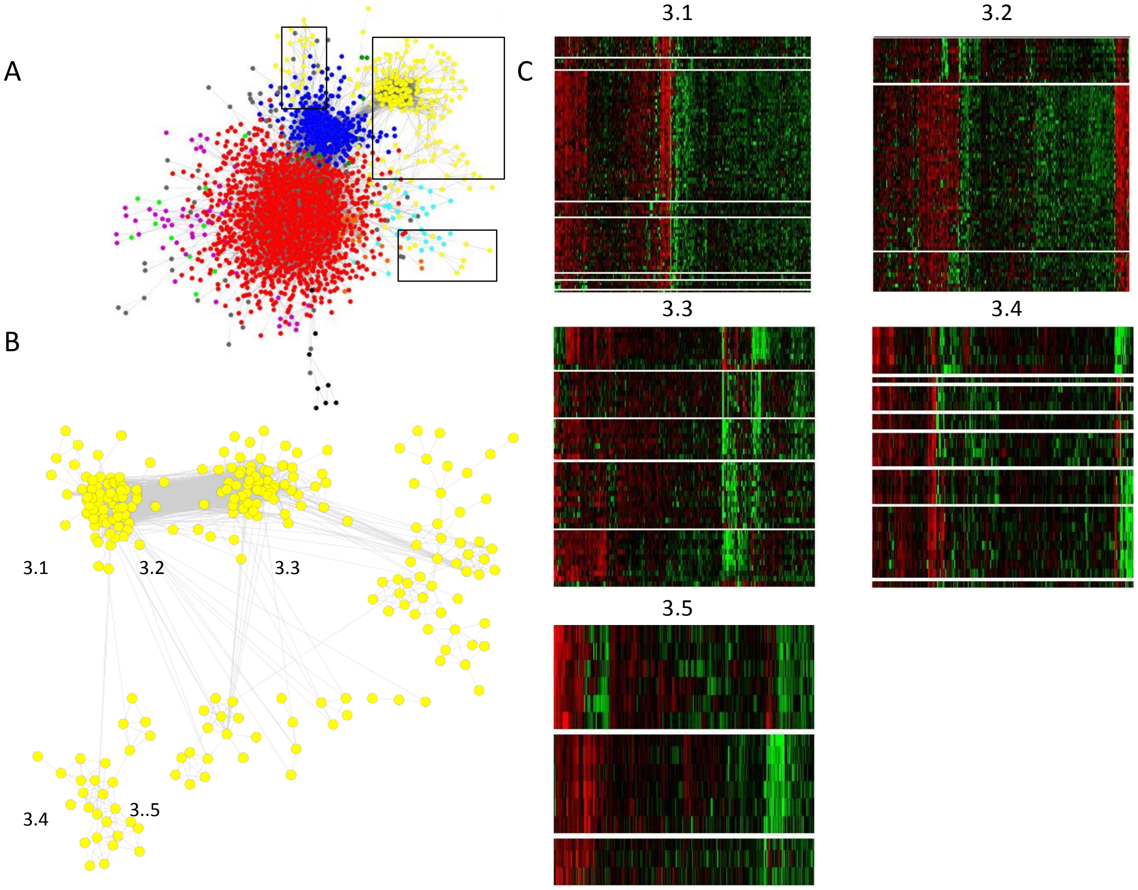


Figure A2. 3 HOPACH on Hillenmeyer Module 3

Panel A A force directed layout of the parent network with the GLay modules mapped on. Module 3 is highlighted in yellow, with the black box encapsulating it. Panel B A further level off modularisation, identifying five sub-modules. Panel C HOPACH clustering identifying the fitness profiles present within each Glay sub-module. White line breaks indicate HOPACH cluster.

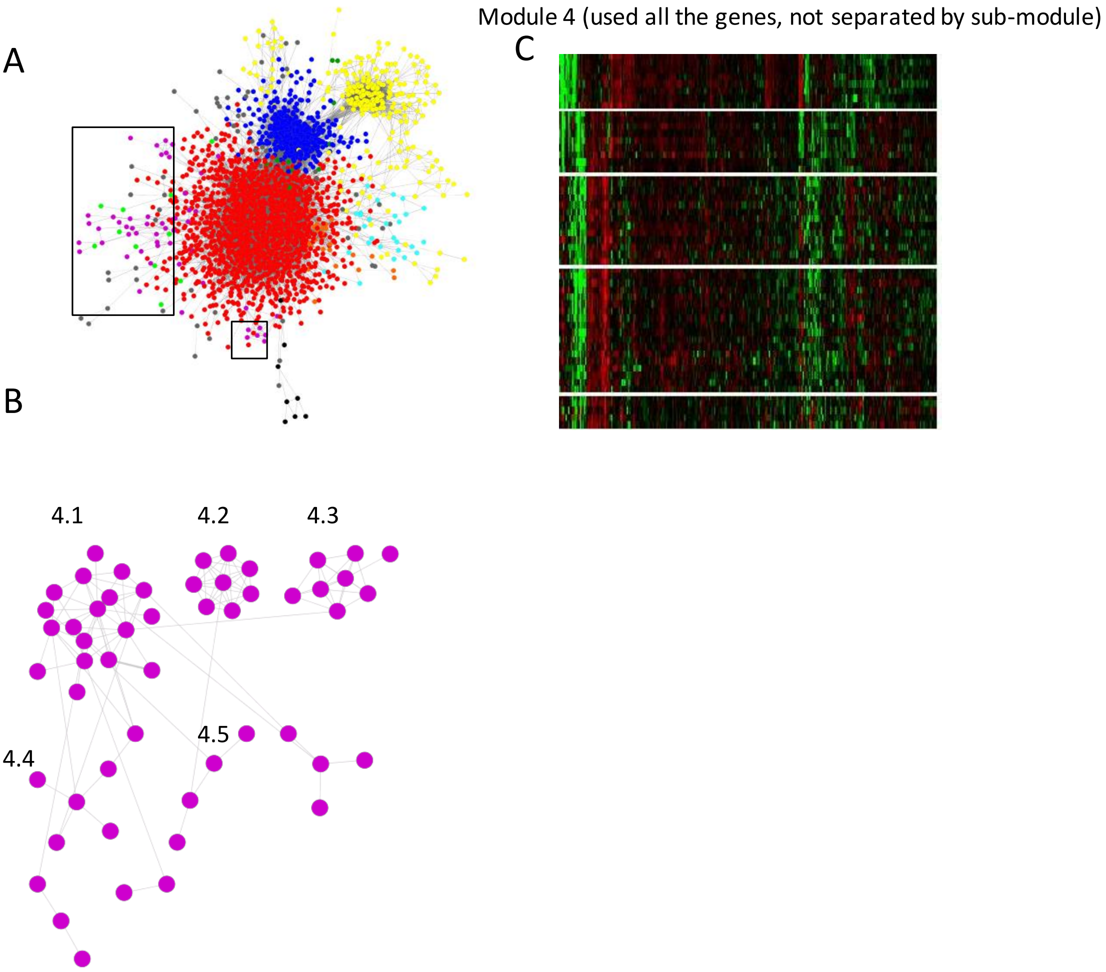


Figure A2. 4 HOPACH on Hillenmeyer Module 4

Panel A A force directed layout of the parent network with the GLay modules mapped on. Module 4 is highlighted in purple, with the black box encapsulating it. Panel B A further level off modularisation, identifying five sub-modules. Panel C HOPACH clustering identifying the fitness profiles present within each Glay sub-module. White line breaks indicate HOPACH cluster.

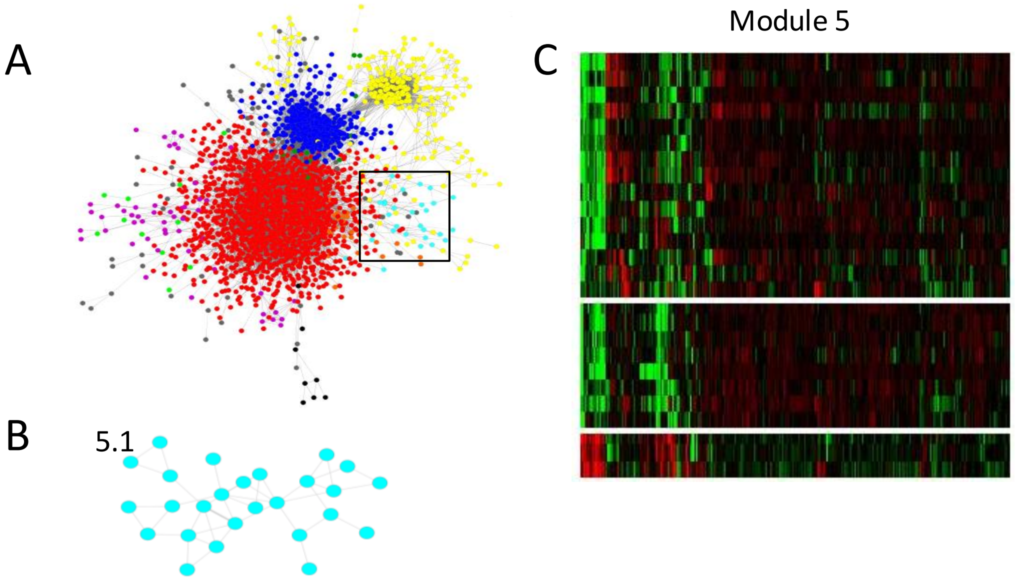


Figure A2. 5 HOPACH on Hillenmeyer Module 5
Panel A A force directed layout of the parent network with the GLay modules mapped on. Module 1 is highlighted in red, with the black box encapsulating it. Panel B The structure of module 5 Panel C HOPACH clustering identifying the fitness profiles present within the module. White line breaks indicate HOPACH cluster.

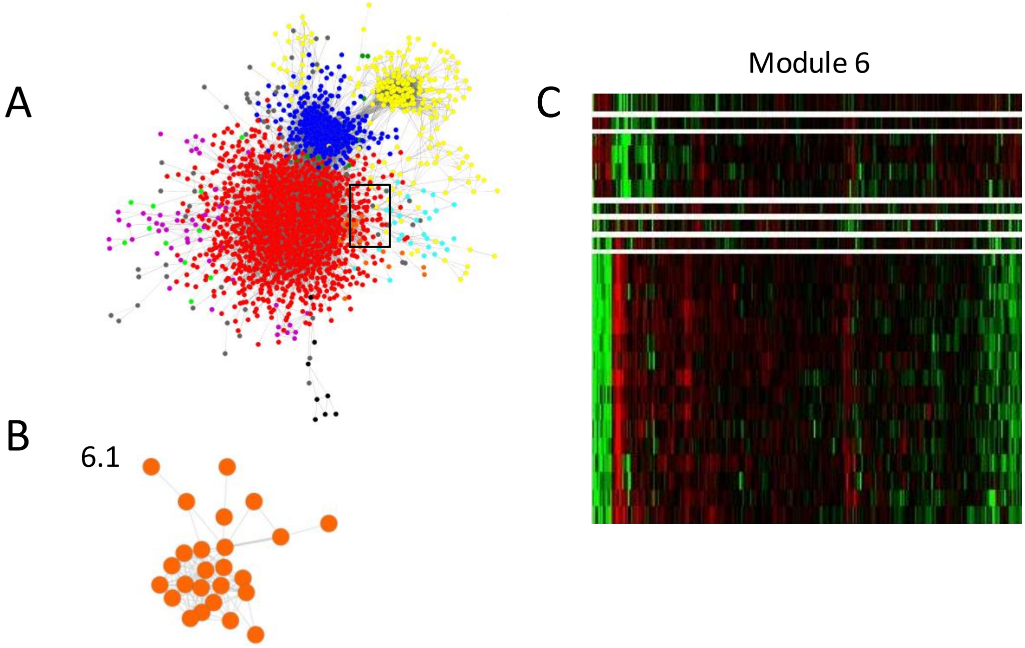


Figure A2. 6 HOPACH on Hillenmeyer Module 6
Panel A A force directed layout of the parent network with the GLay modules mapped on. Module61 is highlighted in orange, with the black box encapsulating it. Panel B The structure of module 5 Panel C HOPACH clustering identifying the fitness profiles present within the module. White line breaks indicate HOPACH cluster.

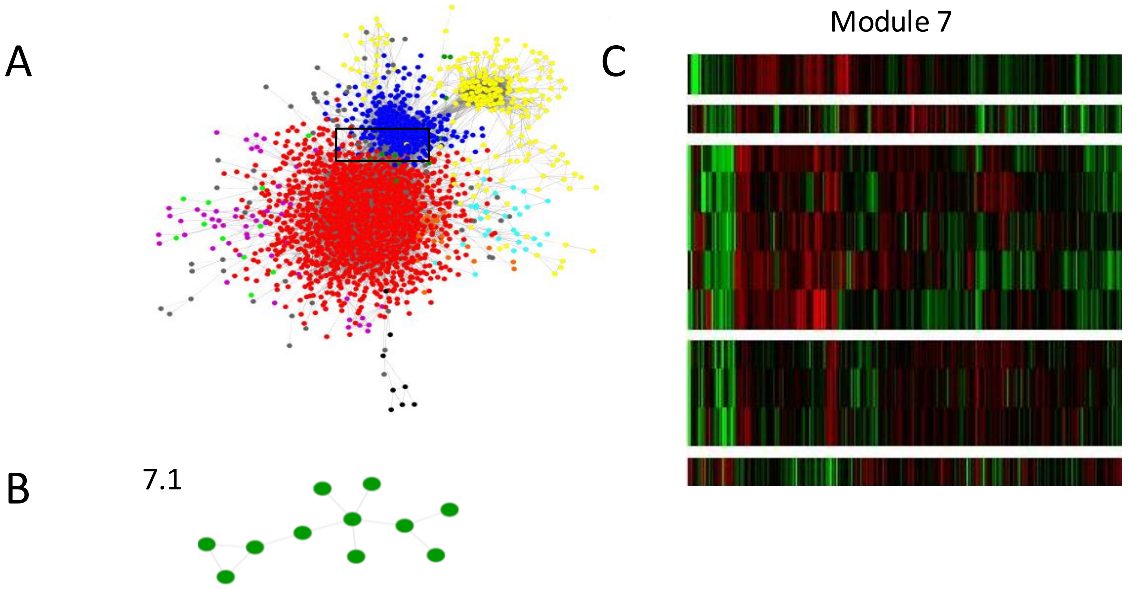


Figure A2. 7 HOPACH on Hillenmeyer Module 7
Panel A A force directed layout of the parent network with the GLay modules mapped on. Module 7 is highlighted in dark green, with the black box encapsulating it. Panel B The structure of module 5 Panel C HOPACH clustering identifying the fitness profiles present within the module. White line breaks indicate HOPACH cluster.

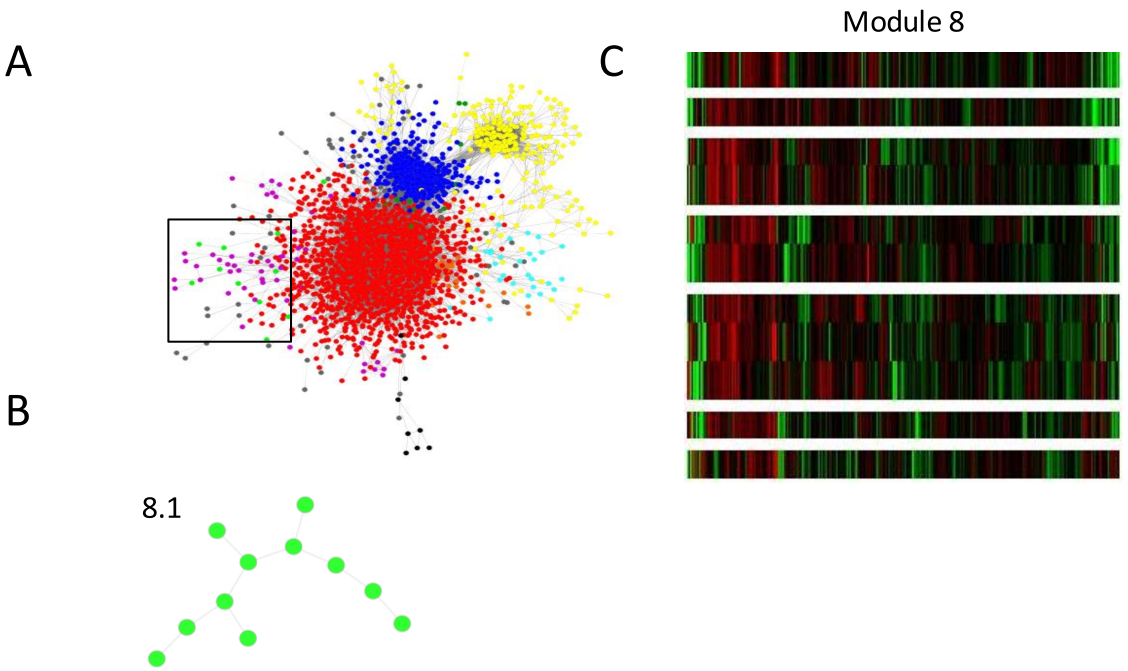


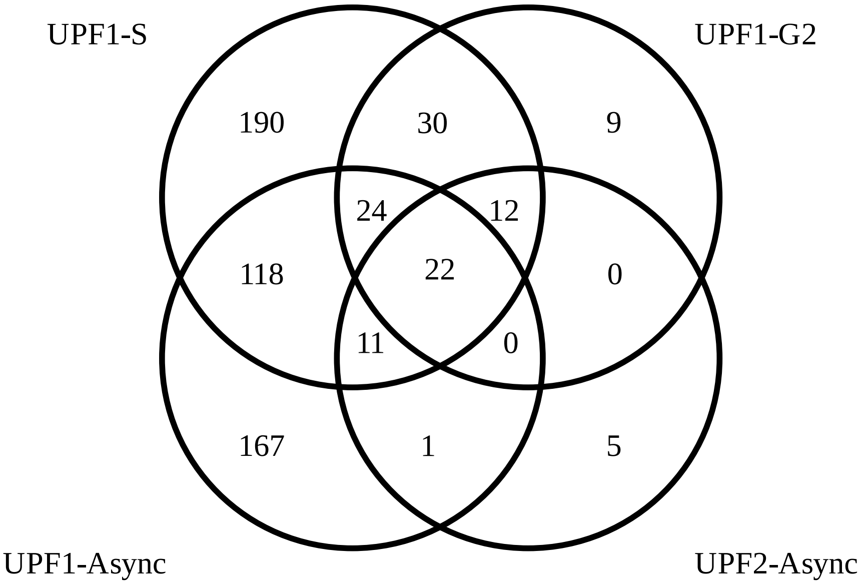
Figure A2. 8 HOPACH on Hillenmeyer Module 8
Panel A A force directed layout of the parent network with the GLay modules mapped on. Module 8 is highlighted in light, with the black box encapsulating it. Panel B The structure of module 5 Panel C HOPACH clustering identifying the fitness profiles present within the module. White line breaks indicate HOPACH cluster.

Sequences producing significant alignment	**Bits score**	**E score**
ref|NP_593140.1| phosphoglycerate mutase family (predicted)	46.6	8e-07
ref|NP_594889.1| monomeric 2,3-bisphosphoglycerate (BPG) - dependent	46.2	1e-06
ref|NP_588471.1| phosphoglycerate mutase family (predicted)	37.7	0.001
sp|P41389.2|MCM5_SCHPO RecName: Full=DNA replication	28.1	1.5
ref|XP_001713071.1| MCM complex subunit Mcm5	28.1	1.5
gb|AAC60568.1| budding yeast CDC46 homolog	28.1	1.6
ref|NP_595279.1| coatomer alpha subunit (predicted)	26.6	4.6
ref|NP_593750.1| central kinetochore associated family protein	25.8	7.1
ref|NP_588313.2| ER protein folding oxidoreductin Ero1b	25.8	7.6
ref|NP_594941.1| 2 OG-Fe(II) oxygenase superfamily protein	25.4	8.8

Table A5. 1 The results of BLASTP: *S. cerevisiae SHB17* against the *S. pombe* genome
Column 1 shows the most significant sequences, column 2 shows the bit score, and column 3 shows the E score. The E score is a significance statistic, the closer to zero, the more significant the result. The results show that *S. pombe* FBP1 is not identified as a significant hit to *S. cerevisiae SHB17*.

Sequences producing significant alignment	**Bits score**	**E score**
dbj|GAA25236.1| K7_Fbp1p [Saccharomyces cerevisiae]	428	9e-149
dbj|GAA25236.1| K7_Fbp1p [Saccharomyces cerevisiae]	426	1e-148
dbj|GAA25236.1| K7_Fbp1p [Saccharomyces cerevisiae]	427	2e-148
gb|EIW07468.1| She4p [Saccharomyces cerevisiae]	217	3e-68
pdb|3OPB|A Chain A, Crystal Structure Of She4p	30.4	3.3
gb|AAC60568.1| budding yeast CDC46 homolog	30.4	3.6
ref|NP_595279.1| coatomer alpha subunit (predicted)	28.9	9.6

Table A5. 2 The results of BLASTP: *S. pombe FBP1* against the *S. cerevisiae* genome
Column 1 shows the most significant sequences, column 2 shows the bit score, and column 3 shows the E score. The E score is a significance statistic, the closer to zero, the more significant the result. The results show that *S. pombe* FBP1 is significantly similar to *S. cerevisiae FBP1*, but not *SHB17*.



Only in UPF1-S & UPF2-Async: 5
Only in UPF1-G2 & UPF1-Async: 3

Figure A7. 1 The overlap of statistically significant genomic regions bound between UPF1 S-phase, UPF1 G2-phase, UPF1-Async and UPF2-Async.

Systematic Name	Description	Feature
SPCC1223.01	ubiquitin-protein ligase E3 (predicted)	protein_coding
SPCC1223.02	no message in thiamine Nmt1	protein_coding
SPBCPT2R1.08c	RecQ type DNA helicase Tlh1	protein_coding
SPAC30D11.13	SUMO conjugating enzyme Hus5	protein_coding
SPATRNAALA.04	tRNA Alanine	tRNA
SPATRNAALA.05	tRNA Alanine	tRNA
SPCTRNAARG.12	tRNA Arginine	tRNA
SPCTRNAASP.06	tRNA Asparagine	tRNA
SPCTRNAASP.07	tRNA Asparagine	tRNA
SPCTRNAARG.13	tRNA Arginine	tRNA
SPATRNAGLU.03	tRNA Glutamic acid	tRNA
SPATRNAGLU.04	tRNA Glutamic acid	tRNA
SPCTRNATHR.08	tRNA Threonine	tRNA
SPCTRNATHR.09	tRNA Threonine	tRNA
SPBTRNAVAL.05	tRNA Valine	tRNA
SPBTRNAVAL.06	tRNA Valine	tRNA
SPBTRNAVAL.07	tRNA Valine	tRNA
SPCTRNAVAL.09	tRNA Valine	tRNA
SPCTRNAVAL.10	tRNA Valine	tRNA
SPNCRNA.10	antisense RNA (predicted)	ncRNA
SPNCRNA.484	non-coding RNA, centromeric (predicted)	ncRNA
SPNCRNA.483	non-coding RNA, centromeric (predicted)	ncRNA

Table A7. 1 Details of the 22 gene overlap between the UPF1 ChIP-chip and UPF2 (shown in Figure A7.1)

LIST OF REFERENCES

[1] P. Kohl, EJ Crampin, TA Quinn, and D. Noble. Systems biology: an approach. *Clinical Pharmacology & Therapeutics*, 88(1):25–33, 2010.

[2] Weiwen Zhang, Feng Li, and Lei Nie. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology*, 156(2):287–301, 2010.

[3] R. Mustacchi, S. Hohmann, and J. Nielsen. Yeast systems biology to unravel the network of life. *Yeast*, 23(3):227–238, 2006.

[4] L.V. Zhang, O.D. King, S.L. Wong, D.S. Goldberg, A.H.Y. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, and F.P. Roth. of biology. *Journal of Biology*, 4:6, 2005.

[5] John Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32:496–501, 2002.

[6] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[7] Zhijin Wu and Rafael A Irizarry. Preprocessing of oligonucleotide array data. *Nature Biotechnology*, 22(6):656–658, 2004.

[8] Zhijin Wu, Rafael A Irizarry, Robert Gentleman, Francisco Martinez Murillo, and Forrest Spencer. A model based background adjustment for oligonucleotide expression arrays. 2004.

[9] Cheng Li and Wing Hung Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36, 2001.

[10] Laurent Gautier, Leslie Cope, Benjamin M Bolstad, and Rafael A Irizarry. affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–315, 2004.

[11] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[12] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987.

[13] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000.

[14] Javier Herrero, Alfonso Valencia, and Joaqun Dopazo. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2):126–136, 2001.

[15] Mark J van der Laan and Katherine S Pollard. A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. *Journal of Statistical Planning and Inference*, 117(2):275–303, 2003.

[16] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[17] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[18] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

[19] Gerhard Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386, 1988.

[20] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[21] B.T.S. Da Wei Huang, R.A. Lempicki, et al. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2008.

[22] B.T. Sherman, R.A. Lempicki, et al. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[23] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[24] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[25] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

[26] Fernando Ortega, Katrin Sameith, Nil Turan, Russell Compton, Victor Trevino, Marina Vannucci, and Francesco Falciani. Models and computational strategies linking physiological response to molecular networks from large-scale data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1878):3067–3089, 2008.

[27] Jing Yu, V Anne Smith, Paul P Wang, Alexander J Hartemink, and Erich D Jarvis. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594–3603, 2004.

[28] Patrik D'haeseleer, Shoudan Liang, and Roland Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707–726, 2000.

[29] Mukesh Bansal, Giusy Della Gatta, and Diego Di Bernardo. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, 22(7):815–822, 2006.

[30] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R.D. Favera, and A. Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7, 2006.

[31] Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1):e8, 2007.

[32] K. Basso, A.A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human b cells. *Nature Genetics*, 37(4):382–390, 2005.

[33] G. Su, A. Kuchinsky, J.H. Morris, F. Meng, et al. Glay: community structure analysis of biological networks. *Bioinformatics*, 26(24):3135–3137, 2010.

[34] Xiaoli Li, Min Wu, Chee-Keong Kwoh, and See-Kiong Ng. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics*, 11(Suppl 1):S3, 2010.

[35] Ioannis A Maraziotis, Konstantina Dimitrakopoulou, and Anastasios Bezerianos. Growing functional modules from a seed protein via integration of protein interaction and gene expression data. *BMC Bioinformatics*, 8(1):408, 2007.

[36] Elizabeth A Winzeler, Daniel D Shoemaker, Anna Astromoff, Hong Liang, Keith Anderson, Bruno Andre, Rhonda Bangham, Rocio Benito, Jef D Boeke, Howard Bussey, et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, 285(5429):901–906, 1999.

[37] Guri Giaever, Angela M Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Veronneau, Sally Dow, Ankuta Lucau-Danila, Keith Anderson, Bruno Andre, et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896):387–391, 2002.

[38] Victoria Smith, David Botstein, and Patrick O Brown. Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proceedings of the National Academy of Sciences*, 92(14):6479–6483, 1995.

[39] Petra Ross-Macdonald, Amy Sheehan, G Shirleen Roeder, and Michael Snyder. A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, 94(1):190–195, 1997.

[40] M.E. Hillenmeyer, E. Fung, J. Wildenhain, S.E. Pierce, S. Hoon, W. Lee, M. Proctor, R.P.S. Onge, M. Tyers, D. Koller, et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, 320(5874):362–365, 2008.

[41] Frank Baganz, Andrew Hayes, Ronnie Farquhar, Philip R Butler, David CJ Gardner, and Stephen G Oliver. Quantitative analysis of yeast gene function using competition experiments in continuous culture. *Yeast*, 14(15):1417–1427, 1998.

[42] Zhijian Li, Franco J Vizeacoumar, Sondra Bahr, Jingjing Li, Jonas Warringer, Frederick S Vizeacoumar, Renqiang Min, Benjamin VanderSluis, Jeremy Bellay, Michael DeVit, et al. Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nature Biotechnology*, 29(4):361–367, 2011.

[43] Kaspar Burger, Bastian Mühl, Thomas Harasim, Michaela Rohrmoser, Anastassia Malamoussi, Mathias Orban, Markus Kellner, Anita Gruber-Eber, Elisabeth Kremmer, Michael Hölzel, et al. Chemotherapeutic drugs inhibit ribosome biogenesis at various levels. *Journal of Biological Chemistry*, 285(16):12416–12425, 2010.

[44] Michael J Buck and Jason D Lieb. Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, 2004.

[45] Philippe Collas. The current state of chromatin immunoprecipitation. *Molecular Biotechnology*, 45(1):87–100, 2010.

[46] Siavash K Kurdistani, Daniel Robyr, Saeed Tavazoie, and Michael Grunstein. Genome-wide binding map of the histone deacetylase rpd3 in yeast. *Nature Genetics*, 31(3):248–254, 2002.

[47] Vishwanath R Iyer, Christine E Horak, Charles S Scafe, David Botstein, Michael Snyder, and Patrick O Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409(6819):533–538, 2001.

[48] Bing Ren, François Robert, John J Wyrick, Oscar Aparicio, Ezra G Jennings, Itamar Simon, Julia Zeitlinger, Jörg Schreiber, Nancy Hannett, Elenita Kanin, et al. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309, 2000.

[49] Daniel Schaarschmidt, Jens Baltin, Isa M Stehle, Hans J Lipps, and Rolf Knippers. An episomal mammalian replicon: sequence-independent binding of the origin recognition complex. *The EMBO Journal*, 23(1):191–201, 2003.

[50] S. De, W. Varsally, F. Falciani, and S. Brogna. Ribosomal proteins' association with transcription sites peaks at tRNA genes in *Schizosaccharomyces pombe*. *RNA*, 17(9):1713–1726, 2011.

[51] Hongkai Ji and Wing Hung Wong. Tilemap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, 21(18):3629–3636, 2005.

[52] W Evan Johnson, Wei Li, Clifford A Meyer, Raphael Gottardo, Jason S Carroll, Myles Brown, and X Shirley Liu. Model-based analysis of tiling-arrays for chip-chip. *Proceedings of the National Academy of Sciences*, 103(33):12457–12462, 2006.

[53] Wei Li, Clifford A Meyer, and X Shirley Liu. A hidden markov model for analyzing chip-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, 21(suppl 1):i274–i282, 2005.

[54] AEA Goffeau, R Aert, ML Agostini-Carbone, A Ahmed, M Aigle, L Alberghina, K Albermann, M[c Albers, M Aldea, D Alexandraki, et al. The yeast genome directory. *Nature*, 387(6632):5–6, 1997.

[55] HW Mewes, K Albermann, M Bähr, D Frishman, A Gleissner, J Hani, K Heumann, K Kleine, A Maierl, SG Oliver, et al. Overview of the yeast genome. *Nature*, 387(6632):7–8, 1997.

[56] J Michael Cherry, Eurie L Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T Chan, Karen R Christie, Maria C Costanzo, Selina S Dwight, Stacia R Engel, et al.

Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1):D700–D705, 2012.

[57] V Wood, R Gwilliam, M-A Rajandream, M Lyne, R Lyne, A Stewart, J Sgouros, N Peat, J Hayles, S Baker, et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature*, 415(6874):871–880, 2002.

[58] Valerie Wood, Midori A Harris, Mark D McDowall, Kim Rutherford, Brendan W Vaughan, Daniel M Staines, Martin Aslett, Antonia Lock, Jürg Bähler, Paul J Kersey, et al. Pombase: a comprehensive online resource for fission yeast. *Nucleic Acids Research*, 40(D1):D695–D699, 2012.

[59] A.R. Borneman, P.J. Chambers, and I.S. Pretorius. Yeast systems biology: modelling the winemaker's art. *Trends in Biotechnology*, 25(8):349–355, 2007.

[60] David Beach, Barbara Durkacz, and Paul Nurse. Functionally homologous cell cycle control genes in budding and fission yeast. *Nature,* 300(5894):706–709, 1982.

[61] Leland H Hartwell. Cell division from a genetic perspective. *The Journal of Cell Biology*, 77(3):627–637, 1978.

[62] I. Herskowitz. Life cycle of the budding yeast *Saccharomyces cerevisiae*. *Microbiological Reviews*, 52(4):536, 1988.

[63] Paul Russell and Paul Nurse. *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*: a look at yeasts divided. *Cell*, 45(6):781–782, 1986.

[64] Matthias Sipiczki. Where does fission yeast sit on the tree of life. *Genome Biol*, 1(2):1–4, 2000.

[65] Leslie Grate and Manuel Ares Jr. Searching yeast intron data at ares lab web site. *Methods in Enzymology*, 350:380–392, 2002.

[66] Nicholas Rhind, Zehua Chen, Moran Yassour, Dawn A Thompson, Brian J Haas, Naomi Habib, Ilan Wapinski, Sushmita Roy, Michael F Lin, David I Heiman, et al. Comparative functional genomics of the fission yeasts. *Science*, 332(6032):930–936, 2011.

[67] T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, et al. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

[68] Z. Hu, P.J. Killion, and V.R. Iyer. Genetic reconstruction of a functional transcriptional regulatory network. *Nature Genetics*, 39(5):683–687, 2007.

[69] Matthew E Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes, Dileepa Diyagama, Andrew Holloway, and Gordon K Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23(20):2700–2707, 2007.

[70] Robert Nadon and Jennifer Shoemaker. Statistical issues with microarrays: processing and analysis. *Trends in Genetics*, 18(5):265–271, 2002.

[71] J. Reimand, J.M. Vaquerizas, A.E. Todd, J. Vilo, and N.M. Luscombe. Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Research*, 38(14):4768–4777, 2010.

[72] V. Pancaldi, F. Schubert, and J. Bähler. Meta-analysis of genome regulation and expression variability across hundreds of environmental and genetic perturbations in fission yeast. *Molecular BioSystems*, 6(3):543–552, 2010.

[73] Dong-Uk Kim, Jacqueline Hayles, Dongsup Kim, Valerie Wood, Han-Oh Park, Misun Won, Hyang-Sook Yoo, Trevor Duhig, Miyoung Nam, Georgia Palmer, et al. Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. *Nature Biotechnology*, 28(6):617–623, 2010.

[74] Tian Xu Han, Xing-Ya Xu, Mei-Jun Zhang, Xu Peng, and Li-Lin Du. Method global fitness profiling of fission yeast deletion strains by barcode sequencing. 2010.

[75] Michael S Samoilov, Xiaodong Wang, Adam P Arkin, and Liming Wang. Inference of gene regulatory networks from genome-wide knockout fitness data. *Bioinformatics*, 29(3):338–346, 2012.

[76] Thiago M Venancio, S Balaji, Lakshminarayan M Iyer, L Aravind, et al. Reconstructing the ubiquitin network: cross-talk with other systems and identification of novel functions. *Genome Biol*, 10(3):R33, 2009.

[77] Anne-Claude Gavin, Markus Bösche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jörg Schultz, Jens M Rick, Anne-Marie Michon, Cristina-Maria Cruciat, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[78] Michael P Washburn, Dirk Wolters, and John R Yates. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, 19(3):242–247, 2001.

[79] Paola Picotti, Bernd Bodenmiller, Lukas N Mueller, Bruno Domon, and Ruedi Aebersold. Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. *Cell*, 138(4):795–806, 2009.

[80] Charles W Schmidt. Metabolomics: what’s happening downstream of DNA. *Environmental Health Perspectives*, 112(7):A410, 2004.

[81] Leonie M Raamsdonk, Bas Teusink, David Broadhurst, Nianshu Zhang, Andrew Hayes, Michael C Walsh, Jan A Berden, Kevin M Brindle, Douglas B Kell, Jem J Rowland, et al. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, 19(1):45–50, 2001.

[82] Andrew R Joyce and Bernhard Ø Palsson. The model organism as a system: integrating’omics’ data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210, 2006.

[83] Mark R Viant. Metabolomics of aquatic organisms: the new’omics’ on the block. *Marine Ecology Progress Series*, 332:301–306, 2007.

[84] Oliver Fiehn. Metabolomics–the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2):155–171, 2002.

[85] Juan I Castrillo, Andrew Hayes, Shabaz Mohammed, Simon J Gaskell, and Stephen G Oliver. An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. *Phytochemistry*, 62(6):929–937, 2003.

[86] Steve O'Hagan, Warwick B Dunn, Marie Brown, Joshua D Knowles, and Douglas B Kell. Closed-loop, multiobjective optimization of analytical instrumentation: gas chromatography/time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. *Analytical Chemistry*, 77(1):290–303, 2005.

[87] Warwick B Dunn, Nigel JC Bailey, and Helen E Johnson. Measuring the metabolome: current analytical technologies. *Analyst*, 130(5):606–625, 2005.

[88] Trey Ideker, Vesteinn Thorsson, Jeffrey A Ranish, Rowan Christmas, Jeremy Buhler, Jimmy K Eng, Roger Bumgarner, David R Goodlett, Ruedi Aebersold, and Leroy Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934, 2001.

[89] Esti Yeger-Lotem, Shmuel Sattath, Nadav Kashtan, Shalev Itzkovitz, Ron Milo, Ron Y Pinter, Uri Alon, and Hanah Margalit. Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):5934–5939, 2004.

[90] Susanne Prinz, Iliana Avila-Campillo, Christine Aldridge, Ajitha Srinivasan, Krassen Dimitrov, Andrew F Siegel, and Timothy Galitski. Control of yeast filamentous-form growth by modules in an integrated molecular network. *Genome Research*, 14(3):380–390, 2004.

[91] Amos Tanay, Roded Sharan, Martin Kupiec, and Ron Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2981–2986, 2004.

[92] Michelle F Clasquin, Eugene Melamud, Alexander Singer, Jessica R Gooding, Xiaohui Xu, Aiping Dong, Hong Cui, Shawn R Campagna, Alexei Savchenko, Alexander F Yakunin, et al. Riboneogenesis in yeast. *Cell*, 145(6):969–980, 2011.

[93] Ekaterina Kuznetsova, Linda Xu, Alexander Singer, Greg Brown, Aiping Dong, Robert Flick, Hong Cui, Marianne Cuff, Andrzej Joachimiak, Alexei Savchenko, et al. Structure and activity of the metal-independent fructose-1, 6-bisphosphatase yk23 from *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, 285(27):21049–21059, 2010.

[94] Benjamin P Tu, Andrzej Kudlicki, Maga Rowicka, and Steven L McKnight. Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science*, 310(5751):1152–1158, 2005.

[95] Sander Granneman and Susan J Baserga. Ribosome biogenesis: of knobs and RNA processing. *Experimental Cell Research*, 296(1):43–50, 2004.

[96] O-Yu Kwon, Kimihiro Ogino, and Hajime Ishikawa. The longest 18s ribosomal RNA ever known. *European Journal of Biochemistry*, 202(3):827–833, 1991.

[97] Jonathan R Warner. The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences*, 24(11):437–440, 1999.

[98] Mikael S Lindström. Emerging functions of ribosomal proteins in gene-specific transcription and translation. *Biochemical and Biophysical Research Communications*, 379(2):167–170, 2009.

[99] Sandip De and Saverio Brogna. Are ribosomal proteins present at transcription sites on or off ribosomal subunits? *Biochemical Society Transactions*, 38(6):1543, 2010.

[100] Ira G Wool. Extraribosomal functions of ribosomal proteins. *Trends in Biochemical Sciences*, 21(5):164–165, 1996.

[101] Jonathan R Warner and Kerri B McIntosh. How common are extraribosomal functions of ribosomal proteins? *Molecular Cell*, 34(1):3–11, 2009.

[102] S. Komili, N.G. Farny, F.P. Roth, and P.A. Silver. Functional specificity among ribosomal proteins regulates gene expression. *Cell*, 131(3):557–571, 2007.

[103] Rudi J Planta. Regulation of ribosome synthesis in yeast. *Yeast*, 13(16):1505–1518, 1997.

[104] Robert R Klevecz, James Bolen, Gerald Forrest, and Douglas B Murray. A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proceedings of the National Academy of Sciences of the United States of America*, 101(5):1200–1205, 2004.

[105] Million Tadege and Cris Kuhlemeier. Aerobic fermentation during tobacco pollen development. *Plant Molecular Biology*, 35(3):343–354, 1997.

[106] Hans Reinke and David Gatfield. Genome-wide oscillation of transcription in yeast. *Trends in Biochemical Sciences*, 31(4):189–191, 2006.

[107] Anna Oliva, Adam Rosebrock, Francisco Ferrezuelo, Saumyadipta Pyne, Haiying Chen, Steve Skiena, Bruce Futcher, and Janet Leatherwood. The cell cycle–regulated genes of *Schizosaccharomyces pombe*. *PLoS Biology*, 3(7):e225, 2005.

[108] Bruce Futcher. Metabolic cycle, cell cycle, and the finishing kick to start. *Genome Biology*, 7(4):107, 2006.

[109] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, N. Barkai, et al. Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31(4):370–378, 2002.

[110] Sarah E Pierce, Eula L Fung, Daniel F Jaramillo, Angela M Chu, Ronald W Davis, Corey Nislow, and Guri Giaever. A unique and universal molecular barcode array. *Nature Methods*, 3(8):601–603, 2006.

[111] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

[112] Yassen Assenov, Fidel Ramrez, Sven-Eric Schelhorn, Thomas Lengauer, and Mario Albrecht. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284, 2008.

[113] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[114] Oscar M Aparicio, Deborah M Weinstein, and Stephen P Bell. Components and dynamics of DNA replication complexes in *S. cerevisiae*: redistribution of mcm proteins and cdc45p during s phase. *Cell*, 91(1):59–69, 1997.

[115] Agnieszka Chacinska, Carla M Koehler, Dusanka Milenkovic, Trevor Lithgow, and Nikolaus Pfanner. Importing mitochondrial proteins: machineries and mechanisms. *Cell*, 138(4):628–644, 2009.

[116] Thomas Lutz, Walter Neupert, and Johannes M Herrmann. Import of small tim proteins into the mitochondrial intermembrane space. *The EMBO Journal*, 22(17):4400, 2003.

[117] Piotr Bragoszewski, Agnieszka Gornicka, Malgorzata E Sztolsztener, and Agnieszka Chacinska. The ubiquitin-proteasome system regulates mitochondrial intermembrane space proteins. *Molecular and Cellular Biology*, 2013.

[118] Thorsten Schäfer, Daniela Strauß, Elisabeth Petfalski, David Tollervey, et al. The path from nucleolar 90s to cytoplasmic 40s pre-ribosomes. *The EMBO Journal*, 22(6):1370–1380, 2003.

[119] Alice Y Dunn, Mark W Melville, and Judith Frydman. Review: cellular substrates of the eukaryotic chaperonin tric/cct. *Journal of Structural Biology*, 135(2):176–184, 2001.

[120] Rajani Srikakulam and Donald A Winkelmann. Myosin ii folding is mediated by a molecular chaperonin. *Journal of Biological Chemistry*, 274(38):27265–27273, 1999.

[121] Wolfgang Hilt, Wolfgang Heinemeyer, and Dieter H Wolf. The proteasome and protein degradation in yeast. In *Intracellular Protein Catabolism*, pages 197–202. Springer, 1996.

[122] Masami Horikoshi, C Kathy Wang, Hiroshi Fujii, James A Cromlish, P Anthony Weil, and Robert G Roeder. Cloning and structure of a yeast gene encoding a general transcription initiation factor tfiid that binds to the tata box. *Nature*, 341(6240):299–303, 1989.

[123] KA Simmen, J Bernues, HD Parry, HG Stunnenberg, A Berkenstam, B Cavallini, JM Egly, and IW Mattaj. Tfiid is required for in vitro transcription of the human u6 gene by RNA polymerase iii. *The EMBO Journal*, 10(7):1853, 1991.

[124] Lucio Comai, Naoko Tanese, and Robert Tjian. The tata-binding protein and associated factors are integral components of the RNA polymerase I transcription factor, sl1. *Cell*, 68(5):965–976, 1992.

[125] Michael C Schultz, Ronald H Reeder, and Steven Hahn. Variants of the tata-binding protein can distinguish subsets of RNA polymerase I, II, and III promoters. *Cell*, 69(4):697–702, 1992.

[126] Ine Schaaff, Stefan Hohmann, and Friedrich K Zimmermann. Molecular analysis of the structural gene for yeast transaldolase. *European Journal of Biochemistry*, 188(3):597–603, 1990.

[127] Alexandra V Andreeva and Mikhail A Kutuzov. Protozoan protein tyrosine phosphatases. *International Journal for Parasitology*, 38(11):1279–1295, 2008.

[128] Jiangling Tu and M Carlson. Reg1 binds to protein phosphatase type 1 and regulates glucose repression in *Saccharomyces cerevisiae*. *The EMBO Journal*, 14(23):5939, 1995.

[129] Pascal Bernard, Kevin Hardwick, and Jean-Paul Javerzat. Fission yeast bub1 is a mitotic centromere protein essential for the spindle checkpoint and the preservation of correct ploidy through mitosis. *The Journal of Cell Biology*, 143(7):1775–1787, 1998.

[130] B Tibor Roberts, Katie A Farr, and M Andrew Hoyt. The *Saccharomyces cerevisiae* checkpoint gene bub1 encodes a novel protein kinase. *Molecular and Cellular Biology*, 14(12):8282–8291, 1994.

[131] Kevin G Hardwick et al. The spindle checkpoint. *Trends in Genetics: TIG*, 14(1):1, 1998.

[132] Kevin G Hardwick, Rong Li, Cathy Mistrot, Rey-Huei Chen, Phoebe Dann, Adam Rudner, and Andrew W Murray. Lesions in many different spindle components activate the spindle checkpoint in the budding yeast *Saccharomyces cerevisiae*. *Genetics*, 152(2):509–518, 1999.

[133] Stephen S Taylor and Frank McKeon. Kinetochore localization of murine bub1 is required for normal mitotic timing and checkpoint response to spindle damage. *Cell*, 89(5):727–735, 1997.

[134] Daniel P Cahill, Christoph Lengauer, Jian Yu, Gregory J Riggins, James KV Willson, Sanford D Markowitz, Kenneth W Kinzler, Bert Vogelstein, et al. Mutations of mitotic checkpoint genes in human cancers. *Nature*, 392(6673):300–303, 1998.

[135] Katie A Farr and M Andrew Hoyt. Bub1p kinase activates the *Saccharomyces cerevisiae* spindle assembly checkpoint. *Molecular and Cellular Biology*, 18(5):2738–2747, 1998.

[136] Audrey Killian, Nathalie Le Meur, Richard Sesboüé, Jeannette Bourguignon, Gaëlle Bougeard, Julien Gautherot, Christian Bastard, Thierry Frébourg, and Jean-Michel Flaman. Inactivation of the rrb1-pescadillo pathway involved in ribosome biogenesis induces chromosomal instability. *Oncogene*, 23(53):8597–8602, 2004.

[137] Tatiana L Iouk, John D Aitchison, Shawna Maguire, and Richard W Wozniak. Rrb1p, a yeast nuclear wd-repeat protein involved in the regulation of ribosome biosynthesis. *Molecular and Cellular Biology*, 21(4):1260–1271, 2001.

[138] Michael F Princiotta, Diana Finzi, Shu-Bing Qian, James Gibbs, Sebastian Schuchmann, Frank Buttgereit, Jack R Bennink, and Jonathan W Yewdell. Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity*, 18(3):343–354, 2003.

[139] Wolfgang Huber, Anja Von Heydebreck, Holger Sültmann, Annemarie Poustka, and Martin Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl 1):S96–S104, 2002.

[140] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[141] David Bentley. The mRNA assembly line: transcription and processing machines in the same factory. *Current Opinion in Cell Biology*, 14(3):336–342, 2002.

[142] Hermann SchaÈgger and Kathy Pfeiffer. Supercomplexes in the respiratory chains of yeast and mammalian mitochondria. *The EMBO Journal*, 19(8):1777–1783, 2000.

[143] Steven Wooding and Hugh RB Pelham. The dynamics of golgi protein traffic visualized in living yeast cells. *Molecular Biology of the Cell*, 9(9):2667–2680, 1998.

[144] Pavel Dolezal, Vladimir Likic, Jan Tachezy, and Trevor Lithgow. Evolution of the molecular machines for protein import into mitochondria. *Science*, 313(5785):314–318, 2006.

[145] M Feuermann, J De Montigny, S Potier, and J-L Souciet. The characterization of two new clusters of duplicated genes suggests a ‘lego’organization of the yeast *Saccharomyces cerevisiae* chromosomes. *Yeast*, 13(9):861–869, 1997.

[146] Roy J Britten. Cases of ancient mobile element DNA insertions that now affect gene regulation. *Molecular Phylogenetics and Evolution*, 5(1):13–17, 1996.

[147] Cédric Feschotte. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, 9(5):397–405, 2008.

[148] Jonathan D Dinman and Reed B Wickner. Ribosomal frameshifting efficiency and gag/gag-pol ratio are critical for yeast m1 double-stranded RNA virus propagation. *Journal of Virology*, 66(6):3669–3676, 1992.

[149] Jin M Kim, Swathi Vanguri, Jef D Boeke, Abram Gabriel, and Daniel F Voytas. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Research*, 8(5):464–478, 1998.

[150] GV Merkulov, KM Swiderek, C Baker Brachmann, and JD Boeke. A critical proteolytic cleavage site near the c terminus of the yeast retrotransposon ty1 gag protein. *Journal of Virology*, 70(8):5548–5556, 1996.

[151] Jeffrey S Smith and Jef D Boeke. An unusual form of transcriptional silencing in yeast ribosomal DNA. *Genes & Development*, 11(2):241–254, 1997.

[152] HR Graack and B Wittmann-Liebold. Mitochondrial ribosomal proteins (mrps) of yeast. *Biochemical Journal*, 329(Pt 3):433, 1998.

[153] Johannes M Herrmann, Michael W Woellhaf, and Nathalie Bonnefoy. Control of protein synthesis in yeast mitochondria: the concept of translational activators. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 2012.

[154] Sophia Y Lunt and Matthew G Vander Heiden. Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. *Annual Review of Cell and Developmental Biology*, 27:441–464, 2011.

[155] F-Nora Vögtle, Julia M Burkhart, Sanjana Rao, Carolin Gerbeth, Jens Hinrichs, Jean-Claude Martinou, Agnieszka Chacinska, Albert Sickmann, René P Zahedi, and Chris Meisinger. Intermembrane space proteome of yeast mitochondria. *Molecular & Cellular Proteomics*, 11(12):1840–1852, 2012.

[156] Johannes M Herrmann and Jan Riemer. The intermembrane space of mitochondria. *Antioxidants & Redox Signaling*, 13(9):1341–1358, 2010.

[157] Peter Uetz, Loic Giot, Gerard Cagney, Traci A Mansfield, Richard S Judson, James R Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, et al. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.

[158] Amy Hin Yan Tong, Marie Evangelista, Ainslie B Parsons, Hong Xu, Gary D Bader, Nicholas Page, Mark Robinson, Sasan Raghibizadeh, Christopher WV Hogue, Howard Bussey, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–2368, 2001.

[159] Tong Ihn Lee, Nicola J Rinaldi, François Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, 2002.

[160] AI Saeed, Vasily Sharov, Joe White, Jerry Li, Wei Liang, Nirmal Bhagabati, J Braisted, M Klapa, T Currier, M Thiagarajan, et al. Tm4: a free, open-source system for microarray data management and analysis. *Biotechniques*, 34(2):374, 2003.

[161] Audrey P Gasch, Paul T Spellman, Camilla M Kao, Orna Carmel-Harel, Michael B Eisen, Gisela Storz, David Botstein, and Patrick O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Science Signaling*, 11(12):4241, 2000.

[162] Stefan Hohmann and Willem H Mager. *Yeast stress responses*, volume 1. Springer Verlag, 2003.

[163] Ajit Varki, Richard D Cummings, Jeffrey D Esko, Hudson H Freeze, Pamela Stanley, Carolyn R Bertozzi, Gerald W Hart, Marilynn E Etzler, John B Lowe, et al. Biological roles of glycans. 2009.

[164] Thomas M Guadagno, Motoaki Ohtsubo, James M Roberts, and Richard K Assoian. A link between cyclin a expression and adhesion-dependent cell cycle progression. *Science*, 262(5139):1572–1575, 1993.

[165] Barry M Gumbiner. Cell adhesion: the molecular basis of tissue architecture and morphogenesis. *Cell*, 84(3):345–357, 1996.

[166] Kathryn E Wellen and Craig B Thompson. Cellular metabolic stress: considering how cells respond to nutrient excess. *Molecular Cell*, 40(2):323–332, 2010.

[167] Stephen Marshall, V Bacote, and RR Traxinger. Discovery of a metabolic pathway mediating glucose-induced desensitization of the glucose transport system. role of hexosamine biosynthesis in the induction of insulin resistance. *Journal of Biological Chemistry*, 266(8):4706–4712, 1991.

[168] Holly V Goodson, Caterina Valetti, and Thomas E Kreis. Motors and membrane traffic. *Current Opinion in Cell Biology*, 9(1):18–28, 1997.

[169] Nobutaka Hirokawa. Kinesin and dynein superfamily proteins and the mechanism of organelle transport. *Science*, 279(5350):519–526, 1998.

[170] VR Simon, SL Karmon, and LA Pon. Mitochondrial inheritance: cell cycle and actin cable dependence of polarized mitochondrial movements in *Saccharomyces cerevisiae*. *Cell Motility and the Cytoskeleton*, 37(3):199–210, 1997.

[171] Kammy L Fehrenbacher, Hyeong-Cheol Yang, Anna Card Gay, Thomas M Huckaba, and Liza A Pon. Live cell imaging of mitochondrial movement along actin cables in budding yeast. *Current Biology*, 14(22):1996–2004, 2004.

[172] Dianna G Fisk, Catherine A Ball, Kara Dolinski, Stacia R Engel, Eurie L Hong, Laurie Issel-Tarver, Katja Schwartz, Anand Sethuraman, David Botstein, and J Michael Cherry. *Saccharomyces cerevisiae* s288c genome annotation: a working hypothesis. *Yeast*, 23(12):857–865, 2006.

[173] Oleg A Barski, Srinivas M Tipparaju, and Aruni Bhatnagar. The aldo-keto reductase superfamily and its role in drug metabolism and detoxification. *Drug Metabolism Reviews*, 40(4):553–624, 2008.

[174] Jan van Riggelen, Alper Yetil, and Dean W Felsher. Myc as a regulator of ribosome biogenesis and protein synthesis. *Nature Reviews Cancer*, 10(4):301–309, 2010.

[175] Qian Dai, Shu-Bing Qian, Hui-Hua Li, Holly McDonough, Christoph Borchers, David Huang, Shinichi Takayama, J Michael Younger, Hong Yu Ren, Douglas M Cyr, et al. Regulation of the cytoplasmic quality control protein degradation pathway by bag2. *Journal of Biological Chemistry*, 280(46):38673–38681, 2005.

[176] Carmen Garrido, Mathilde Brunet, Celine Didelot, Yael Zermati, Elise Schmitt, and Guido Kroemer. Heat shock proteins 27 and 70: anti-apoptotic proteins with tumorigenic properties. *Cell Cycle*, 5(22):2592–2601, 2006.

[177] Jean-Claude Walser, Bing Chen, and Martin E Feder. Heat-shock promoters: targets for evolution by p transposable elements in drosophila. *PLoS Genetics*, 2(10):e165, 2006.

[178] Nathan J Bowen, I King Jordan, Jonathan A Epstein, Valerie Wood, and Henry L Levin. Retrotransposons and their recognition of pol ii promoters: a comprehensive survey of the transposable elements from the complete genome sequence of *Schizosaccharomyces pombe*. *Genome Research*, 13(9):1984–1997, 2003.

[179] Gabriella Rustici, Juan Mata, Katja Kivinen, Pietro Lió, Christopher J Penkett, Gavin Burns, Jacqueline Hayles, Alvis Brazma, Paul Nurse, and Jürg Bähler. Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics*, 36(8):809–817, 2004.

[180] Alexander S Spirin. Ribosome as a molecular machine. *FEBS letters*, 514(1):2–10, 2002.

[181] A Tzagoloff and Alan M Myers. Genetics of mitochondrial biogenesis. *Annual Review of Biochemistry*, 55(1):249–285, 1986.

[182] Hai Pan and David L Smith. Quaternary structure of aldolase leads to differences in its folding and unfolding intermediates. *Biochemistry*, 42(19):5713–5721, 2003.

[183] Andreas Heine, John G Luz, Chi-Huey Wong, and Ian A Wilson. Analysis of the class i aldolase binding site architecture based on the crystal structure of 2-deoxyribose-5-phosphate aldolase at 0.99 å resolution. *Journal of Molecular Biology*, 343(4):1019–1034, 2004.

[184] Peter T Erskine, Natalie Senior, Sarah Awan, Richard Lambert, Gareth Lewis, Ian J Tickle, M Sarwar, Paul Spencer, Paul Thomas, Martin J Warren, et al. X-ray structure of 5-aminolaevulinate dehydratase, a hybrid aldolase. *Nature Structural & Molecular Biology*, 4(12):1025–1031, 1997.

[185] Christine A Raines, Julie C Lloyd, Nicola M Willingham, Susan Potts, and Tristan A Dyer. cDNA and gene sequences of wheat chloroplast sedoheptulose-1, 7-bisphosphatase reveal homology with fructose-1, 6-bisphosphatases. *European Journal of Biochemistry*, 205(3):1053–1059, 1992.

[186] Hengming Ke, Cheleste M Thorpe, Barbara A Seaton, Frank Marcus, and William N Lipscomb. Molecular structure of fructose-1, 6-bisphosphatase at 2.8-a resolution. *Proceedings of the National Academy of Sciences*, 86(5):1475–1479, 1989.

[187] HM Ke, YP Zhang, and William N Lipscomb. Crystal structure of fructose-1, 6-bisphosphatase complexed with fructose 6-phosphate, amp, and magnesium. *Proceedings of the National Academy of Sciences*, 87(14):5243–5247, 1990.

[188] Hengming Ke, Cheleste M Thorpe, Barbara A Seaton, William N Lipscomb, and Frank Marcus. Structure refinement of fructose-1, 6-bisphosphatase and its fructose 2, 6-bisphosphate complex at 2.8 å resolution. *Journal of Molecular Biology*, 212(3):513–539, 1990.

[189] Chamel M Khoury, Zhao Yang, Xiao Yu Li, Marissa Vignali, Stanley Fields, and Michael T Greenwood. A tsc22-like motif defines a novel antiapoptotic protein family. *FEMS Yeast Research*, 8(4):540–563, 2008.

[190] Peter J Winn, Mai Zahran, JN Battey, Yanxiang Zhou, Rebecca C Wade, and Amit Banerjee. Structural and electrostatic properties of ubiquitination and related pathways. *Frontiers in Bioscience: A Journal and Virtual Library*, 12:3419, 2007.

[191] Debdyuti Mukhopadhyay and Howard Riezman. Proteasome-independent functions of ubiquitin in endocytosis and signaling. *Science*, 315(5809):201–205, 2007.

[192] Cecile M Pickart and Michael J Eddins. Ubiquitin: structures, functions, mechanisms. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1695(1):55–72, 2004.

[193] Greg Brown, Alexander Singer, Vladimir V Lunin, Michael Proudfoot, Tatiana Skarina, Robert Flick, Samvel Kochinyan, Ruslan Sanishvili, Andrzej Joachimiak, Aled M Edwards, et al. Structural and biochemical characterization of the type ii fructose-1, 6-bisphosphatase glpx from *Escherichia coli*. *Journal of Biological Chemistry*, 284(6):3784–3792, 2009.

[194] Ashok N Hegde and Aaron DiAntonio. Ubiquitin and the synapse. *Nature Reviews Neuroscience*, 3(11):854–861, 2002.

[195] Rebecca L Welchman, Colin Gordon, and R John Mayer. Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nature Reviews Molecular Cell Biology*, 6(8):599–609, 2005.

[196] Toru Higuchi, Yoshinori Watanabe, and Masayuki Yamamoto. Protein kinase a regulates sexual development and gluconeogenesis through phosphorylation of the zn finger transcriptional activator rst2p in fission yeast. *Molecular and Cellular Biology*, 22(1):1–11, 2002.

[197] Kim D Pruitt and Donna R Maglott. Refseq and locuslink: Ncbi gene-centered resources. *Nucleic Acids Research*, 29(1):137–140, 2001.

[198] Kevin P O'Brien, Maido Remm, and Erik LL Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, 33(suppl 1):D476–D480, 2005.

[199] Robert P Perry. Balanced production of ribosomal proteins. *Gene*, 401(1):1–3, 2007.

[200] Jonathan R Warner. Synthesis of ribosomes in *Saccharomyces cerevisiae*. *Microbiological Reviews*, 53(2):256–271, 1989.

[201] Saverio Brogna, Taka-Aki Sato, and Michael Rosbash. Ribosome components are associated with sites of transcription. *Molecular Cell*, 10(1):93–104, 2002.

[202] Patricia A Schroder and Melissa J Moore. Association of ribosomal proteins with nascent transcripts in *S. cerevisiae*. *RNA*, 11(10):1521–1529, 2005.

[203] Fengyi Wan, D Eric Anderson, Robert A Barnitz, Andrew Snow, Nicolas Bidere, Lixin Zheng, Vijay Hegde, Lloyd T Lam, Louis M Staudt, David Levens, et al. Ribosomal protein s3: a kh domain subunit in nf-κb complexes that mediates selective gene regulation. *Cell*, 131(5):927–939, 2007.

[204] Jian-Quan Ni, Lu-Ping Liu, Daniel Hess, Jens Rietdorf, and Fang-Lin Sun. Drosophila ribosomal proteins are associated with linker histone h1 and suppress gene transcription. *Genes & Development*, 20(14):1959–1973, 2006.

[205] Yun Wah Lam, Angus I Lamond, Matthias Mann, and Jens S Andersen. Analysis of nucleolar protein dynamics reveals the nuclear degradation of ribosomal proteins. *Current Biology*, 17(9):749–760, 2007.

[206] John W Nicol, Gregg A Helt, Steven G Blanchard Jr, Archana Raja, and Ann E Loraine. The integrated genome browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, 25(20):2730–2731, 2009.

[207] Christian Heichinger, Christopher J Penkett, Jürg Bähler, and Paul Nurse. Genome-wide characterization of fission yeast DNA replication origins. *The EMBO Journal*, 25(21):5171–5179, 2006.

[208] Alison L Pidoux and Robin C Allshire. Kinetochore and heterochromatin domains of the fission yeast centromere. *Chromosome Research*, 12(6):521–534, 2004.

[209] Tom Volpe, Vera Schramke, Georgina L Hamilton, Sharon A White, Grace Teng, Robert A Martienssen, and Robin C Allshire. RNA interference is required for normal centromere function infission yeast. *Chromosome Research*, 11(2):137–146, 2003.

[210] Brian T Wilhelm, Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J Penkett, Jane Rogers, and Jürg Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243, 2008.

[211] Eun Shik Choi, Annelie Strålfors, Araceli G Castillo, Mickaël Durand-Dubief, Karl Ekwall, and Robin C Allshire. Identification of noncoding transcripts from within cenp-a chromatin at fission yeast centromeres. *Journal of Biological Chemistry*, 286(26):23600–23607, 2011.

[212] Sergey Melnikov, Adam Ben-Shem, Nicolas Garreau de Loubresse, Lasse Jenner, Gulnara Yusupova, and Marat Yusupov. One core, two shells: bacterial and eukaryotic ribosomes. *Nature Structural & Molecular Biology*, 19(6):560–567, 2012.

[213] Kushal Nivriti Rugjee, Subhendu Roy Chaudhury, Khalid Al-Jubran, Preethi Ramanathan, Tina Matina, Jikai Wen, and Saverio Brogna. Fluorescent protein tagging confirms the presence of ribosomal proteins at drosophila polytene chromosomes. *PeerJ*, 1:e15, 2013.

[214] Giorgio Dieci, Roberta Ruotolo, Priscilla Braglia, Christophe Carles, Andrea Carpentieri, Angela Amoresano, and Simone Ottonello. Positive modulation of RNA polymerase III transcription by ribosomal proteins. *Biochemical and Biophysical Research Communications*, 379(2):489–493, 2009.

[215] Mu-Shui Dai, Xiao-Xin Sun, and Hua Lu. Ribosomal protein l11 associates with c-myc at 5 s rrna and trna genes and regulates their expression. *Journal of Biological Chemistry*, 285(17):12587–12594, 2010.

[216] Derick G Wansink, Ody CM Sibon, Fons FM Cremers, Roel van Driel, and Luitzen de Jong. Ultrastructural localization of active genes in nuclei of a431 cells. *Journal of Cellular Biochemistry*, 62:10–18, 1996.

[217] Dean A Jackson, A Bassim Hassan, Rachel J Errington, and Peter R Cook. Visualization of focal sites of transcription within human nuclei. *The EMBO Journal*, 12(3):1059, 1993.

[218] Zhihong Cheng, Denise Muhlrad, Meng Kiat Lim, Roy Parker, and Haiwei Song. Structural and functional insights into the human upf1 helicase core. *The EMBO journal*, 26(1):253–264, 2006.

[219] Anirban Bhattacharya, Kevin Czaplinski, Panayiota Trifillis, Feng He, Allan Jacobson, and Stuart W Peltz. Characterization of the biochemical properties of the human upf1 gene product that is involved in nonsense-mediated mRNA decay. *RNA*, 6(9):1226–1235, 2000.

[220] Yao-Fu Chang, J Saadi Imam, and Miles F Wilkinson. The nonsense-mediated decay RNA surveillance pathway. *Annu. Rev. Biochem.*, 76:51–74, 2007.

[221] Pamela Nicholson, Hasmik Yepiskoposyan, Stefanie Metze, Rodolfo Zamudio Orozco, Nicole Kleinschmidt, and Oliver Mühlemann. Nonsense-mediated mRNA decay in human cells:

mechanistic insights, functions beyond quality control and the double-life of nmd factors. *Cellular and Molecular Life Sciences*, 67(5):677–700, 2010.

[222] Wazeer Varsally and Saverio Brogna. Upf1 involvement in nuclear functions. *Biochemical Society Transactions*, 40(4):778, 2012.

[223] Saverio Brogna, Preethi Ramanathan, Jikai Wen, et al. Upf1 p-body localization. *Biochemical Society Transactions*, 36(4):698–700, 2008.

[224] Lu´sa Romão, Ângela Inácio, Susana Santos, Madalena Ávila, Paula Faustino, Paula Pacheco, and João Lavinha. Nonsense mutations in the human β-globin gene lead to unexpected levels of cytoplasmic mRNA accumulation. *Blood*, 96(8):2895–2901, 2000.

[225] Saverio Brogna and Jikai Wen. Nonsense-mediated mrna decay (nmd) mechanisms. *Nature Structural & Molecular Biology*, 16(2):107–113, 2009.

[226] Guramrit Singh, Steffen Jakob, Mark G Kleedehn, and Jens Lykke-Andersen. Communication with the exon-junction complex and activation of nonsense-mediated decay by human upf proteins occur in the cytoplasm. *Molecular Cell*, 27(5):780–792, 2007.

[227] Nicolas Kuperwasser, Saverio Brogna, Ken Dower, and Michael Rosbash. Nonsense-mediated decay does not occur within the yeast nucleus. *RNA*, 10(12):1907–1915, 2004.

[228] Pamela A Frischmeyer and Harry C Dietz. Nonsense-mediated mrna decay in health and disease. *Human Molecular Genetics*, 8(10):1893–1900, 1999.

[229] Jan Rehwinkel, Ivica Letunic, Jeroen Raes, Peer Bork, and Elisa Izaurralde. Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. *RNA*, 11(10):1530–1544, 2005.

[230] Claus M Azzalin and Joachim Lingner. The human RNA surveillance factor upf1 is required for s phase progression and genome stability. *Current Biology*, 16(4):433–439, 2006.

[231] Jeffrey N Dahlseid, Jodi Lew-Smith, Michael J Lelivelt, Shinichiro Enomoto, Amanda Ford, Michelle Desruisseaux, Mark McClellan, Neal Lue, Michael R Culbertson, and Judith Berman. mRNAs encoding telomerase components and regulators are controlled by upf genes in *Saccharomyces cerevisiae*. *Eukaryotic Cell*, 2(1):134–142, 2003.

[232] Sophie Rozenzhak, Eva Mejá-Ramrez, Jessica S Williams, Lana Schaffer, Jennifer A Hammond, Steven R Head, and Paul Russell. Rad3atr decorates critical chromosomal domains with γh2a to protect genome integrity during s-phase in fission yeast. *PLoS genetics*, 6(7):e1001032, 2010.

[233] Rachel Lyne, Gavin Burns, Juan Mata, Chris J Penkett, Gabriella Rustici, Dongrong Chen, Cordelia Langford, David Vetrie, and Jürg Bähler. Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC genomics*, 4(1):27, 2003.

[234] Katharine Compton Abruzzi, Scott Lacadie, and Michael Rosbash. Biochemical analysis of trex complex recruitment to intronless and intron-containing yeast genes. *The EMBO journal*, 23(13):2620–2631, 2004.

[235] Miguel A Rodrguez-Gabriel, Stephen Watt, Jürg Bähler, and Paul Russell. Upf1, an RNA helicase required for nonsense-mediated mrna decay, modulates the transcriptional response to oxidative stress in fission yeast. *Molecular and Cellular Biology*, 26(17):6347–6356, 2006.

[236] Alexander B Steever, Achim Wach, Peter Phlippsen, and John R Pringle. Heterologous modules for efficient and versatile pcr-based gene targeting in *Schizosaccharomyces pombe*. *Yeast*, 14:943–951, 1998.

[237] Thomas R Cech. Beginning to understand the end of the chromosome. *Cell*, 116(2):273–279, 2004.

[238] Olaf Isken and Lynne E Maquat. The multiple lives of NMD factors: balancing roles in gene and genome regulation. *Nature Reviews Genetics*, 9(9):699–712, 2008.

[239] Bela Novak, Attila Csikasz-Nagy, Bela Gyorffy, Kathy Chen, John J Tyson, et al. Mathematical model of the fission yeast cell cycle with checkpoint controls at the g1/s, g2/m and metaphase/anaphase transitions. *Biophysical Chemistry*, 72(1):185–200, 1998.

[240] Stephanie S Pao, Ian T Paulsen, and Milton H Saier. Major facilitator superfamily. *Microbiology and Molecular Biology Reviews*, 62(1):1–34, 1998.

[241] Shiv IS Grewal and Sarah CR Elgin. Transcription and RNA interference in the formation of heterochromatin. *Nature*, 447(7143):399–406, 2007.

[242] Hugh P Cam, Tomoyasu Sugiyama, Ee Sin Chen, Xi Chen, Peter C FitzGerald, and Shiv IS Grewal. Comprehensive analysis of heterochromatin-and RNAi-mediated epigenetic control of the fission yeast genome. *Nature Genetics*, 37(8):809–819, 2005.

[243] Kristin C Scott, Caroline V White, and Huntington F Willard. An RNA polymerase III-dependent heterochromatin barrier at fission yeast centromere 1. *PLoS One*, 2(10):e1099, 2007.

[244] Gurumurthy D Shankaranarayana, Mohammad R Motamedi, Danesh Moazed, and Shiv IS Grewal. Sir2 regulates histone h3 lysine 9 methylation and heterochromatin assembly in fission yeast. *Current Biology*, 13(14):1240–1246, 2003.

[245] Elizabeth Pennisi. How the genome readies itself for evolution. *Science*, 281(5380):1131–1134, 1998.

[246] Kuheli Chakrabarty, Sampad Narayan Gupta, Gourab Kanti Das, and Sukhendu Roy. Theoretical studies on the pyridoxal-5'-phosphate dependent enzyme dopa decarboxylase: Effect of thr 246 residue on the co-factor-enzyme binding and reaction mechanism. 2012.

[247] Jikai Wen and Saverio Brogna. Splicing-dependent nmd does not require the ejc in *Schizosaccharomyces pombe*. *The EMBO Journal*, 29(9):1537–1551, 2010.

[248] L Michael Carastro, Cheng-Keat Tan, Manuel Selg, Hans-Martin Jack, Antero G So, and Kathleen M Downey. Identification of delta helicase as the bovine homolog of hupf1: demonstration of an interaction with the third subunit of DNA polymerase delta. *Nucleic Acids Research*, 30(10):2232–2243, 2002.

[249] Claus M Azzalin, Patrick Reichenbach, Lela Khoriauli, Elena Giulotto, and Joachim Lingner. Telomeric repeat–containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science*, 318(5851):798–801, 2007.

[250] Bik K Tye. Mcm proteins in DNA replication. *Annual Review of Biochemistry*, 68(1):649–686, 1999.

[251] Gerald T Ankley, Richard S Bennett, Russell J Erickson, Dale J Hoff, Michael W Hornung, Rodney D Johnson, David R Mount, John W Nichols, Christine L Russom, Patricia K Schmieder, et al. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry*, 29(3):730–741, 2010.

[252] Vincent J Kramer, Matthew A Etterson, Markus Hecker, Cheryl A Murphy, Guritno Roesijadi, Daniel J Spade, Julann A Spromberg, Magnus Wang, and Gerald T Ankley. Adverse outcome pathways and ecological risk assessment: Bridging to population-level effects. *Environmental Toxicology and Chemistry*, 30(1):64–76, 2011.

[253] Edward J Perkins, J Kevin Chipman, Stephen Edwards, Tanwir Habib, Francesco Falciani, Ronald Taylor, Graham Van Aggelen, Chris Vulpe, Philipp Antczak, and Alexandre Loguinov. Reverse engineering adverse outcome pathways. *Environmental Toxicology and Chemistry*, 30(1):22–38, 2011.

[254] William J Jo, Alex Loguinov, Henri Wintz, Michelle Chang, Allan H Smith, Dave Kalman, Luoping Zhang, Martyn T Smith, and Chris D Vulpe. Comparative functional genomic analysis identifies distinct and overlapping sets of genes required for resistance to monomethylarsonous acid (mmaiii) and arsenite (asiii) in yeast. *Toxicological Sciences*, 111(2):424–436, 2009.

[255] Michael Waisberg, Pius Joseph, Beverley Hale, and Detmar Beyersmann. Molecular and cellular mechanisms of cadmium carcinogenesis. *Toxicology*, 192(2):95–117, 2003.

[256] Florianne Monnet-Tschudi, Corina Boschat, Anne Corbaz, and Paul Honegger. Involvement of environmental mercury and lead in the etiology of neurodegenerative diseases. *Reviews on Environmental Health*, 21(2):105–118, 2006.

[257] Štefan Fujs, Zoltán Gazdag, Borut Poljšak, Vekoslava Stibilj, Radmila Milacic, Miklós Pesti, Peter Raspor, and Martin Batic. The oxidative stress response of the yeast candida intermedia to copper, zinc, and selenium exposure. *Journal of Basic Microbiology*, 45(2):125–135, 2005.

[258] Sandeep K Sharma, Pierre Goloubinoff, and Philipp Christen. Heavy metal ions are potent inhibitors of protein folding. *Biochemical and Biophysical Research Communications*, 372(2):341–345, 2008.

[259] Tomás R Guilarte, Christopher D Toscano, Jennifer L McGlothan, and Shelley A Weaver. Environmental enrichment reverses cognitive and molecular deficits induced by developmental lead exposure. *Annals of Neurology*, 53(1):50–56, 2003.

[260] Therese Jacobson, Clara Navarrete, Sandeep K Sharma, Theodora C Sideri, Sebastian Ibstedt, Smriti Priya, Chris M Grant, Philipp Christen, Pierre Goloubinoff, and Markus J Tamás.

Arsenite interferes with protein folding and triggers formation of protein aggregates in yeast. *Journal of Cell Science*, 2012.

[261] Pierre J Dilda, Gabriel G Perrone, Amanda Philp, Richard B Lock, Ian W Dawes, and Philip J Hogg. Insight into the selectivity of arsenic trioxide for acute promyelocytic leukemia cells by characterizing *Saccharomyces cerevisiae* deletion strains that are sensitive or resistant to the metalloid. *International Journal of Biochemistry and Cell Biology*, 40(5):1016–1029, 2008.

[262] Xuewen Pan, Stefanie Reissman, Nick R Douglas, Zhiwei Huang, Daniel S Yuan, Xiaoling Wang, J Michael McCaffery, Judith Frydman, and Jef D Boeke. Trivalent arsenic inhibits the functions of chaperonin complex. *Genetics*, 186(2):725–734, 2010.

[263] Sandeep K Sharma, Philipp Christen, Pierre Goloubinoff, et al. Disaggregating chaperones: an unfolding story. *Current Protein & Peptide Science*, 10(5):432, 2009.

[264] Kenneth H Falchuk. The molecular basis for the role of zinc in developmental biology. In *Molecular and Cellular Effects of Nutrition on Disease Processes*, pages 41–48. Springer, 1998.

[265] Defeng Wu, Arthur I Cederbaum, et al. Alcohol, oxidative stress, and free radical damage. *Alcohol Research and Health*, 27:277–284, 2003.

[266] G Bertin and D Averbeck. Cadmium: cellular effects, modifications of biomolecules, modulation of DNA repair and genotoxic consequences (a review). *Biochimie*, 88(11):1549–1559, 2006.

[267] DB Vinh and David G Drubin. A yeast tcp-1-like protein is required for actin function in vivo. *Proceedings of the National Academy of Sciences*, 91(19):9116–9120, 1994.

[268] Tim C Huffaker, M Andrew Hoyt, and David Botstein. Genetic analysis of the yeast cytoskeleton. *Annual Review of Genetics*, 21(1):259–284, 1987.

[269] Yaping Chen and Peter W Piper. Consequences of the overexpression of ubiquitin in yeast: elevated tolerances of osmostress, ethanol and canavanine, yet reduced tolerances of cadmium, arsenite and paromomycin. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1268(1):59–64, 1995.

[270] Edward McEwen, Nancy Kedersha, Benbo Song, Donalyn Scheuner, Natalie Gilks, Anping Han, Jane-Jane Chen, Paul Anderson, and Randal J Kaufman. Heme-regulated inhibitor kinase-mediated phosphorylation of eukaryotic translation initiation factor 2 inhibits translation, induces stress granule formation, and mediates survival upon arsenite exposure. *Journal of Biological Chemistry*, 280(17):16925–16933, 2005.

[271] Rachid Mazroui, Rami Sukarieh, Marie-Eve Bordeleau, Randal J Kaufman, Peter Northcote, Junichi Tanaka, Imed Gallouzi, and Jerry Pelletier. Inhibition of ribosome recruitment induces stress granule formation independently of eukaryotic initiation factor 2α phosphorylation. *Molecular Biology of the Cell*, 17(10):4212–4219, 2006.

[272] Bernd Bukau and Arthur L Horwich. The hsp70 and hsp60 chaperone machines. *Cell*, 92(3):351–366, 1998.

[273] Julie E Archer, Leticia R Vega, and Frank Solomon. Rbl2p, a yeast protein that binds to β-tubulin and participates in microtubule function in vivo. *Cell*, 82(3):425–434, 1995.

[274] Andrea Mauri, Viviana Consonni, Manuela Pavan, and Roberto Todeschini. Dragon software: An easy approach to molecular descriptor calculations. *Match*, 56(2):237–248, 2006.

[275] Kristin Breitschopf, Eyal Bengal, Tamar Ziv, Arie Admon, and Aaron Ciechanover. A novel site for ubiquitination: the n-terminal residue, and not internal lysines of myod, is essential for conjugation and degradation of the protein. *The EMBO Journal*, 17(20):5964–5973, 1998.

[276] Kenneth Matthew Scaglione, Venkatesha Basrur, Naila S Ashraf, John R Konen, Kojo SJ Elenitoba-Johnson, Sokol V Todi, and Henry L Paulson. The ubiquitin conjugating enzyme (e2) ube2w ubiquitinates the n-terminus of substrates. *Journal of Biological Chemistry*, 2013.

[277] Nitnipa Soontorngun, Marc Larochelle, Simon Drouin, François Robert, and Bernard Turcotte. Regulation of gluconeogenesis in *Saccharomyces cerevisiae* is mediated by activator and repressor functions of rds2. *Molecular and Cellular Biology*, 27(22):7895–7905, 2007.

[278] Joseph L DeRisi, Vishwanath R Iyer, and Patrick O Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, 1997.

[279] Tae-Sung Kim, Chang-Young Jang, Hag Dong Kim, Jae Yung Lee, Byung-Yoon Ahn, and Joon Kim. Interaction of hsp90 with ribosomal proteins protects from ubiquitination and proteasome-dependent degradation. *Molecular Biology of the Cell*, 17(2):824–833, 2006.

[280] Elena S Tasheva and Donald J Roufa. Regulation of human rps14 transcription by intronic antisense RNAs and ribosomal protein s14. *Genes & Development*, 9(3):304–316, 1995.

[281] Ralph J DeBerardinis, Nabil Sayed, Dara Ditsworth, and Craig B Thompson. Brick by brick: metabolism and tumor cell growth. *Current Opinion in Genetics & Development*, 18(1):54–61, 2008.

[282] Laszlo G Boros, Joaquim Puigjaner, Marta Cascante, Wai-Nang Paul Lee, James L Brandes, Sara Bassilian, Fouza I Yusuf, Robert D Williams, Pete Muscarella, W Scott Melvin, et al. Oxythiamine and dehydroepiandrosterone inhibit the nonoxidative synthesis of ribose and tumor cell proliferation. *Cancer Research*, 57(19):4242–4248, 1997.

[283] László G Boros, Marta Cascante, and Wai-Nang Paul Lee. Metabolic profiling of cell growth and death in cancer: applications in drug discovery. *Drug Discovery Today*, 7(6):364–372, 2002.

[284] James J-L Wang, Peter R Martin, and Charles K Singleton. A transketolase assembly defect in a wernicke-korsakoff syndrome patient. *Alcoholism: Clinical and Experimental Research*, 21(4):576–580, 1997.

[285] Stephen P Miller, Gulshan R Anand, Elizabeth J Karschnia, Graeme I Bell, David C LaPorte, and Alex J Lange. Characterization of glucokinase mutations associated with maturity-onset diabetes of the young type 2 (mody-2): different glucokinase defects lead to a common phenotype. *Diabetes*, 48(8):1645–1651, 1999.

[286] Alexandre Wajngot, Visvanathan Chandramouli, William C Schumann, Karin Ekberg, Paul K Jones, Suad Efendic, and Bernard R Landau. Quantitative contributions of gluconeogenesis to glucose production during fasting in type 2 diabetes mellitus. *Metabolism*, 50(1):47–52, 2001.

[287] Qun Dang, Srinivas Rao Kasibhatla, K Raja Reddy, Tao Jiang, M Rami Reddy, Scott C Potter, James M Fujitaki, Paul D van Poelje, Jingwei Huang, William N Lipscomb, et al. Discovery of potent and specific fructose-1, 6-bisphosphatase inhibitors and a series of orally-bioavailable phosphoramidase-sensitive prodrugs for the treatment of type 2 diabetes. *Journal of the American Chemical Society*, 129(50):15491–15502, 2007.