DETECTING AND UNDERSTANDING DIFFERENTIATION OF MICROARRAY EXPRESSION DATA

by

Chao Cheng

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(COMPUTATIONAL BIOLOGY AND BIOINFORMATICS)

December 2006

Dedication

to my parents

Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Lei Li, for his patience and guidance during my graduate study. Without his constant support and encouragement, the completion of the thesis would not be possible. I feel privileged to have had the opportunity to work closely and learn from him.

I am very grateful to Professor Valter D. Longo, Dr. Paola Fabrizio and Dr. Min Wei and all the other members in Professor Longo's Lab. They provide us high quality Microarray data sets and most of my knowledge about ageing is learned from them. Their enthusiasm to science and research will always benefit me in my future professional life.

I would like to thank Professor Fengzhu Sun and Professor Tim Chen and all the previous and current members in their research group: Dr. Minghua Deng, Dr. Kui Zhang, Dr. Lei Zhuge, Dr. Xiaoma Tu, Dr. Rui Jiang, Dr. Debo Dutta, Dr. Hyun-ju Lee, Dr. Quansong Ruan, Dr. Bingwen lu, Dr. Xiting Yan, Dr. Yunhu Wan, Dr. Hua Yang, ZhiDong Tu, Li Wang, Linqi Zhou. I appreciate the help and suggestions from them in the past few years.

I would like to thank to some other colleagues and friends: Keyan Zhao, Kangyu Zhang, Min Xu, Ming Li, Huanying Ge, Yu Huang, Hua Bai, et al., for their help in these years.

Finally, I would like to thank my little sister and my parents for their love, support and encouragement. I would also like to thank my grandmother, who passed away in 2002, soon after I began my study in USC. I feel sorry that I did not accompany her in her last minute. This thesis is dedicated to her.

Table of Contents

List of Tables

List of Figures

Abstract

This thesis consists of three parts, reflecting three levels of Microarray data analysis.

In the first part, we introduce a new normalization method for Affymetrix oligonucleotide based arrays. Our perspective is to find a transformation that matches the distributions of hybridization levels of those probes corresponding to undifferentiated genes between arrays. We address two important issues. First, array-specific spatial patterns exist due to uneven hybridization and measurement process. Second, in some cases a substantially large portion of genes are differentially expressed between a target and a reference array. For the purpose of normalization we need to identify a subset that excludes those probes corresponding to differentially expressed genes and abnormal probes due to experimental variation. Least trimmed squares (LTS) is a natural choice to achieve this goal. Substantial differentiation is protected in LTS by setting an appropriate trimming fraction. To take into account any spatial pattern of hybridization, we divide each array into sub-arrays and normalize probe intensities within each sub-array. We illustrate the problem and solution through an Affymetrix spike-in data set with defined perturbation and a data set of primate brain expression.

In the second part, we describe a novel method to identify substantially perturbed genes in the treatment/control time course data sets. It is often difficult to compare expression patterns of a gene of two time courses for the following reasons: (1) the number of sampling time points may be different or hard to be aligned between the

treatment and the control time courses; (2) estimation of the function that describes the expression of a gene in a time course is difficult and error-prone due to the limited number of time points. We propose a novel method to identify the differentially expressed genes between two time courses which avoids direct comparison of gene expression patterns of the two time courses. This method does not require alignment between the two time courses to be compared. Instead of attempting to "align" and compare the two time courses directly, we first convert the treatment and control time courses into two neighborhood systems that reflect the underlying relationships between genes. We then identify the differentially expressed genes by comparing the two gene relationship networks from the two neighborhood systems. To verify our method, we apply it to several treatment-control time course data sets. The results are consistent with the previous results and also give some new biologically meaningful findings.

In the third part, we describe our integrative analysis of Microarray data from long-lived yeast mutants. To understand gene expression change in these mutants from a systematic perspective, we combine Microarray data with many other data sources, such as literatures, Gene Ontology, KEGG, and so on. Our results show that these long-lived mutants share some common features in gene expression changes. Gene categories involved in basal transcription, translation and ion transportation tend to be down-regulated. The glycolysis/gluconeogenisis pathway is significantly activated, whereas the oxidative phosphorylation pathway and the citric acid cycle pathway are somehow repressed. These findings may shed light on the underlying mechanisms of longevity of these mutants.

Chapter 1

Introduction

1.1 Gene expression and Microarray technologies

In the world of bio-molecules, proteins play the key roles as structural components, enzymes, antibodies, and so on. Genes in DNA molecules carry the encoding information for proteins. The flow of this encoding information from genes to proteins involves two stages: transcription and translation. As shown in Figure 1.1, in transcription, a DNA segment that constitutes a gene in the DNA molecules is transcribed into a single stranded sequence of RNA, called messenger RNA (mRNA). Then in translation, the mRNA is translated into a sequence of amino acids which finally become a protein after some modifications. To study the biological system quantitatively, several techniques have been developed to measure the expression levels of mRNAs and proteins. These techniques include Western Blot, Enzyme-Linked ImmunoSorbent Assay (ELISA), Mass Spectrometry (MS) and Protein Microarrays for protein expression measuring and serial Analysis of Gene Expression (SAGE), Northern Blot, quantitative RT-PCR, and DNA Microarrays for mRNA expression measuring. Although expression levels for mRNA and proteins are both of interest in biological studies, this thesis will focus on DNA Microarrays data.

To measure the expression levels of genes using the DNA Microarray techniques, hundreds of thousands of DNA probes are immobilized on a small glass, plastic, or nylon membrane which is called an array. These probes are designed to stand for certain amount of genes. mRNAs from the sample cells are hybridized with the probes on the

array. So by measuring the intensity of the mRNAs hybridized with the probes, we can have the expression levels of the genes that we're interested in. This technique enables us to measure the expression values for hundreds of thousands of genes simultaneously so that we can observe the changes in genes' expression systematically. Also with the aid of the Microarray techniques, we are able to design more intricate experiments to predict gene function, infer gene regulatory networks, understand disease mechanisms, et al.

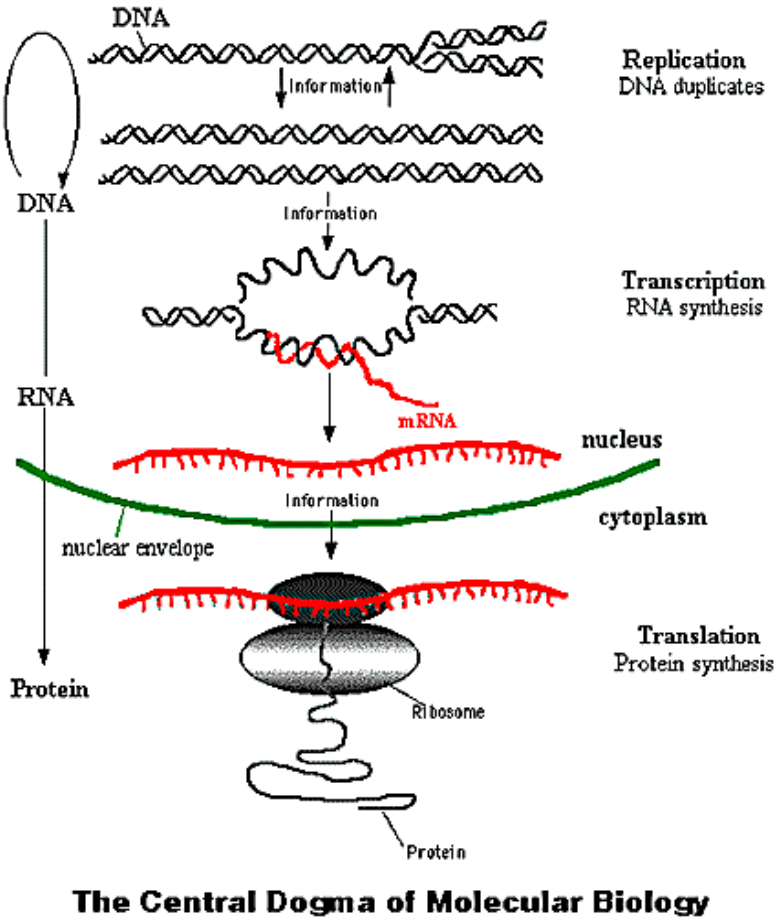

Figure 1.1: A scheme of mRNA in gene transcription and protein translation. *Picture is copied from http://www.accessexcellence.org*.

1.2 Types of Microarrays

There are a number of microarray technologies for large scale gene expression measuring. Among them, cDNA arrays and oligonucleotide arrays are the most popular approaches. Although they use the same principle, they differ in many aspects.

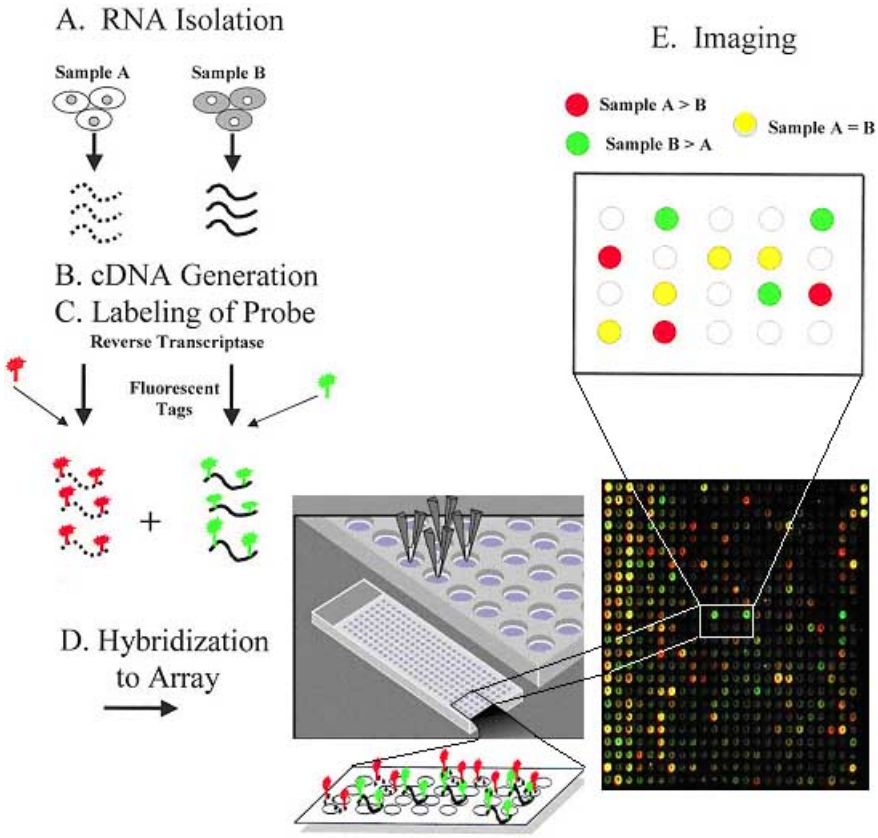

Figure 1.2: Illustration of cDNA array experiment. *From http://www.fao.org*.

In a typical cDNA array experiment, mRNAs from two different samples are extracted and reverse-transcribed into cDNAs which are labeled with dyes of different colors if they're in different samples. Then equal amount of labeled cDNA samples are mixed together and hybridized with the probes on the array. The probes are spotted cDNA of hundreds of nucleotides in length. After the hybridization, a laser scanner measures dye fluorescence of each color at a fine grid of pixels. Higher fluorescence

indicates higher amount of hybridized cDNA and hence higher gene expression in the corresponding sample [SSDB95, DIB97]. The experiment procedure described above is also illustrated in Figure 1.2. After the scanning, we typically have two intensities for spotted cDNA of two colors and two intensities for the background of two colors. So there're at least four quantities for each probe on the cDNA array. Sometimes, these are accompanied with quantities that measure the quality of the spot, e.g. the variability of the pixel intensity. Since samples are labeled with different colors and hybridize competitively to the same set of probes, the cDNA array is also called two-channel array. The two channel array allows measurement of the relative gene expression in the two samples, i.e. the ratios of the two colors for each spot.

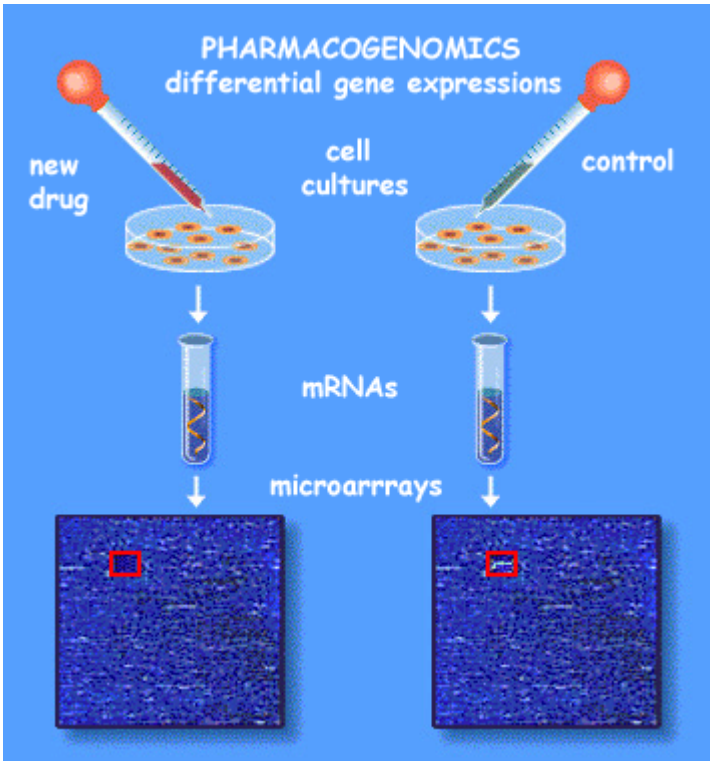

Figure 1.3: Illustration of oligonucleotide array experiment. *From http://fig.cox.miami.edu*

The oligonucleotide arrays are available commercially from several companies, such as Affymetrix, Illumina, NimbleGen, Agilent et al. Although they use different techniques, they have one thing in common: the short oligonucleotide sequences are used as probes. For example, in Affymetrx array, each gene is represented by one or more probe sets, each composed of 11-20 pairs of 25bps long oligonucleotide. Each pair consists of a perfect match and a mismatch. The mismatch is created by changing the middle (13th) base of the perfect match sequence to reduce the specific binding of mRNA for that gene. The goal of mismatch is to control experimental variation and nonspecific binding of other mRNAs with the probe [Aff01]. Unlike the two-channel cDNA array, oligonucleotide array is often one-channel: mRNA from only one sample is prepared, labeled with a fluorescent dye, and hybridized to the probes on an array. After the hybridization, arrays are scanned, and images are produced and analyzed to obtain a fluorescence intensity value for each probe. In the probe set level, the typical output for a probe set includes two vectors of intensity readings, one for perfect matches and the other for mismatches. The experiment procedure using oligonucleotide arrays is also illustrated in Figure 1.3.

1.3 Microarray data analysis

Despite the high throughput and high efficiency of microarray technologies, high level of noises and complex experimental artifacts are associated with microarray data, which emphasizes the requirement for statistical and data analytic techniques for all stage of experimentation. Microarray data analysis can roughly be classified into three levels: low, middle, and high level, according to the stage of experimentation and involvement of other data sources.

Low level of data analysis, also termed as signal extraction, includes image analysis, gene filtering, background correction, probe level analysis and gene summarization for oligonucleotide arrays, as well as between-array normalization and removal of artifacts for comparisons across arrays. These kinds of data analysis are performed at the early stage of microarray experimentation. For example, normalization and summarization are often performed to obtain expression values of genes from raw data set collected from Affymetrix gene chips. In Chapter 2, we will give a brief introduction to several prevalent normalization methods as well as a new method we proposed for Affymetric arrays, called Sub-Sub.

Middle level of data analysis includes selection of differentially expressed genes between experimental conditions, clustering/classification of biological samples or genes, construction of gene co-expression network, et al. For instance, in order to understand the mechanism of a type of cancer, say lung cancer, we are interested in: (1) what are the physiologically different between the cells in the tumors and in normal lung tissues? (2) Which genes show expression change in the tumor cell compared with normal cell? That is, we try to associate the physiological difference with gene expression changes so that we can shed light on the mechanism of lung cancer in a molecule level. When two different conditions are considered, such as disease/non-disease, we also denote one as treatment and the other as control. Differential expression between treatment and control can be investigated from a static or temporal viewpoint. In a static experiment design, snapshots of gene expression levels are taken without considering the temporal effect. Whereas in a temporal experiment design, also called a time course design, the gene expression across several time points are measured. To identify differentially expressed genes in a static experiment design, a number of approaches have been proposed, including the two-sample t-test (T-test), the Wilcoxon rank sum test (WRST), significance analysis of microarrays (SAM) [TTC01], and relative entropy

based method [YDFQ05]. Several approaches have also developed to identify differentially expresssed genes between time courses in a temporal experiment design. In Chapter 3, we will first introduce some of them, then we will describe a novel method we proposed, which is called MARD analysis.

High level of microarray data analysis includes those approaches that integrate microarray data sets from different platforms or combine microarray data with other data sources, such as Chip-Chip results, Gene Ontology information, pathway information, and so on. Great success has been achieved in the past few years by performing high level microarray data analysis. For example, Segal et al. presented an integrated analysis of 1,975 published microarrays spanning 22 tumor types. They described expression profiles in different tumors in terms of the behavior of modules, which are gene sets that act in concert to carry out a specific function. Using a simple unified analysis, they extracted modules and characterized gene-expression profiles in tumors as a combination of activated and deactivated modules. Activation of some modules was found to be specific to particular types of tumor, whereas other modules were shared across a diverse set of clinical conditions, which suggests the existence of common tumor progression mechanisms [SFKR04]. In another paper, Subramanian et al. described a method called Gene Set Enrichment Analysis (GSEA) [STM$^+$05]. This method focuses on gene sets which are groups of genes that share common biological function, chromosomal location, or regulation. Construction of gene sets is based on information collected from literatures and other sources of data sets. In Chapter 4, we will also apply some integrative analysis to our ageing project and show how the analysis exposes a common mechanism for longevity in four long-lived yeast mutants.

Chapter 2

Sub-array normalization for oligonucleotide array

The high density oligonucleotide array has been widely used in biological studies. Analysis of oligonucleotide array data includes several steps: image processing, background correction, normalization, PM correction, and probe set summary. In this chapter, we first describe the importance of normalization, and introduce several approaches available for oligonucleotide array normalization. Then we propose a novel normalization method called Sub-Sub (sub-array normalization subject to differentiation). Our method allows a substantial differentiation of genes between a target and a reference array. To evaluate the performance of Sub-Sub normalization, we apply it to both simulated data sets and real data sets.

2.1 Background

As one of the commercial standards, Affymetric GeneChips® uses 11-20 probe pairs, which are short oligonucleotides of 25 bp, to represent each gene, and as a whole they are called a probe set [Aff01, LDB+96]. Each probe pair consists of a perfect match (PM) and a mismatch (MM) probe that differ only in the middle (13th) base. MM probes are designed to measure the non-specific binding. Ideally, probes are arranged on a chip in a random fashion. But in customized arrays, this is not always true. RNA samples are prepared, labeled, and hybridized with arrays. Then arrays are scanned and images are

produced and analyzed to generate an intensity value for each probe. These intensities represent how much hybridization occurred for each oligonucleotide. Up to now, each probe is represented by an intensity value. In order to obtain the expression level for a probe set, we need to combine the intensities of probes corresponding to it. This process is what so called summarization.

In many of the applications of high density nucleotide arrays, the goal is to seek the differentiation of mRNA expression among different samples. For example, the identification of differentially expressed genes in tumor with respect to normal tissue helps to understand the mechanism of cancers. The variations between samples that are informative are referred to as "interesting variation". However, the gene expression levels measured by microarrays also include variations introduced during the experiment processes: RNA extraction, fluorescence labeling, hybridization and scanning. These variations are referred to as "obscuring variation" [ZKM$^+$05]. Direct comparison of data from different arrays can lead to misleading results and incorrect conclusions due to the "obscuring variation". However, this effect will be alleviated if the arrays can be appropriately normalized [DYCS02, IBC$^+$03]. The purpose of normalization is to minimize the "obscuring variation" between arrays so that the expression levels of genes measured by different arrays are comparable. Therefore, normalization is one of the critical steps for microarray data analysis.

2.2 Normalization approaches for oligonucleotide array

Affymetrix developed a normalization method called scaling normalization, which scales the intensities so that each array has the same average intensity. It is performed after summarization in the probe set level. In the probe level (before summarization), several normalization approaches for oligonucleotide arrays have been proposed.

Among these normalization methods, "constant" simply carries out scaling normalization in the probe level [IBC+03]. The others can be roughly categorized into two classes. The first class, including "loes" and "contrasts", is based on M vs A methodology, which achieves normalization of a target array against a reference array by correcting for non-central and non-linear bias observed in M-A plot [BIApS03, Ast03]. The second class, such as "quantiles" and "qspline" [BIApS03, WJJ+02], correct for the nonlinear bias seen in Q-Q plot. In this section, we will introduce several normalization approaches that are most frequently used for oligonucleotide arrays.

2.2.1 Loess normalization

This approach is proposed by Dudoit et al. [DYCS02] which is originally applied to perform within slide normalization for the two color channels of cDNA array. It is based on the M vs A methodology where M is the difference in log expression values and A is the average of the log expression values. Bolstad et al. generalized this approach to normalize probe intensities from two arrays [BIApS03]. The underlying rationale is that very few genes will have different expressions in two arrays. So an M vs A plot for the normalized data should have a point cloud centered around the $M = 0$ axis.

For any two arrays i and j with probe k's intensities $x_k i$ and $x_k j$, where $k = 1, \cdots, p$, the M and A are defined as $M_k = \log_2(x_k i / x_k j)$ and $A_k = \log_2(x_k i x_k j)$. Loess, a local regression method [Cle79], is used to fit a normalization curve to the M vs A plot for the Ms and As of all probes. If the fitted M for probe k by the normalization curve are $\hat{M}_k$, then the normalization adjustment can be formulated as $M'_k = M_k - \hat{M}_k$. And the normalized probe intensities are given by $x'_k i = 2^{A_k + M'_k/2}$ and $x'_k j = 2^{A_k - M'_k 2}$. For a data set with more than two arrays, the normalization is carried out in a pairwise manner. To do a within slide normalization, the two channels are treated as two arrays.

2.2.2 Contrast normalization

This approach is first introduced by Astrand [Ast03]. It is also based on the M vs A methodology, but this method transforms the data into a set of contrasts before the normalization.

Suppose one has a data set with k arrays, where each array contains n probes. Let the $n \times k$ matrix Y denote the probe intensities of these arrays. First the data Y is transformed into a log scale and the basis is changed as the following: $Z = [x, y_1, \cdots, y_{k-1}] = \log Y \cdot M'$, where M is an orthonormal $k \times k$ matrix called transformation matrix. The first row of M is always the 1-vector times $\sqrt{1/k}$ and then it follows that the other rows are a set of orthonormal contrast. In the transformed basis, a normalization curves is fit using loess for each of the $n-1$ y_i with respect to x. The data is then adjusted by using a smooth transformation which adjusts the normalization curve so that it lies along the horizontal. Data in the normalized state is obtained by transforming back to the original basis and exponentiating.

2.2.3 Quantiles normalization

Quantile normalization was first introduced by Bolstad et al. in 2003 [BIApS03]. The goal of the method is to achieve the same distribution of probe intensities for each array in the data set. If two data vectors have the same distribution, the Q-Q plot of them is a straight diagonal line. This concept can be extended to n dimensions: if all n data vectors have the same distribution, then if we plot the quantiles of them in a n dimension space, we'll also get a straight diagonal line. Therefore, one could make a set of arrays have the same distribution of intensities by projecting the points of the n dimensional quantile plot onto the diagonal.

In practice, this is simply achieved by taking the mean quantile and substituting the value in the original data set by this mean quantile. To do this, one can use the following algorithm where X is the $n \times k$ matrix of the original probe intensities(n probes and k arrays):

1. Given k array each with n probes, form X of dimension $n \times k$ where each array is a column

2. Sort each column of X to give X_{sort}

3. Take the means across each row of X_{sort} and assign this mean to each element in the row to get X'_{sort}

4. Get $X_{normalized}$ by rearranging each column of X'_{sort} to have the same ordering as the original X

2.2.4 Qspline normalization

This approach is proposed by Workman et al. [WJJ+02]. It fits a smoothing B-spline between the quantiles of probe intensities on the array(x) and those on the reference array (v). The splines are then used as intensity-dependent normalization functions on the probe intensities of x. After the normalization, the probe intensities of all arrays share the same distribution with the reference array. The reference array can be any array in the data or the mean "array" calculated from multiple arrays.

2.2.5 Invariantset normalization

This approach was first used in the dChip software by Li et al. [LW01a, LW01b]. The normalization is based on a set of probes that belong to non-differentially expressed genes. This set of probes is called "invariant set". To identify the "invariant set", an

iterative procedure is applied. Specifically, one starts with all n PM probe in an array. If the probe's proportion rank difference(absolute rank difference in two arrays divided by n) is small enough, it is kept for the new set. In this way, a new set of probes is selected, and the same procedure is applied to the new set iteratively, until the number of probes in the new set does not decrease anymore. Then based on the invariant set, loess is used to fit a normalization curve to relate the reference array to an array to be normalized.

2.3 Introduction to Sub-Sub normalization

The development of our normalization method, Sub-Sub (**Sub**-array normalization **sub**ject to differentiation), was motivated by two important issues that must be considered in oligonucleotide array normalization: fraction of the differentially expressed genes and spatial effect of the arrays. Consistently, the first "Sub" in "Sub-Sub" aims to deal with the spatial effect by dividing the whole array into sub-arrays. The second "Sub" refers to "subject to differentiation", which means that our method allows for large fraction of differentially expressed genes.

2.3.1 Differentiation fraction

Among all the existing normalization methods introduced above, most of them have the following three assumptions about the data. First, most genes are not differentially expressed; Second, the number of up-regulated genes roughly equals the number of down-regulated genes; Third, the above two assumptions hold across the signal-intensity range. However in the reality, these assumptions are not always true. So we should consider normalization that is resistant to violation of these assumptions [BIApS03, ZAZL01, WJJ$^+$02].

When we compare two arrays in which a substantially large portion of genes are differentially expressed, we need to identify a "base" subset for the purpose of normalization. This subset should exclude those probes corresponding to differentially expressed genes and abnormal probes due to experimental variation. A similar concept "invariant set" has been defined in [SLEW01, TOR$^+$01, KCM02]. To identify the base subset, we use least trimmed squares (LTS) [RL87] which can estimate the transformation simultaneously. The substantially large portion of genes that are differentially expressed are protected in LTS by setting an appropriate trimming fraction. The exact LTS solution is computed by a fast and stable algorithm we developed recently [Li04].

2.3.2 Spatial pattern

Array-specific spatial patterns may exist due to uneven hybridization and measurement process. For example, reagent flow during the washing procedure after hybridization may be uneven; scanning may be non-uniform. We have observed different spatial patterns from one array to another. To take this into account, we divide each array into sub-arrays so that each of them consist of a few hundred probes. The probe intensities are normalized within each sub-array. Other spatial normalization methods such as that in [WJJ$^+$02] only consider the spatial effect in background. In comparison, we try to adjust for spatial effect both in background and in scale. We show that match of distribution at the array-level can be achieved by normalization at the sub-array level to a great extent. In cDNA arrays, local subgrid normalization has been proposed by [vKv$^+$03].

2.4 Methods

2.4.1 Statistical principle of normalization

Suppose we have two arrays: one reference and one target. Denote the measured fluorescence intensities from the target and reference arrays by $\{U_j, V_j\}$ respectively. Denote true concentrations of specific binding molecules by ($\tilde{U}_j$, $\tilde{V}_j$). Ideally, we expect that (U_j, V_j)=($\tilde{U}_j$, $\tilde{V}_j$). But in practice, measurement bias exists due to uncontrolled factors so we need a normalization procedure to adjust measurement. Now let's have another look at normalization. Consider a system with ($\tilde{U}_j$ $\tilde{V}_j$) as input and (U_j, V_j) as output. Let $\mathbf{h} = (h_1, h_2)$ be the system function that accounts for all uncontrolled biological and instrumental bias; namely,

$$\begin{cases} U_j &= h_1(\tilde{U}_j)\,, \\ V_j &= h_2(\tilde{V}_j)\,. \end{cases}$$

The goal is to reconstruct the input variables ($\tilde{U}_j$, $\tilde{V}_j$) based on the output variables (U_j, V_j). It is a blind inversion problem [Li03], in which both input values and the effective system are unknown. The general idea is to find a transformation that matches the distributions of input and output. This leads us to the question: what is the joint distribution of true concentrations ($\tilde{U}_j$, $\tilde{V}_j$)? First, let us assume that the target and reference array are biologically undifferentiated. Then the differences between the target and reference are purely caused by random variation and uncontrolled factors. In this ideal case, it is reasonable to assume that the random variables $\{(\tilde{U}_j, \tilde{V}_j), j = 1, \cdots\})$ are independent samples from a joint distribution $\tilde{\Psi}$ whose density centers around the straight line $\tilde{U} = \tilde{V}$, namely, $E(\tilde{V}|\tilde{U}) = \tilde{U}$. The average deviations from the straight line measures the accuracy of the experiment. If the effective measurement system $\mathbf{h}$ is not an identity one, then the distribution of the output, denoted by Ψ, could be different from $\tilde{\Psi}$. An appropriate estimate $\hat{\mathbf{h}}$ of the transformation should satisfy the following:

the distribution $\hat{\mathbf{h}}^{-1}(\Psi)$ matches $\tilde{\Psi}$, which centers around the line $\tilde{V} = \tilde{U}$. In other words, the right transformation straightens out the distribution of Ψ.

Next we consider the estimation problem. Roughly speaking, only the component of h_1 relative to h_2 is estimable. Thus we let $v = h_2(\tilde{v})$. In addition, we assume that h_1 is a monotone function. Denote the inverse of h_1 by g, then we expect the following to be valid.

$$E[\tilde{V}|\tilde{U}] = \tilde{U}, \quad \text{or} \quad E[V|g(U)] = g(U)\,.$$

Proposition 1 *Suppose the above equation is valid. Then g is the minimizer of* $\min_l E(V - l(U))^2$.

According to the well known fact of conditional expectation, $E[V|g(U)] = g(U)$ minimizes $E[V - l_1(g((U))]^2$ with respect to l_1. Next write $l_1(g(U)) = l(U)$. This fact suggests that we estimate g by minimizing $\sum_j (v_j - g(u_j))^2$. When necessary, we can impose smoothness on g by appropriate parametric or non-parametric forms.

2.4.2 Differentiation fraction and undifferentiated probe set

Next we consider a more complicated situation. Suppose that a proportion λ of all the genes are differentially expressed while other genes are not except for random fluctuations. Consequently, the distribution of the input is a mixture of two components. One component consists of those undifferentiated genes, and its distribution is similar to $\tilde{\Psi}$. The other component consists of the differentially expressed genes and is denoted by $\tilde{\Gamma}$. Although it is difficult to know the form of $\tilde{\Gamma}$ as *a priori*, its contribution to the input is at most λ. The distribution of the input variables $(\tilde{U}_j, \tilde{V}_j)$ is the mixture $(1-\lambda)\,\tilde{\Psi} + \lambda\,\tilde{\Gamma}$. Under the system function $\mathbf{h}$, $\tilde{\Psi}$ and $\tilde{\Gamma}$ are transformed respectively into distributions denoted by Ψ and Γ; That is, $\Psi = \mathbf{h}(\tilde{\Psi})$, $\Gamma = \mathbf{h}(\tilde{\Gamma})$. This implies that the distribution of the output (U_j, V_j) is $(1-\lambda)\,\Psi + \lambda\,\Gamma$. If we can separate the two components Ψ and Γ,

then the transformation $\mathbf{h}$ of some specific form could be estimated from the knowledge of $\tilde{\Psi}$ and Ψ.

2.4.3 Spatial pattern and sub-arrays

Normalization can be carried out in combination with a stratification strategy. For cDNA arrays, researchers have proposed to group spots according to the layout of array-printing so that data within each group share a more similar bias pattern. And then normalization is applied to each group. This is referred to as within-print-tip-group normalization. On a high-density oligonucleotide array, tens of thousands of probes are laid out on a chip. To take into account any plausible spatial variation in h, we divide each chip into sub-arrays, or small squares, and carry out normalization for probes within each sub-array. To get over any boundary effect, we allow sub-arrays to overlap. A probe in a overlapping regions gets multiple adjusted values from sub-arrays it belongs to, and we take their average.

2.4.4 Parameterization

Since each sub-array contains only a few hundred probes, we choose to parameterize the function g by a simple linear function $\alpha + \beta\, u$, in which the background α and scale β represent respectively uncontrolled additive and multiplicative effects.

2.4.5 Simple least trimmed squares

Our target solution consists of two parts: 1. identify the "base" subset of probes; 2. estimate the parameters in the linear model. We adopt least trimmed squares to solve the problem. Starting with a trimming fraction ρ, set $h = [n(1-\rho)] + 1$. For any (α, β), define $r(\alpha, \beta)_i = v_i - (\alpha + \beta\, u_i)$; Let $H_{(\alpha,\beta)}$ be a size-h index set that satisfies the

following property: $|r(\alpha,\beta)_i| \leq |r(\alpha,\beta)_j|$, for any $i \in H_{(\alpha,\beta)}$ and $j \notin H_{(\alpha,\beta)}$. Then the least trimmed squares estimate (LTS) minimizes

$$\sum_{i \in H_{(\alpha,\beta)}} r(\alpha,\beta)_i^2 \,,$$

The solution of LTS can be characterized by either the parameter (α, β) or the size-h index set H. It is this dual form that we find it ideal for our purpose. Statistically, LTS is a robust solution for regression problems. On one hand, it can achieve any given breakdown value by setting a proper trimming fraction. On the other hand, it has $\sqrt{n}$-consistency and asymptotic normality under some conditions. In addition, the LTS estimator is regression, scale, and affine equivariant [RL87]. Despite its good properties, LTS has not been widely used because no practically good algorithm exists to implement computation. Recently we developed a fast and stable algorithm to compute the exact LTS solution to simple linear problems [Li04]. On an average desktop PC, it solves LTS for a data set with several thousand points in two seconds.

A LTS solution naturally associates with a size-h index set. By setting a proper trimming fraction ρ, we expect the corresponding size-h set is a subset of the undifferentiated probes explained earlier. Obviously, the trimming fraction ρ should be larger than the differentiation fraction λ.

2.4.6 Multiple arrays and reference

In the case of multiple arrays, the strategy of normalization hinges on the selection of reference. In some experiments, a master reference can be defined. For example, the time zero array can be set as a reference in a time course experiment. In experiments of comparing tumor and normal tissues, the normal sample can serve as a reference. In other cases, the median array or mean array are options for references. Another strategy

is: first, randomly choose two arrays, one reference and one target, for normalization; use the normalized target array from the last normalization as the reference for the next normalization; iterate this procedure until all arrays have been normalized once; and repeat this loop for several runs. Hereafter we adopt the median polishing method in RMA [IBC+03] to summarize expression levels from multiple arrays.

The direct result of normalization is the calibration of relative expression levels of an array with respect to a reference. Suppose we have an ideal reference array with known concentrations of binding molecules for all probes. Then in theory, we can measure the absolute expression values of any sample as long as we can normalize its hybridization arrays with the reference.

2.4.7 Implementation and Sub-Sub normalization

We have developed a module to implement the normalization method describe above, referred as SUB-SUB normalization. The core code is written in C, and we have an interfaces with Bioconductor in R [Net, Bio]. The input of this program is a set of Affymetrix CEL files and output are their CEL files after normalization. Three parameters need to be specified: sub-array size, overlapping size and trimming fraction. The sub-array size specified the size of the sliding window. The overlapping size controls the smoothness of window-sliding. Trimming fraction specifies the break down value in LTS. The normalized CEL files generated by the program could be directly read in by the "affy" package in Bioconductor for further processing such as PM correction, summarization and so on. An experiment with an expected higher differentiation fraction should be normalized with a higher trimming fraction.

2.5 Evaluation of Sub-Sub on Spike-in data set

2.5.1 Affymetrix Spike-in data set

Affymetrix Spike-in data set provides a standard to evaluate the performance and effectiveness of a normalization method. This data set includes 14 groups of arrays, and each of them consists of 4 replicates, from Affymetrix HG-U95 chips. Fourteen genes are spiked-in on these arrays at given concentrations and in a cyclic fashion known as the Latin square design. The data are available from *http://www.affymetrix.com/support/technical/sample_data/datasets.affx*. Out of the 14 groups of arrays, we chose two groups. The first group contains four arrays: *1521m99hpp_av06, 1521n99hpp_av06, 1521o99hpp_av06* and *1521p99hpp_av06*. The second group also contains four arrays: *1521q99hpp_av06, 1521r99hpp_av06, 1521s99hpp_av06 and 1521t99hpp_av06*. Later we will abbreviate these arrays by M, N, O, P, Q, R, S, T. As a result, the concentrations of thirteen spiked-in genes are two-fold lower in the second group than the first group. The concentrations of the remaining spike-in gene are respectively 0 and 1024 in the two groups. In addition, two other genes are so controlled that their concentrations are also two-fold lower in the second group compared to the first one.

As we have claimed, one issue that Sub-Sub aims to deal with is the large differentiation fraction between RNA samples. To test the robustness of Sub-Sub normalization when there's a substantial differentiation fraction, we generate an artificial data set with relatively large fraction of differentiation by perturbing the HG-U95 spike-in data set. Namely, we randomly choose 20% genes and increase their corresponding probe intensities by 2.5 fold in the four arrays in the second group. We also generate other two perturbed Spike-in data sets with 1.5 and 1.25 fold increase.

2.5.2 Parameter selection

Before the normalization, three parameters need to be specified: sub-array size, overlapping size and trimming fraction. The sub-array size specified the size of the sliding window. The overlapping size controls the smoothness of window-sliding. Trimming fraction specifies the break down value in LTS. We will describe the effect of the three parameters one by one in the following context.

Sub-array size

Effect of sub-array size on Sub-Sub normalization is shown in Figure 2.1. We perform Sub-Sub normalization using different sub-array size that vary from 20×20 to 80×80. Overlapping size and trimming fraction are fixed to 0 and 80%, respectively. To make the results comparable, the same PM correction (PM only) and summarization (medianpolish) method are used after normalization. To do this, the "affy" package in Bioconductor is used. In Figure 2.1, M-A plots (log ratios of expressions between the two groups versus the abundance) are summarized from the eight arrays. For a given pair of arrays, each from a group, the M and A values for all probe sets are calculated. Pairwise comparison results in 16 M and A values for each probe set. These 16 M and A values are averaged and then used for M-A plot shown in Figure 2.1. Note that we will always use PM correction and medianpolish summarization throughout this chapter unless specified. As can be seen in Figure 2.1, the effect of sub-array size on Sub-Sub normalization is small. Sub-array sizes ranging from 20×20 to 80×80 result in similar M-A plots. In general, the smaller the sub-array size is, the more accurately we can capture the spatial bias while the less number of probes are left for estimation of linear relation. Thus, we need to trade off between bias and variation. From our experiments, a sub-array size from 20×20 to 80×80 works well for Affymetrix HG-U95 and HG-U133 chips.

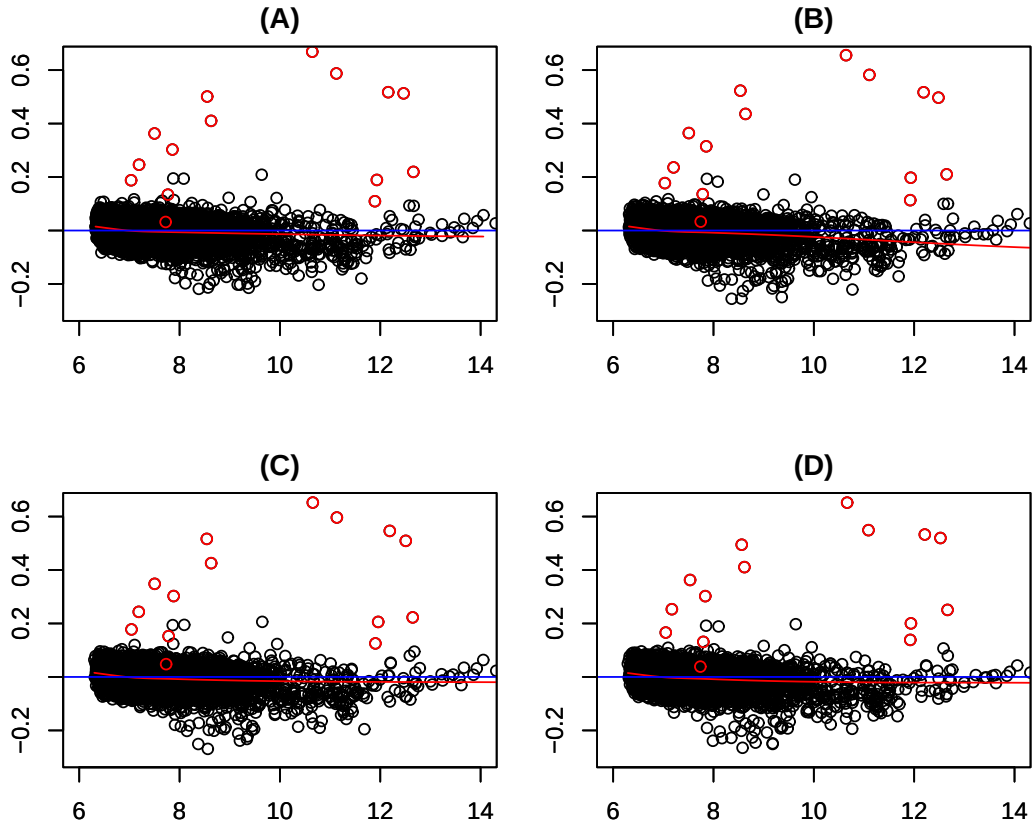

Figure 2.1: Effect of sub-array size on Sub-Sub normalization. M-A plots of Spike-in data are shown after Sub-Sub normalization:(A) sub-array size is 80×80; (B) sub-array size is 60×60; (C) sub-array size is 40×40; (D) sub-array size is 20×20; In all the cases, overlapping sizes are set to 0 and trimming fractions are set to 20%. Spike-in genes are shown in red.

Overlapping size

Effect of overlapping size on Sub-Sub normalization is shown in Figure 2.2. We fix the sub-array size and trimming fraction to be 20×20 and 20% respectively. Different overlapping sizes (0, 5, 10, 15) are used for normalization. As shown, we found the effect of the overlapping size on normalization is also small. Our recommendation is half of the sub-array size. For example, it is 10 if sub-array size is 20×20. According to

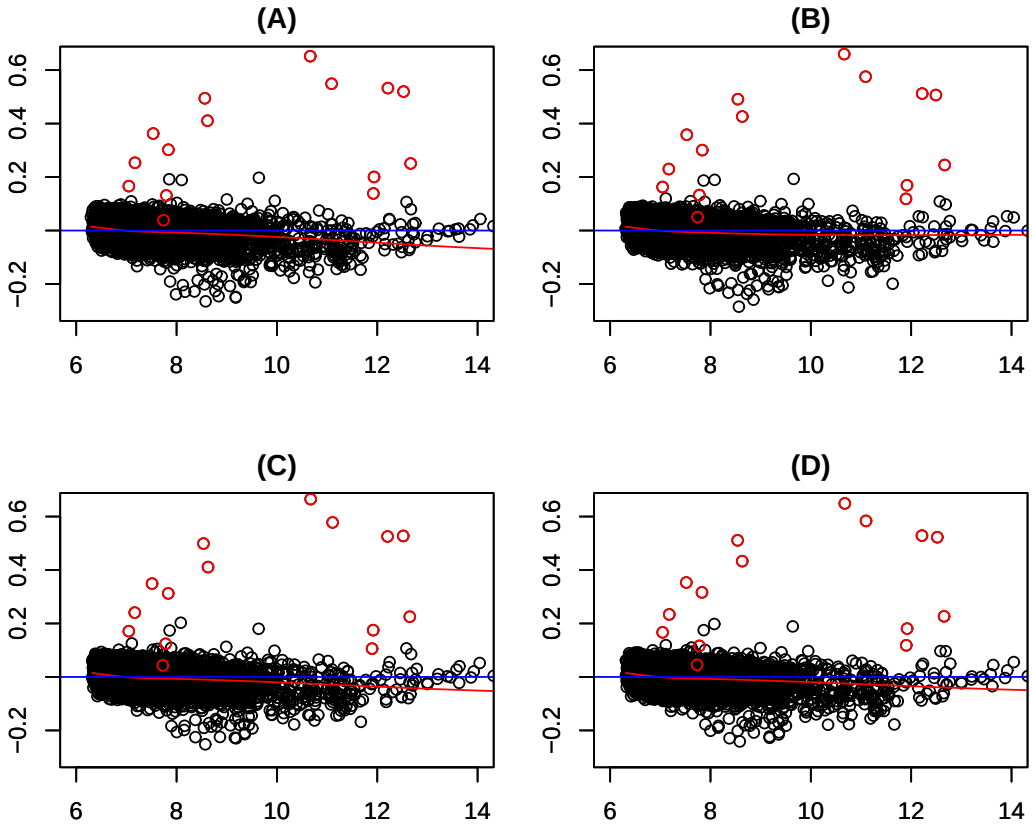

Figure 2.2: Effect of overlapping size on Sub-Sub normalization. M-A plots of Spike-in data are shown after Sub-Sub normalization: (A) overlapping size is 0; (B) overlapping size is 5; (C) overlapping size is 10; (D) overlapping size is 15. In all the cases, sub-array sizes are set to 20×20 and trimming fractions are set to 20%. Spike-in genes are shown in red.

our experience, it can even be set to 0 (no overlapping between adjacent sub-arrays) to speed up computation without obvious changing the normalization.

Trimming fraction

Effect of trimming fraction on Sub-Sub normalization is shown in Figure 2.3. A trimming fraction ranging from 0 to 30% is used while sub-array size and overlapping size

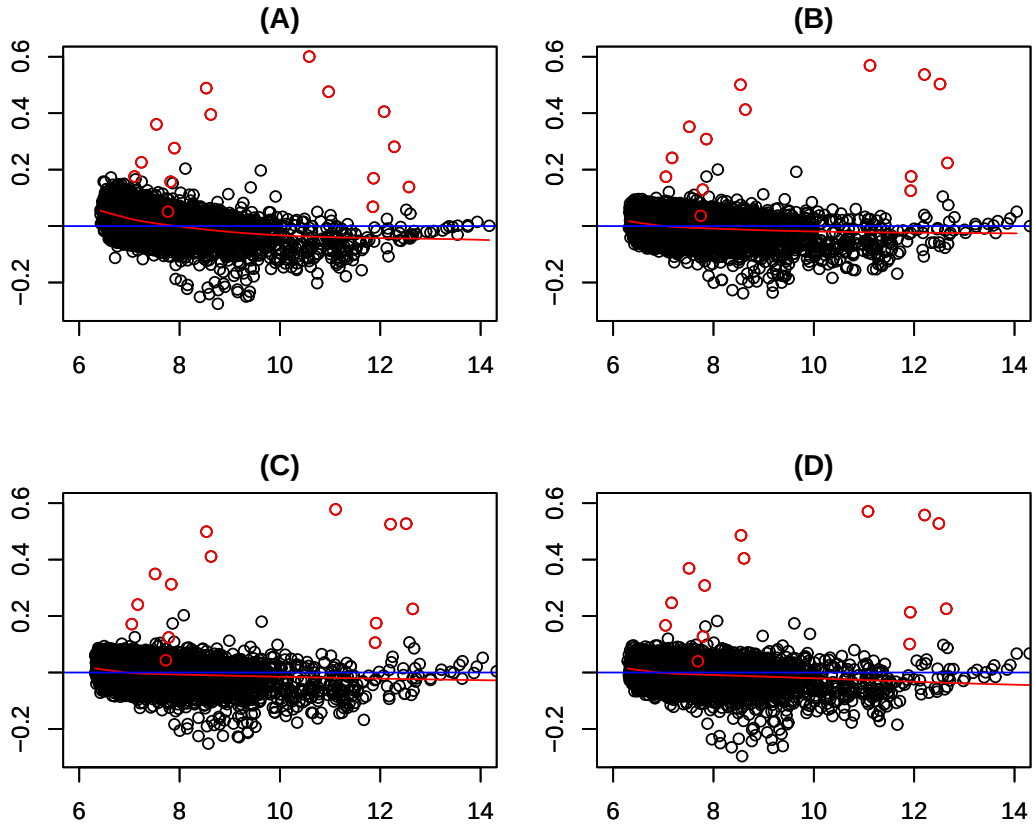

Figure 2.3: Effect of trimming fraction on Sub-Sub normalization. M-A plots of Spike-in data are shown after Sub-Sub normalization:(A) trimming fraction is 0; (B) trimming fraction is 10%; (C) trimming fraction is 20%; (D) trimming fraction is 30%. In all the cases, sub-array sizes are set to 20×20 and overlapping sizes are set to 10. Spike-in genes are shown in red.

are fix to 20×20 and 10, respectively. As can be seen in Figure 2.3, when the trimming fraction is 0, an obvious nonlinear pattern can be observed in the M-A plot, which gives a points cloud with a "banana" shape (Figure 2.3A). In this case, the LTS degenerates into an ordinary linear regression method, which is not robust to outliers any more. As a consequence, accurate estimation of linear relations in each sub-array can not be guaranteed. When we gradually increase the trimming fraction, the nonlinear pattern is removed from the M-A plots(see 2.3B-D). The selection of trimming fraction should

depend on which samples to be compared in the experiments and the quality of microarray data. For an experiment with 20% differentiated genes, we should set a trimming fraction larger than 20%. Again we need a trade off between robustness and accuracy when selecting the trimming fraction. On one hand, to avoid break-down of LTS, we prefer large trimming fractions. On the other hand, we want to keep as many probes as possible to achieve accurate estimates of α and β. Without *a priori*, we can try different trimming fractions and look for a stable solution. We recommend 50% to be the starting value for the try.

In the Affymetrix Spike-in data set, the majority of genes have constant expression levels across all the arrays. Trimming fraction is mainly used to protect the ill hybridized probes rather than probes corresponding to differentially expressed genes. Thus a relatively small trimming fraction of 20% is enough to achieve a good result.

It should be noted that the effects of the three parameters: sub-array size, overlapping size and trimming fraction, are under separate investigation in above sections. Also we have tried many combinations of these three parameters on several data sets. In some reasonable range, the interaction between the parameters is negligible. Our results indicate that the trimming fraction matters substantially to the normalization. The selection of sub-array size is relatively flexible. Effective normalization could be expected for a large range of sub-array size such as from 10×10 to 80×80. On the other hand, dividing array into sub-arrays is required to deal with the spatial effect. As a matter of fact, stratification by spatial neighborhood and selection of break down value in LTS do contribute a great deal to the normalization. Overlapping size has a little contribution in this data set.

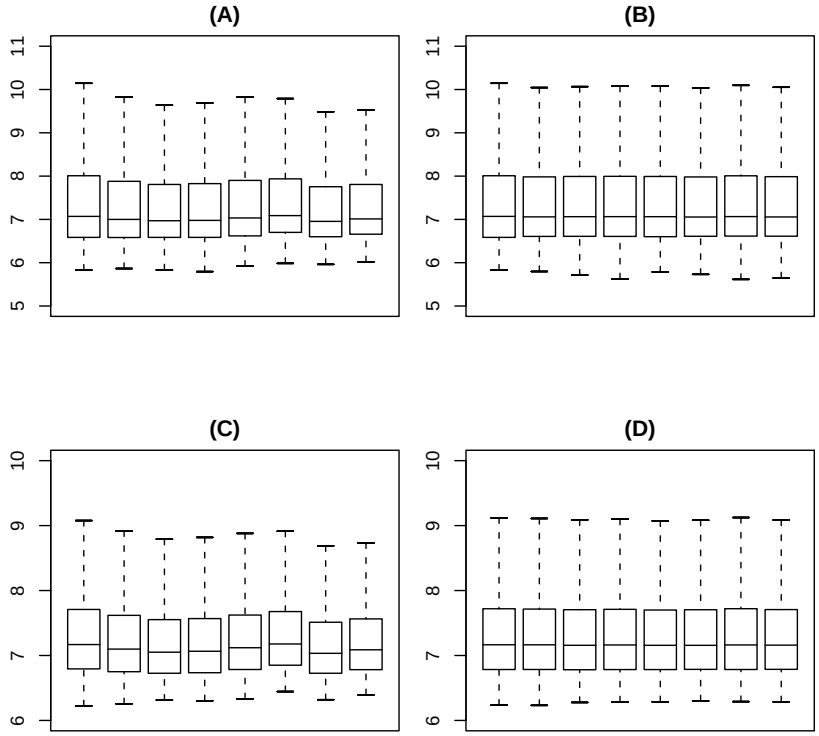

Figure 2.4: Box-plots of log transformed expression measurements for the 8 arrays from Spike-in data set in probe and probe set level before and after Sub-Sub normalization. (A) box-plot of probe intensities before normalization. (B) box-plot of probe intensities after Sub-Sub normalization. (C) box-plot of probe set expression values before normalization. (B) box-plot of probe set expression values after Sub-Sub normalization.

2.5.3 Global assessment of normalization

Figure 2.4A and 2.4B show the box plots of log transformed probe intensities before and after Sub-Sub normalization, respectively. Before normalization, the probe intensities from the eight arrays are different with each other. For example, the probe intensities from the 6th array are generally higher than those from other arrays. Obviously, this is an artificial result caused by "obscuring " variation between arrays, since we know that expression levels of the majority of genes are the same on the eight arrays except for the 14 spike-in genes. After Sub-Sub normalization, all the arrays have almost the

same median of intensities. That is, the "obscuring" variation between arrays has been significantly reduced so that expression levels after the normalization can be compared between different arrays. The effectiveness of Sub-Sub normalization to reduce "obscuring" normalization is also shown in the probe set level (see Figure 2.4C and 2.4D).

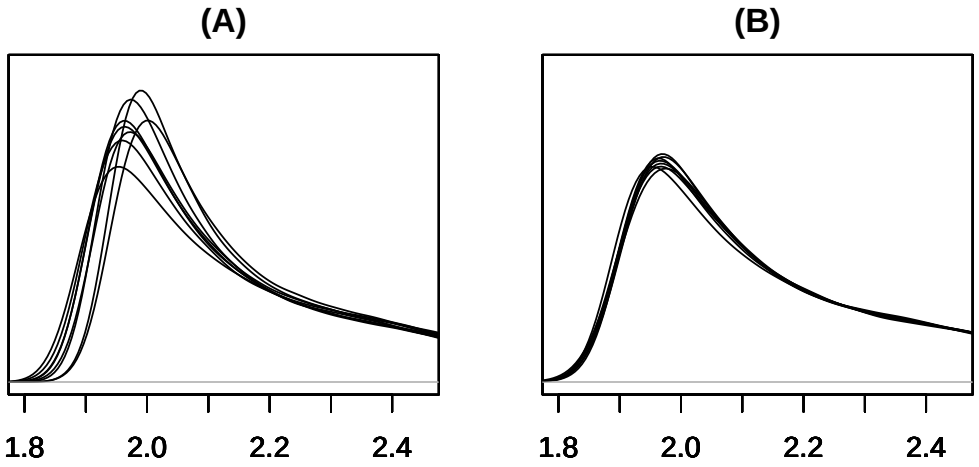

Figure 2.5: Distribution of probe intensities on the 8 arrays from Spike-in data before (A) and after (B) Sub-Sub normalization.

The distribution of probe intensities before and after Sub-Sub normalization for each array are shown in Figure 2.5. The probe intensities from the eight arrays have different distributions before normalization (see Figure 2.5A). Sub-Sub normalization results in similar distributions for probe intensities from all arrays. Although the Sub-Sub normalization doesn't attempt to match the marginal distributions purposely as the "quantile" normalization does, it does achieve similar marginal distributions between arrays.

The effectiveness of Sub-Sub normalization on the Spike-in data set is also revealed by the M-A plots. Figure 2.6A and 2.6B show the M-A plots before and after the Sub-Sub normalization, respectively. Sub-Sub normalization removes the non-linear pattern seen in the M-A plot. After Sub-Sub normalization, the point cloud centers around the horizontal $M = 0$, which is what we expect to achieve by normalization.

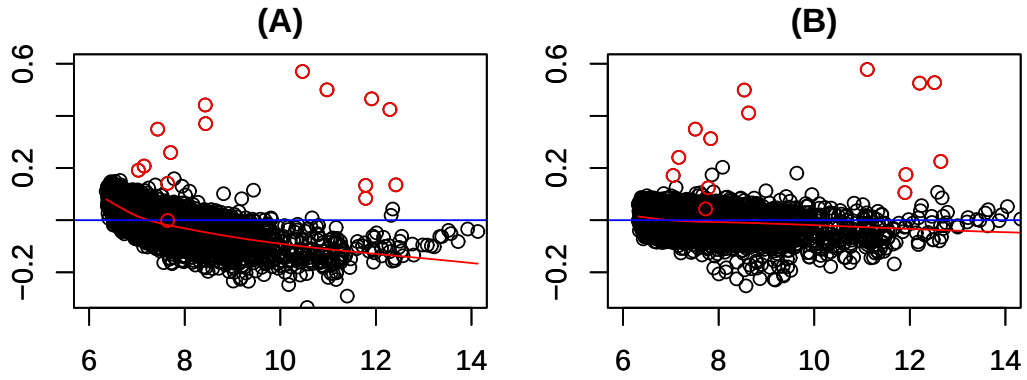

Figure 2.6: M-A plots for the Spike-in data set before (A) and after (B) Sub-Sub normalization.

2.5.4 Detection of spatial patterns

We then investigate the existence of spatial pattern. The HG-U95 chips have 640×640 spots on them. We divided each array into sub-arrays with a size of 20×20. We run simple LTS regression on the target with respect to the reference for each sub-array. This results in an intercept matrix and a slope matrix of size 32×32, representing the spatial difference between target and reference in background and scale. We first take Array M as the common reference. In Figure 2.7, the slope matrices of Array P and M are shown in the subplots at top left and top right respectively. Their histograms are shown in the subplots at bottom left and bottom right. Two quite different patterns are observed. Similar phenomenon exists in patterns of α. The key observation is that spatial patterns are array-specific and unpredictable to a great extent. This justifies the need of adaptive normalization.

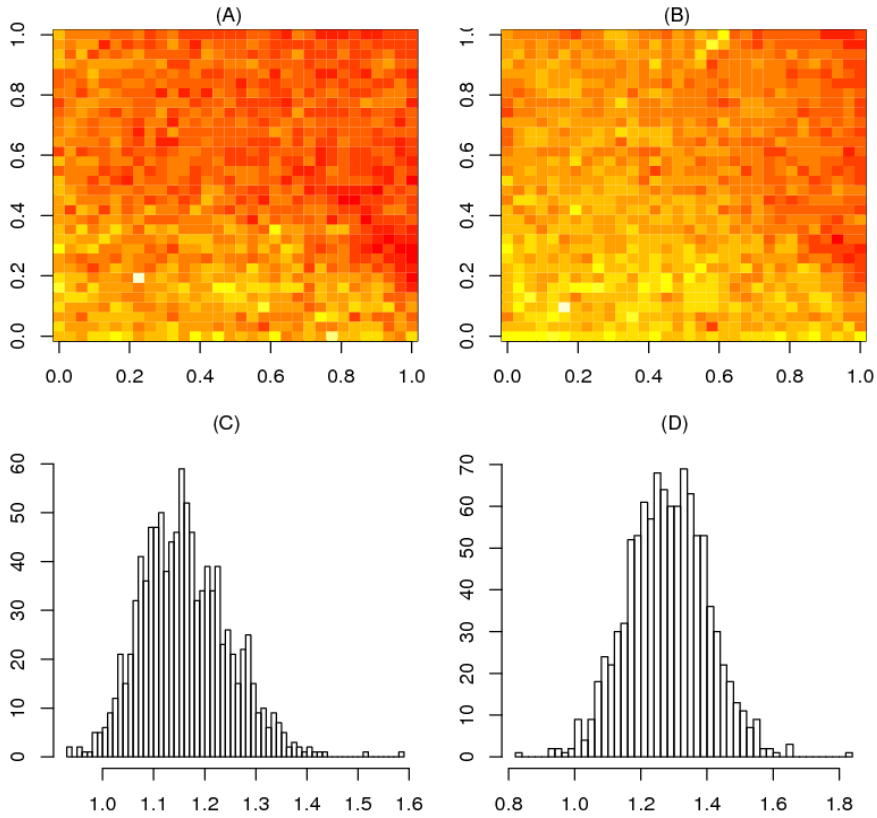

Figure 2.7: The slope matrices of two arrays show different spatial patterns in scale. The common reference is Array M. (A) Array P versus M; (B) Array N versus M. Their histograms are shown at bottom in (C) and (D) correspondingly.

2.5.5 Robustness to large differentiation fraction

As we have mentioned, one of the motivations to design Sub-Sub normalization is to deal with the differentiation of genes between samples. Sub-Sub normalization protects substantial differentiation genes by selecting an appropriate trimming fraction in LTS.

To test this, we generate several artificial data sets with relatively large fraction of differentiation by perturbing the Affymetrix HG-U95 Spike-in data set. We randomly choose 20% genes and increase their corresponding probe intensities by n fold(n=2.5, 1.5, and 1.25) in the four arrays of the second group. We then run SUB-SUB normalization on the perturbed data set with various trimming fractions. The results are shown in Figure **??** for four trimming fractions, 30%, 20%, 10%, and 0%. The normalization is satisfactory when the trimming fraction is less than 20% (see Figure 2.8A and 2.8B). When the trimming fraction is larger than 20%, the real differentiation fraction, Sub-Sub does not achieve a good normalization as revealed by the "banana" shaped point cloud in the M-A plot (see Figure 2.8). Again, this suggests the importance of choosing an appropriate trimming fraction for LTS.

We have also tried the other two perturbed Spike-in data set with 1.5 and 1.25 fold up-regulation for 20% randomly choose genes. Similar results are obtained as shown in Figure 2.9. These results indicate that Sub-Sub normalization is effective for data sets with large differentiation fractions as long as an appropriate trimming fraction is chosen.

2.6 Evaluation of Sub-Sub on real data sets

2.6.1 Microarray data sets

Yeast sir2△/wt data

The data set was introduced by Fabrizio et al. [FGB$^+$05]. To study the function of Sir2 in yeast ageing process, RNA samples were extracted from sir2△ and wild type strain in duplication, and hybridized with Affymetrix YG-S98 chips. This leads to four arrays,

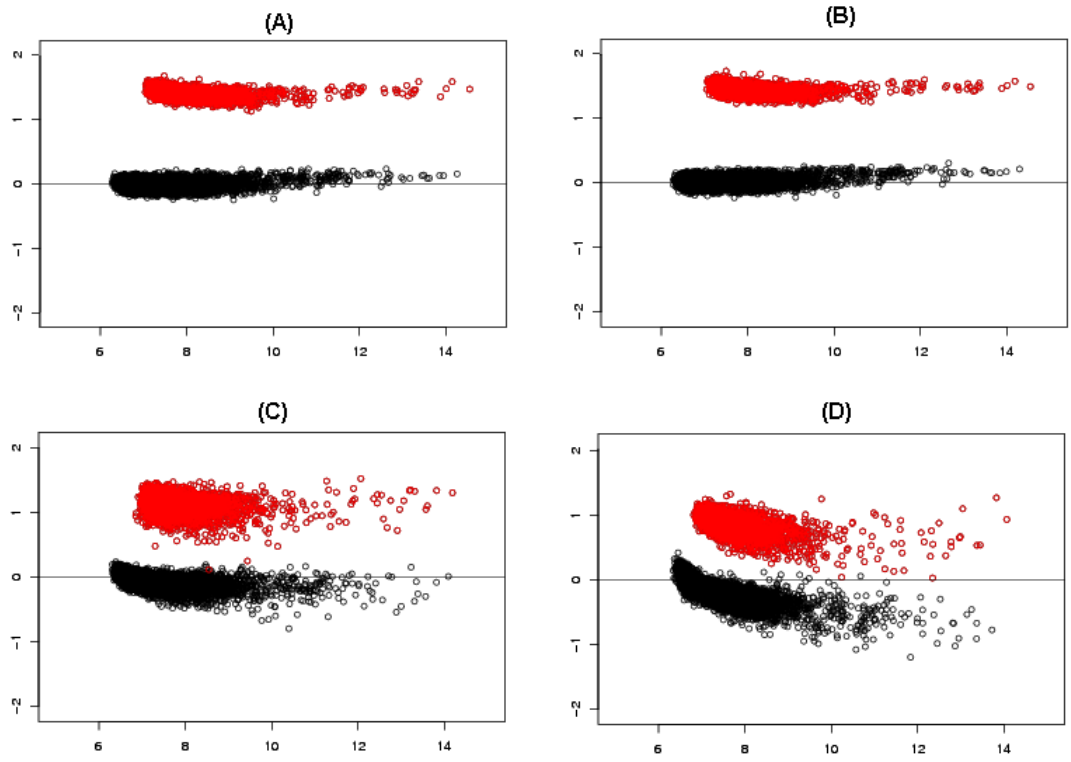

Figure 2.8: M-A plots for perturbed Spiked-in data set (n=2.5) after Sub-Sub normalization. 20% randomly selected genes are artificially up-regulated by 2.5 fold in Array Q, R, S and T. The differentially expressed genes are marked red, and un-differentially expressed genes are in black. The trimming fraction in the subplots are (A) 30%; (B) 20%; (C) 10%; (D) 0%.

two corresponding to sir2$\triangle$ and the other two corresponding to wild type. The YG-S98 chip has 534×534 spots on it. We will use this data set as an example of gene differentiation.

Yeast technical replicates data

The data set was downloaded from Affymetrix web site as an sample data set [Aff]. It includes two technical replicates of YG-S98 arrays: Yeast-2-121501 and Yeast-2-121502. Technical replicates are obtained by hybridizing the same RNA sample to

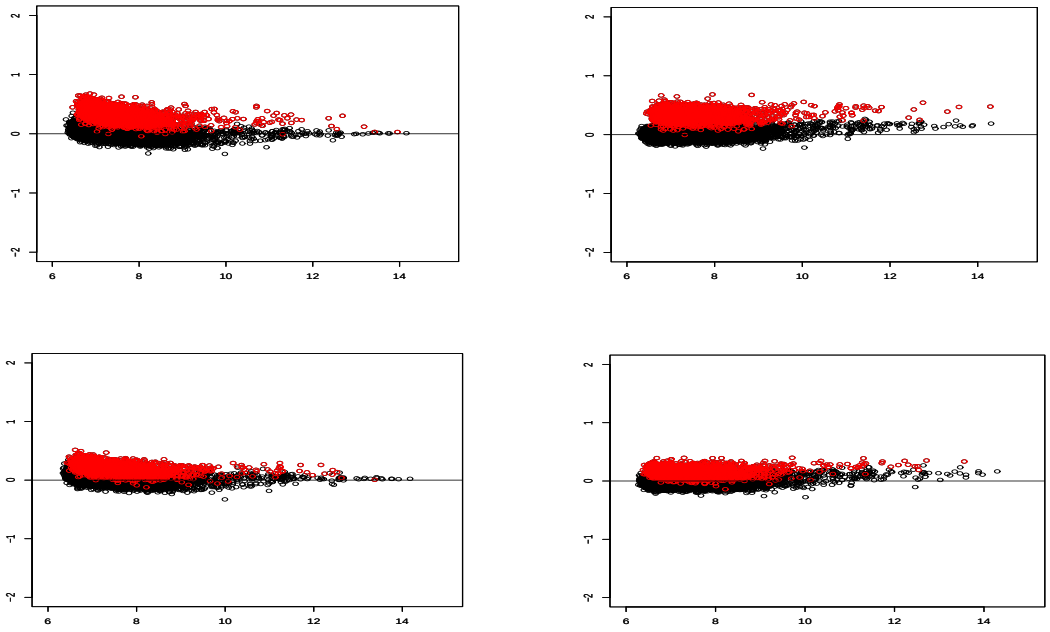

Figure 2.9: M-A plots of perturbed spike-in data set (n=1.5 and n=1.25) after Sub-Sub normalization. 20% randomly selected genes are artificially up-regulated by 1.5 fold (A and B) and 1.25 fold (C and D) in Array Q,R,S and T. The trimming fractions are: (A) 0%; (B) 30%; (C) 0%; (D) 30%.

multiple arrays. The "obscuring" variations can only be introduced after hybridization. So we expect no biological variation between technical replicates since they are from exactly the same sample. Technical replicates are different from what are so called biological replicates. The latter ones are hybridization results of different RNA samples that are prepared separately from the same biological sample, i.g. a tumor sample from a patient. Technical replicates enable us to compare the performance of normalization methods by measuring the variation reduction.

Primate brain expression data

Expression profiles offer a way to study the difference between humans and their closest evolutionary relatives. Unfortunately, gene chips are only available commercially for a

limited number of species. For example, there is still no commercial gene chip for chimpanzee. To measure the expression profiles for chimpanzees, we have to do cross species hybridization, that is, hybridize chimpanzee RNA samples with human chips. The Primate brain expression data is one of this type of data which is available from *http://email.eva.mpg.de/∼khaitovi/supplement1.html* [EKK+02]. Two brain samples are extracted from each of three humans, three chimpanzee and one orangutan. In what follows we only show results on two human individuals (HUMAN 1 and HUMAN 2), one chimpanzee (CHIMP. 1) , and the orangutan (ORANG). The mRNA expression levels were measured by hybridizing them with the Affymetrix human chip HG-U95.

2.6.2 Results

Example of differentiation

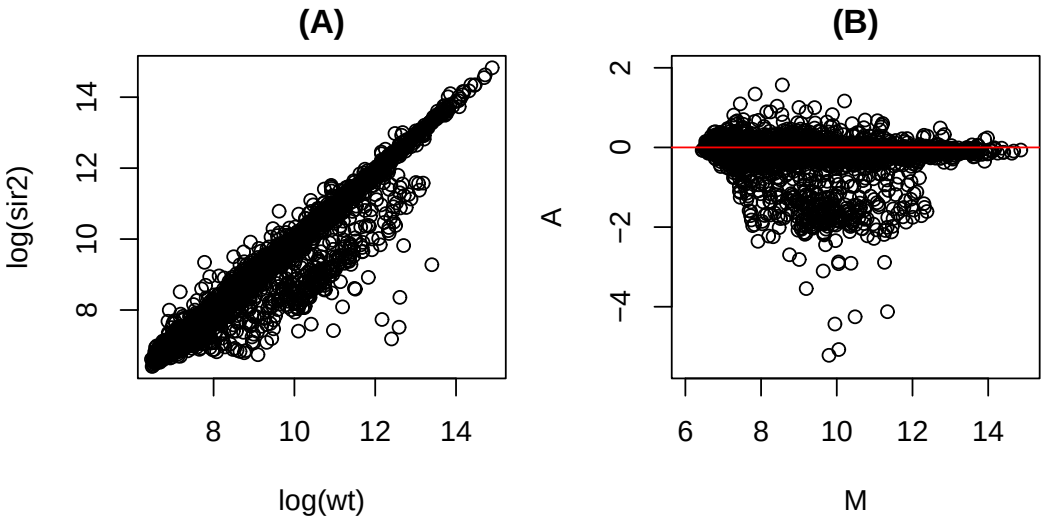

Figure 2.10: An example of gene differentiation. (A) Scatter plot of log transformed expressions for probe sets in wild type versus those in sir2△. (B) The corresponding M-A plot.

Unlike most of the other existing normalization methods, Sub-Sub does not assume that the majority of genes are not differentially expressed. That is, Sub-Sub allow a fraction of genes to be differentially expressed between arrays. In addition, the number of up-regulated genes and the number of down-regulated genes are not assumed to be equal. In section 2.5, we have shown the effectiveness of Sub-Sub to deal with data set that has substantial differentiation fraction using the perturbed Affymetrix Spike-in data set. In the perturbed Spike-in data, we simulate a situation in which a certain fraction of genes are differentially expressed, while log ration of the other genes are close to 0. Consequently, two clusters appears in the point cloud as shown in the M-A plot.

One may ask can this appear in a real microarray data set? The answer is"yes". As shown in Figure 2.10, Yeast sir2$\triangle$/wt data set provides us a good example. Figure 2.10A shows the scatter plot of log transformed expressions for probe sets in wild type versus those in sir2$\triangle$. Obviously, there are two clusters that appear in the scatter plot. The major cluster corresponds to genes that are not differentially expressed in sir2$\triangle$ with respect to wild type. Whereas, the other cluster corresponds to differentially expressed genes. These two clusters can be observed more easily in the M-A plot as shown in Figure 2.10B. We investigate the gene cluster that is down-regulated in sir2$\triangle$. It turns out that most of these genes are involved in the yeast pheromone pathway [RNM$^+$00, WD04]. Thus deletion of sir2 results in the repression of the pheromone pathway. From another point of view, this example indicates that it is reasonable for Sub-Sub normalization to protect differentially expressed genes and outliers using LTS.

Variation reduction by Sub-Sub normalization

Stratification is a statistical technique to reduce variation. Sub-array normalization can be regarded as a way of stratification. We normalize the yeast array 2-121502 versus

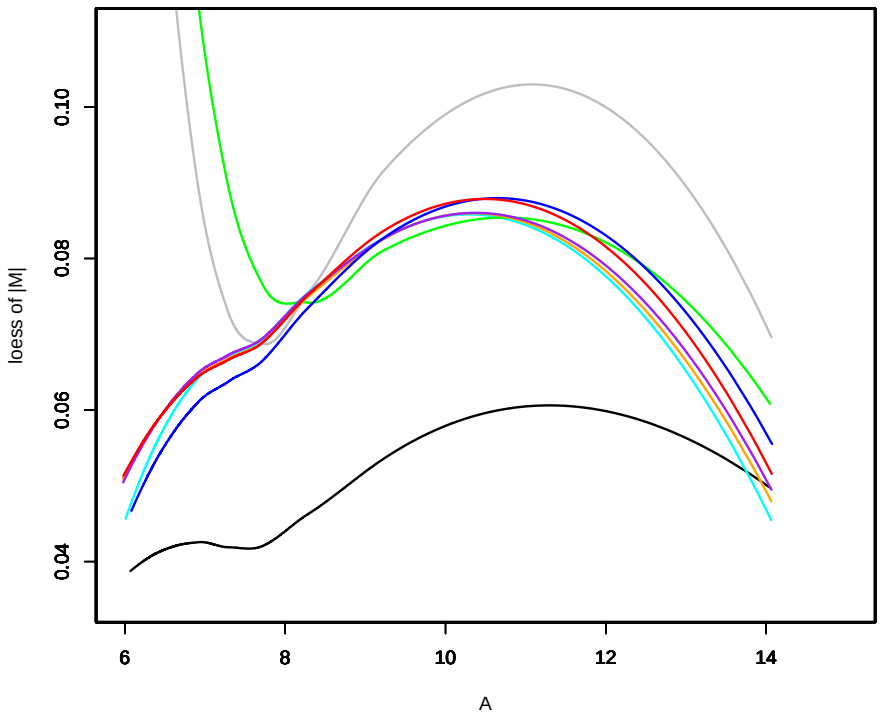

Figure 2.11: The lowess curves of $|M|$ versus A values by various normalization methods. Gray: no normalization; black: sub-sub; red: quantiles; green: constant; purple: contrasts; blue: invariant-set; orange: loess; cyan: qspline. In Sub-Sub, sub-array size, overlapping size and trimming fraction are set to 20×20, 10 and 20%, respectively.

2-121501 by various normalization methods available from "affy" package in Bioconductor. Since the two arrays are technical replicates, the difference between them is due to experimental variation. In the resulting M-A plots, we fit lowess [Cle79] curves to the absolute values of M, or $|M|$. These curves measure the variation between the two arrays after normalization, see Figure 2.11. The sub-array normalization achieves the minimal variation. Since variation is reduced, signal to noise ratio is enhanced and power of significance tests is increased.

Generally, a smaller sub-array size captures more spatial bias and therefore leads to more variation reduction in Sub-Sub normalization. Figure 2.12 shows the effect of sub-array size on variation reduction. As can be seen, with the decrease of sub-array size, more variation reduction is achieved.

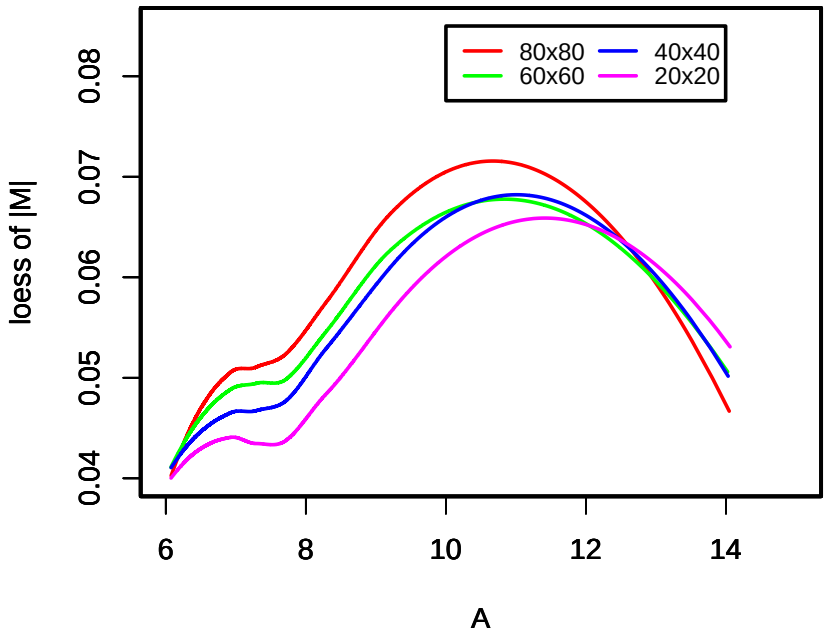

Figure 2.12: Effect of sub-array size on variation reduction. Trimming fraction and overlapping size are set to 20% and 0, respectively.

Trimming fraction is also important for variation reduction in Sub-Sub normalization. On one hand, a trimming fraction must be large enough to protect the differentially expressed genes and outliers in the data. On the other hand, larger trimming fraction results in less number of probes left for estimation in LTS. Thus, we need to trade off between bias and variation. Definitely, there is no differentiation between the yeast technical replicates. So a small trimming fraction should be used for Sub-Sub normalization. As expected, the variation decreases gradually as the trimming fraction increase from 10% to 50% (see Figure 2.13). However, a non-zero trimming fraction is required to protect the influence of outliers in the data. So as shown, if a trimming fraction of 0% is

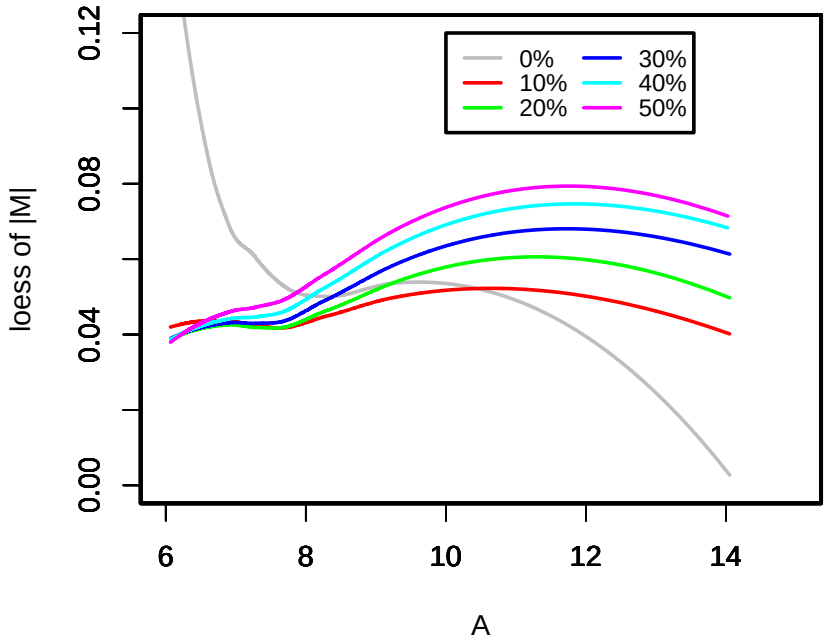

Figure 2.13: Effect of trimming fraction on variation reduction. Sub-array size and overlapping size are set to 20×20 and 10, respectively.

used, we do not achieve consistent variation reduction across the whole intensity range. In the range with small intensities, the worst variation reduction is obtained.

Primate brain expression data

Compared to other primate brains such as chimpanzee and orangutan, a relatively high percentage of genes are differentially expressed in human brains, and most of them are up-regulated in human brains [CLZ+03, GG03]. Moreover, the chimpanzee and orangutan samples are hybridized with human HG-U95 chips, so it is reasonable to assume: if there were any measurement bias in primate mRNA expressions compared

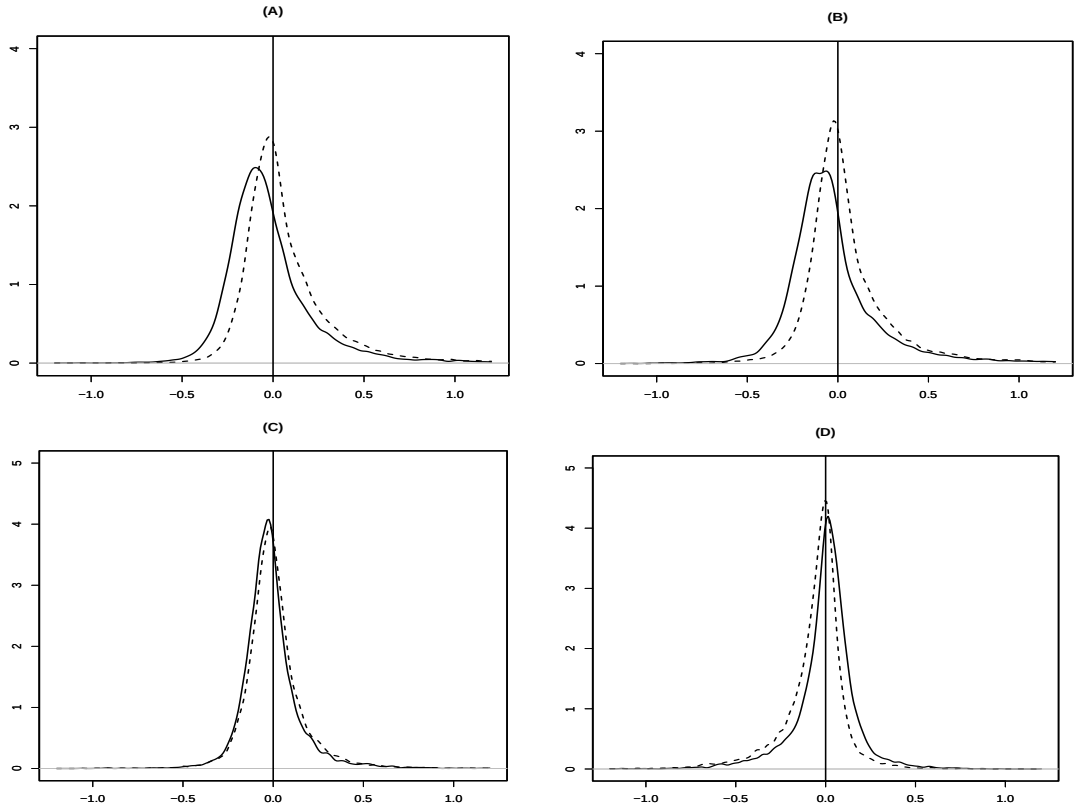

Figure 2.14: The densities of expression log-ratios between: (A) HUMAN 1 versus ORANG.; (B) HUMAN 2 versus ORANG.; (C) HUMAN 1 versus CHIMP. 1; (D) HUMAN 1 versus HUMAN 2. The results from SUB-SUB normalization (trimming fraction is 20%) and quantile normalization are represented by dotted and solid line respectively.

to humans, it would be downward bias. Figure 2.14 shows the density functions of log-ratios of gene expressions for four cases: HUMAN 1 versus ORANG.; HUMAN 2 versus ORANG.; HUMAN 1 versus CHIMP. 1 and HUMAN 1 versus HUMAN 2. In Figure 2.14, the density curves of the normalized densities by SUB-SUB(trimming fraction is 20%) and by quantile normalization are plotted in dotted and solid line respectively. When comparing humans with primates, the distribution from the SUB-SUB method shifts to the right than that from the quantile method. This is more obvious in the cases of humans versus orangutan, which are more genetically distant from each other than other cases do; see Figure 2.14A and Figure 2.14B. As expected, the distributions skew

to the right and the long tails on the right might have a strong influence on the quantile normalization, which aims to match marginal distributions from humans and primates in a global fashion.

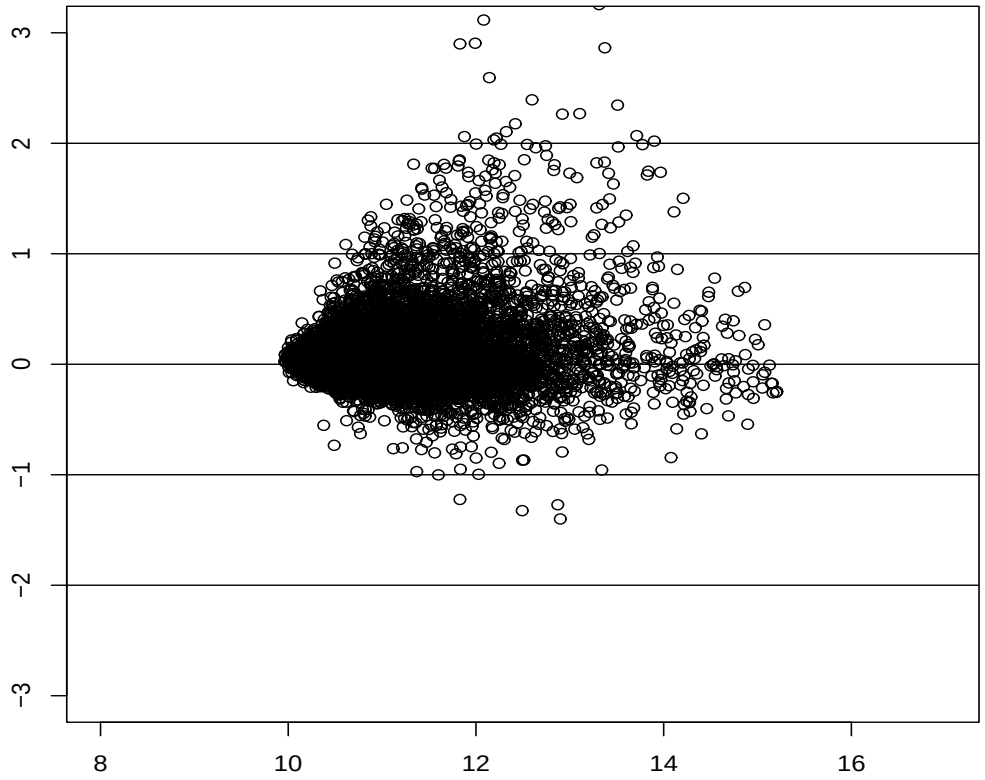

Figure 2.15: M-A plot of HUMAN versus ORANG after normalization. The sub-array size, overlapping size and trimming fraction are set to 20×20, 10 and 30% for Sub-Sub normalization, respectively.

However, in the cases of HUMAN 1 versus ORANG. and HUMAN 2 versus ORANG., the modes corresponding to the quantile method are in the negative territory while the modes corresponding to SUB-SUB method are closer to zero. The results from SUB-SUB normalization seems to be more reasonable. Furthermore, the difference in the case of HUMAN 1 versus HUMAN 2 is more distinct than that in the case of HUMAN 1 versus CHIMP. 1; see the two subplots at the bottom in Figure 2.14. The analysis in [EKK$^+$02] also indicates that HUMAN 2 differs more from other human

samples than the latter differ from the chimpanzee samples. We checked the M-A plot of HUMAN2 versus ORANG after SUB-SUB normalization (see Figure 2.15), and observed that HUMAN 2 has more up-regulated genes than down-regulated genes compared to ORANG.

2.7 Discussion

2.7.1 External controls

In cDNA arrays, some designs use external RNA controls to monitor global messenger RNA changes, [vKv$^+$03]. In our view, external RNA controls play the role of undifferentiated probe sets. To carry out local normalization, we need a quite large number of external controls for each subgrid. In current Affymetrix arrays, this is not available.

2.7.2 Differentiation fraction

In many microarray experiments, the primary goal is to identify differentially expressed genes. But the differentiation fraction may be quite different from one case to another. Following are three cases in which a large fraction of genes may be differentially expressed between two samples. First, in the study on the life span of yeast, we compare expression profiles of a wild type strain with another such as $sch9\triangle$. The metabolism in the knock-out strain is greatly reduced and this leads to life span extension [FPP$^+$01]. Second, gene chips for some organisms are not available. And cross-species hybridization is a useful strategy for comparative functional genomics. The comparison of brain expressions of humans versus primates discussed earlier is one such example. Third, to reduce the cost, some customized arrays are designed to include only probes of hundreds of genes that are related to a specific biological pathway. SUB-SUB normalization

uses LTS to identify a ”base” subset of probes for adjusting difference in background and scale. In theory, the method can be applied to microarray experiments with differentiation fractions as high as 50%. In addition, our method does not assume an equal percentage of up- and down-regulated genes. In the mean time, LTS keeps the statistical efficiency advantage of least squares.

2.7.3 Non-linear array transformation versus linear sub-array transformation

To eliminate the non-linear phenomenon seen in M-A plots or Q-Q plots, methods such as *lowess*, qspline and quantile normalization use non-linear transformation at the global level [WJJ$^+$02, BIApS03, YDL$^+$02]. In comparison, we apply a local strategy in SUB-SUB normalization. One array is split into sub-arrays and a simple linear transformation is fitted for each sub-array. With an appropriate sub-array size and trimming fraction, the nonlinear feature observed in M-A plots is removed by linear sub-array transformation to a great extent. We speculate that the nonlinear phenomenon is partially caused by spatial variation. One simulation study also supports this hypothesis, but further investigation is required. Next we give one remark regarding nonlinearity. In normalization, we adjust the intensities of a target array compared to those of a reference. Even though the dye effect is a nonlinear function of spot intensities, a linear transformation may be a good approximation as long as the majority of probe intensities from the target and reference are in the same range and thus have similar nonlinear effect. Occasionally when the amount of mRNA from two arrays are significantly different, slight nonlinear pattern is observed even after sub-array normalization. To fix the problem, we can apply global lowess after the sub-array normalization. Alternatively, to protect the substantial differentiation, we can apply a global LTS normalization subject to a differentiation fraction once more.

2.7.4 Transformation

The variance stabilization technique was proposed in relation to normalization [DHHR02, HHS+02]. We have tested SUB-SUB normalization on the log-scale of probe intensities, but the result is not as good as that obtained on the original scale. After normalization, a summarization procedure reports expression levels using the probe intensities. we have tried the median polishing method [IBC+03] on the log-scale. Alternatively, we can do a similar job on the original scale using MBEI [LW01b].

2.7.5 Usage of mis-match probes

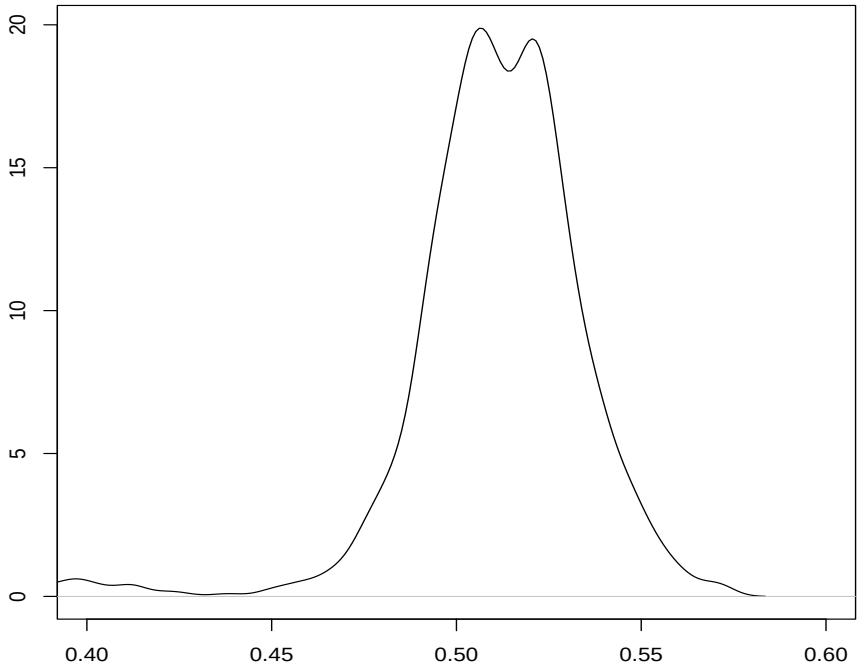

Figure 2.16: Histogram of percentages of MM probes in subsets associated with LTS.

Some studies suggested only using perfect-match probes in Affymetrix chips [WI04]. We checked the contribution of mismatch probes and perfect match probes to the subsets associated with LTS regressions from all sub-arrays. Figure 2.16 shows the distribution of the percentage of mis-match probes in the subsets identified by LTS. Our

result shows that mis-match probes contribute slightly more than perfect match probes in LTS regression, mostly in the range 46-56%.

2.7.6 Diagnosis

The detection of bad arrays is a practical problem in the routine data analysis of microarrays. In comparison with the obvious physical damages such as bubbles and scratches, subtle abnormalities in hybridization, washing and optical noise are more difficult to detect. By checking the values of α and β in LTS across sub-arrays, we can detect bad areas on one array and save information from the rest of the areas. Consequently, we can report partial hybridization result instead of throwing away an entire array.

2.8 Improve performance of Sub-Sub by PLTS

2.8.1 Limitation of LTS

LTS is the basis of Sub-Sub normalization. In each sub-array, LTS is performed to estimate the normalization relation between a target and a reference array. However, LTS has some drawbacks in nature when used for normalization. Basically, LTS is a robust method to solve linear regression problems. For a simple linear regression model:$y = \alpha + \beta x + \varepsilon$ with n observation (x_i, y_i), if we denote the squares residuals in an ascending order by $|r^2(\alpha, \beta)\_(1) \leq |r^2(\alpha, \beta)\_(2) \leq \cdots |r^2(\alpha, \beta)\_(n)$. Then the LTS estimation of coverage h, $\hat{\alpha}, \hat{\beta}$, are obtained by

$$\min_{\alpha,\beta} \sum_{i=1}^{\lfloor nh \rfloor} |r^2(\alpha, \beta)|_{(i)}.$$

As shown, LTS estimates the regression coefficients based only on data points with the smallest residuals, so it is robust to the outliers. That is, it achieves a more accurate

estimation of the linear relation between x and y, when there are some outliers included in the data.

In an ordinary LTS, vertical offsets are used: $r(\alpha, \beta) = y_i - \alpha - \beta x_i$. The vertical offset only takes the errors from response variable y into account. Consequently, extreme values in x_is and y_is are not equally treated. Extreme values in x_is are ignored in some sense. But for a normalization problem, the selection of reference is often arbitrary. It is more reasonable to treat the reference and the target array equally. To do this, we need to take into errors from both x and y into account. To address this issue, we designed a new method called PLTS (Perpendicular Least Trimmed Squares). The new method is based on the algorithm proposed by Li [Li04] but with some revisions. In PLTS, the vertical offset is replaced by perpendicular offset: $r(\alpha, \beta) = (y_i - \alpha - \beta x_i)/\sqrt{1+\beta^2}$. Note that the same formula is used by Total Least Squares in a error-in-variables model [vHL02]. The vertical offset measures the perpendicular distance from a data point (x_i, y_i) to the regression line, and therefore takes the errors from both x and y into account (see Figure 2.17).

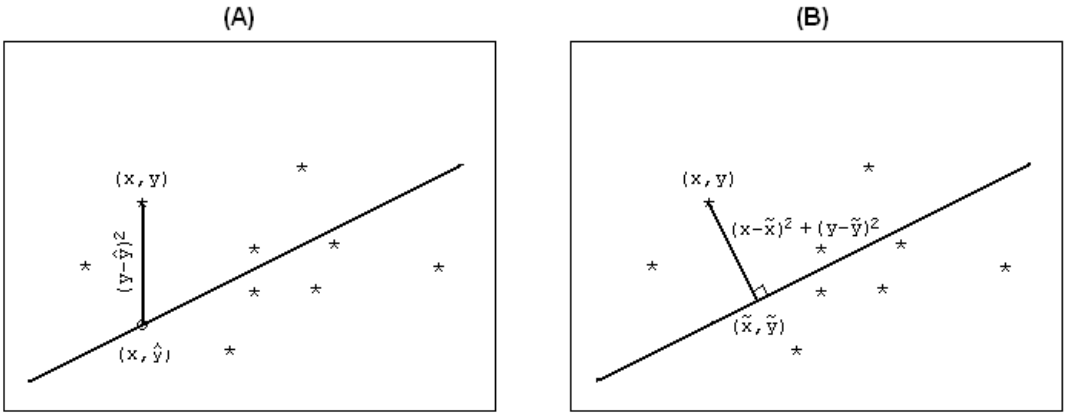

Figure 2.17: Comparison of vertical offset and perpendicular offset. (A)vertical offset used in LTS; (B)perpendicular offset used in PLTS.

2.8.2 LTS versus PLTS

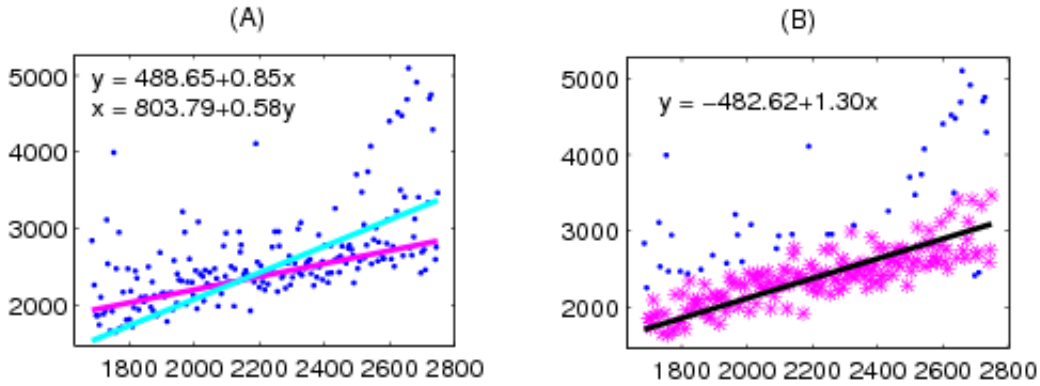

Figure 2.18: PLTS is symmetric with respect to x and y. (A)LTS; (B)PLTS. In (A), the magenta line and cyan line are the best fitted lines of regression $y \sim x$ and $x \sim y$, respectively. In (B), outliers are marked as blue points. For both LTS and PLTS, a trimming fraction of 30% is used.

We developed an algorithm to solve PLTS on the basis of Li's work [Li04]. One of the key properties of PLTS is that it is symmetric with respect to x and y. To illustrate this property, we apply both LTS and PLTS with the same trimming fraction (30%) on a data with 200 observations. As shown in Figure 2.18, for LTS, two different regression lines are obtained. One is for regression of y on x, the other is for regression of x on y. But for PLTS, the same regression line is achieved no matter x or y is used as the response variable. This property is useful when PLTS is applied to Sub-Sub normalization. By using PLTS, we achieve the same normalization no matter which array is chosen as the reference. Of more importance, PLTS takes outliers from both target and reference array into account, a more accurate estimation would be expected.

To test whether PLTS achieve a more accurate estimation of linear relations between variables than LTS, we simulate a data with of 1000 in size using the following procedure. First, we generate a vector $X = [1, 2, \cdots, 1000]$. Then we generate another

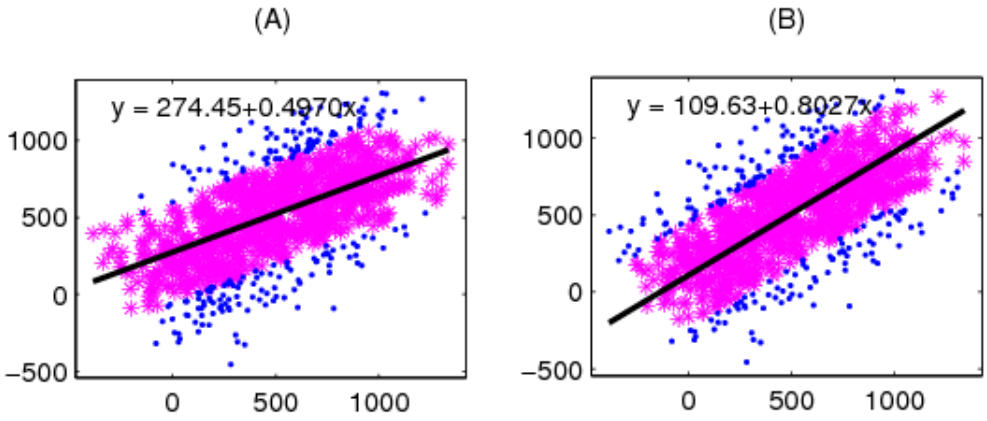

Figure 2.19: Comparison of LTS with PLTS using simulated data with errors in both x and y. (A) LTS; (B) PLTS. Magenta stars mark the data points in the subset. Blue dots indicate the identified outliers.

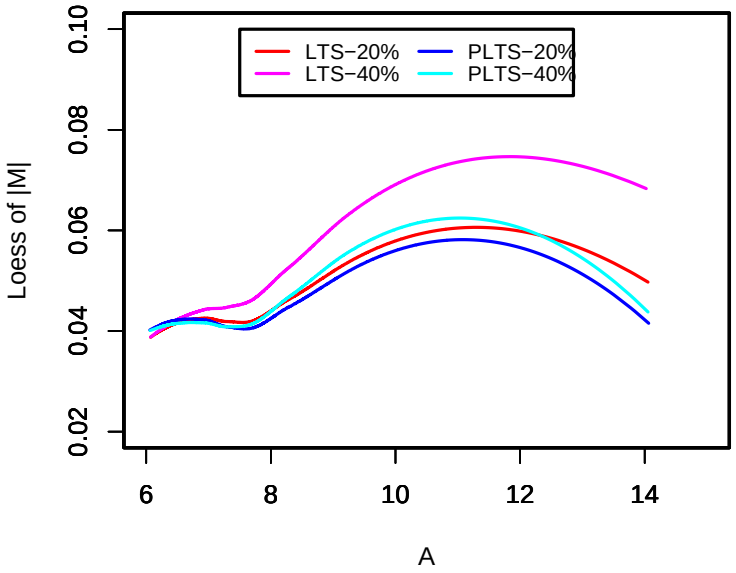

Figure 2.20: PLTS achieves more variation reduction than LTS in Sub-Sub normalization. Sub-array size and overlapping size are set to 20×20 and 10, respectively.

vector Y, where $Y_i = 100 + 0.8X_i$ for all $i = 1, 2, \cdots, 1000$. Finally, we introduce errors for both X and Y by adding a random number ε_x and ε_y to each X_i and Y_i, where $\varepsilon_x \sim N(0, 200)$ and $\varepsilon_y \sim N(0, 200)$. This procedure generates 1000 observations (x_y, y_i) with the underlying relation $Y = 100 + 0.8X$. Figure 2.19A and Figure 2.19B show the regression lines estimated by LTS and PLTS respectively using the same trimming fraction: 20%. As can be seen, the regression line estimated by LTS is $y = 279.45 + 0.4970x$. It deviate from the real relation between x and y. Those data points with larger values in vertical direction are more likely to be identified as outliers, which results in a regression line with a slope smaller than the real value. Whereas, PLTS achieves a regression line: $y = 109.63 + 0.8027x$. Both the slope and intercept of the regression line are close to the real ones.

2.8.3 Application of PLTS on Sub-Sub

If PLTS is able to capture the relations between two variables more accurately, we would expect more variation reduction after Sub-Sub normalization for the Yeast technical replicates data set. We compare the results of Sub-Sub normalization using LTS and PLTS. As expected, PLTS does reduce more variation between the replicates than LTS with different trimming fractions (see Figure 2.20). For a larger trimming fraction, the improvement in performance is more significant.

Chapter 3

Identification of perturbed genes between time courses

Microarray technologies have been applied to a wide range of biological studies including large-scale linkage analysis, association, and copy number studies et al. Among all these applications, the most frequently used application is to detect differentially expressed genes between control and treatment/disease samples. For example, to understand the mechanisms of carcinogenesis, we can compare the gene expressions in tumors with those in the normal tissues to identify the differentially expressed genes. Differential expression between treatment and control condition can be investigated from both a static and temporal viewpoint. In a static experiment design, snapshots of gene expression levels are taken without considering the temporal effect. Whereas in a temporal experiment design, which is also called a time course design, the gene expression across several time points are measured. In this chapter, we introduce a novel method, called MARD (Mean Absolute Rank Difference) analysis, which is developed to identify differentially expressed genes between treatment and control time courses.

3.1 Introduction

Microarray techniques have been widely applied to identify genes that have different expression under various biological conditions. In many cases, we regard one condition

as treatment and the other one as control, which leads to the definition of treatment-control experiment design. Differential expression between treatment and control condition can be investigated from a static or a temporal viewpoint. In a static experiment design, snapshots of gene expression levels are taken without considering the temporal effect. While in a temporal experiment design the gene expression across several time points are measured. Since the regulation of gene expression is a dynamic process, usually a temporal design provides more biological information than a static design.

It has been shown that most available approaches for the static data are not directly applicable for the time-course data [BJ04, SXL+05]. Existing data analysis methods for the time course data often focus on identifying special expression patterns across the time points [WKS04]. For example, clustering analysis is often performed on a time course data to identify gene clusters with interesting expression patterns [ESBB98, LL03]. On the other hand, several approaches have been proposed to compare different time courses and identify differentially expressed genes between them. If the sampling time points can be "aligned" between the treatment and control time courses, we can identify differentially expressed genes by direct comparison of the gene expression patterns under the two conditions. Available methods include the fold-change analysis [YSG+02], order-restricted statistics [PLL+03], the analysis of variance [PYL+03], and one-sample multivariate empirical Bayes statistic [TpS06].

However, two difficulties exist in the analysis of microarray data with temporal design. First, the sampling time points is generally different from one study to another [BJ04]. As a consequence, it is hard to integrate data from different studies. Second, a treatment may alter the "life-clock" pace of an organism. For example, it has been reported that the knockout of gene sch9 extends yeast life-span by three folds [FPP+01]. In this case it is difficult to align the time-scales of individuals under treatment and

control conditions. When the sampling time points can not be "aligned" between treatment and control, various interpolation techniques are often used as a preprocessing step before the direct comparison. For example, Bar-Joseph et al. [BJGS+03] proposed to represent gene expression patterns in treatment and control by B-spline curves and then compute a global difference measure between these two curves. Recently, Storey et al. [SXL+05] proposed to represent gene expression trajectories by a natural cubic spline and then use goodness-of-fit test for differentially expressed genes detection. However, in either situation, the time points of the two time courses must be "aligned".

In this chapter, we propose a novel method to identify differentially expressed genes between control and treatment time courses which does not require "aligned" time points in the two time courses. The method is proposed for the following considerations. (1) The relationships between genes can be estimated from microarray time course data. Functionally associated genes tend to have similar expression patterns. So we can construct a gene relationship network out of a microarray data set, where each node is a gene and each edge links two genes with similar expression patterns. Two gene relationship networks can be constructed from control and treatment time courses which may be different from each other. (2) Due to the robustness of cell system [LLYLT04, ASBL99], we may expect the gene relationship network to be also robust. Namely, we may expect that the majority of gene relationships are only marginally affected by a nonlethal treatment. Otherwise, dramatic change in the whole relationship network may cause lethal effect. (3) If a gene is substantially affected by a treatment, we would expect a dramatic change of the gene between the two relationship networks constructed from control and treatment time courses. So we estimate the effect of the treatment on a gene indirectly by investigating the gene's neighborhood change between the two relationship networks constructed from the treatment and control time courses. Namely, if the neighbors of a gene in the two networks change dramatically, we regard this gene as substantially

affected gene by the treatment. Based on these three considerations, we design a statistic called MARD (Mean Absolute Rank Difference) to measure the effect of a treatment on a gene. Since we compare the two constructed relationship networks instead of directly comparing the expression patterns, the problem of sampling scheme differences between treatment and control is not an issue for our approach.

3.2 Available approaches

3.2.1 Static analysis based methods

Many approaches have been introduced to identify genes that are differentially expressed between two experiments with a static expression design. However they can not be applied directly to compare time courses , due to differences in sampling rate and variation in the timing of biological processes [BJGS+03]. Previously proposed approaches for identifying differentially expressed genes between time courses essentially applied static analysis methods. These ad hoc methods are not generally applicable, or only applicable for a specific data set. These methods include cluster analysis [ZSV+00], generalized singular value decomposition [ABB03], point-wise comparison [NRS+02, HLM+01], and customer-tailored models [XOZ02]. Although these approaches have achieved some success, they suffer from many problems. Cluster analysis identifies gene clusters in which a large portion of genes change in expression. But it fails to detect differentially expressed genes that belong to clusters for which most genes do not change [ZSV+00]. Generalized singular value decomposition can be used to detect difference between various sets of gene sets but it is not applicable to comparing individual genes. Moreover, this method requires that the two time courses being compared have the same number of time points , which is not the case in general [ABB03]. Direct point-wise comparison between samples in two time courses does

not take the dynamic nature of the time course experiments into account. Further more, it can not distinguish the real gene expression changes from the random noises. In addition, for two time courses that have different timing scales, direct comparison is impossible. Custom-tailored models require significant assumptions about he shape of the expression profiles being compared, e.g. following linear or quadratic models, and therefore does not provide a general solution to time course comparisons [XOZ02]. In most cases, it is hard to justify using a highly specific model. Even if some genes are known to change in a certain way over time, e.g. a sinusoidal model foe the cell cycle time courses, using such a specific model for the shape of expression files may result in failure of detecting changes in many genes that are differentially expressed but not behave in the way that is assumed by the model [XOZ02].

3.2.2 ANOVA method

Park et al. provided a detail description about application of statistical tests for identifying differentially expressed genes in time course microarray experiments [PYL+03]. Two-way ANOVA model is applied to detect differentially expressed genes between two time courses with aligned time points, where each time course is obtained from a group of samples (treated or control).

Let y_{ikln} represent the expression level of gene n in replication l from group i at time k. The following models, M1 and M2, are considered for data set with and without replications, respectively.

$M_1 : y_{ikln} = \mu_n + \alpha_{in} + \beta_{kn} + (\alpha\beta)_{ikn} + \varepsilon_{ikln}$,

$M_2 : y_{ikln} = \mu_n + \alpha_{in} + \beta_{kn} + \varepsilon_{ikln}$,

where $i = 1, 2; k = 1, \cdots, K; l = 1, \cdots, L$; and $n = 1, \cdots, N$. The gene effects μ_n capture the overall mean expression value for gene n across the arrays. The α_{in} terms account for gene specific group effect representing overall differences between

the treated and control group. The β_{ikn} account for time effects that capture differences in the overall expression in the samples from different time points. The $(\alpha\beta)_{ikn}$ terms account for the interaction effect between group and time. Note that the interaction term can not be estimated if there is no replication in a experiment. To identify genes that are differentaily expressed between treatment and control groups, we are interested in genes that show significant interaction effect $(\alpha\beta)_{ikn}$ in model M_1 or significant group effect α_{in} in model M_2. Testing significance of these effects can be achieved by the calculation of F-statistics for each gene.

3.2.3 Continuous representation based method

Another method proposed by Bar et al in 2003 is based on the continuous representation of time courses [BJGS$^+$03]. Gene expression profiles in both the treatment and control time courses are represented as continuous curves using B-splines. To address the time shift and time scale problems between the two time courses, liner warping function is used to obtain an optimal alignment by adjusting shifting and stretching parameters to minimize a global error function [BJGG$^+$03]. For a given gene, its expression profiles in the control and treatment time course are denoted as C_1 and C_2, respectively. Then the problem of detecting expression difference of the gene in the two time courses has been converted into the following hypothesis testing problem:

H_0: C_2 is a noisy realization of C_1,

H_1: C_1 and C_2 are independent.

The test is performed using the log likelihood ratio statistics written as

$$2 \log \frac{p(C_2|C_1, H_1)}{p(C_2|C_1, H_0)}.$$

The log likelihood ratio statistics measures the ability of the hypothesis to explain the difference between the two curves.

To compute $p(C_2|C_1, H_0)$, the noise in the individual measurements is assumed to be normally distributed with mean 0 and variance σ^2. Denote the actual expression values measured in the control and treatment experiment as Y_1 and Y_2, respectively. Due to the difference in sampling rate and temporal expression variations, Y_1 and Y_2 can not be compared directly. Therefore, the spline curve is sampled at the time points in the control experiment to obtain a set of expression values, denoted as $\mathrm{Y'}_2$. Now $\mathrm{Y'}_2$ is comparable to Y_1 because they have the same sampling rate. Thus, we can set

$$p(C_2|C_1, H_1) = p(Y_2'|\sigma^2, H_1) = \frac{1}{(2\pi\sigma^2)^{m/2}}.$$

To compute $p(C_2|C_1, H_1)$, the definition of global difference between two expression curves C_1 and C_2 is introduced as

$$e^2 = D(C_1, C_2) = \frac{\int_{v_s}^{v_e} [C_2(t) - C_1(t)]^2 dt}{v_e - v_s},$$

where v_s and v_e are the start and end of the interval in which the two curves can be compared. Then we can set $p(C_2|C_1, H_1) = p(e^2|Y_1, \sigma^2, H_0)$. To calculate $p(e^2|Y_1, \sigma^2, H_0)$, one replaces it with the maximum-likelihood assignment of $p(e^2|Y_1, \sigma^2, H_0)$, which can be computed by finding a curve C with a global distance of (e^2) from C_1 that maximizes the probability of C being a noisy realization of C_1. That is, only a global error value e^2 that can not be adequately explained by the best (maximum-likelihood) curve C will be considered significant.

3.2.4 EDGE method

Storey et al. introduced the method called EDGE (extraction of differential gene expression) in 2005 [SXL$^+$05]. Under the null hypothesis, the method assumed that there is no differentially expressed genes between the treatment and control time courses. That is, the treated and control groups have the same average expression profile of all the genes. Therefore, we can use a single cubic curve to fit the combined group. Under alternative hypothesis, the method fits a cubic curves in each group seperately. Fitted values under the null and alternative hypothesis are calculated for each observed value. The residuals of the fitting are then obtained by subtracting the fitted values from the observed values. Denoting the sum of squares of the residuals obtained from the null hypothesis and alternative hypothesis as SS_i^0 and SS_i^1, respectively, a statistic for gene i is constructed as

$$F_i = \frac{SS_i^0 - SS_i^1}{SS_i^1}.$$

This statistic compares the goodness of fit of the model under the null hypothesis with that under the alternative hypothesis. It is a quantification of evidence for differential expression between the treatment and control time courses. The larger it is the more differentially expressed the gene appears to be.

3.3 Description of MARD analysis

In this thesis we mainly focus on two-channel cDNA arrays, but the main idea can be extended to other types of arrays.

Given a data set from treatment-control time course design, suppose that it measures the expression levels of n genes at K_1 time points/samples under control condition and K_2 time points/samples under treatment condition. Let's denote the gene

expression levels under the control and treatment condition as $Y^{(1)} = \left(y_{gk}^{(1)}\right)_{n\times K_1}$ and $Y^{(2)} = \left(y_{gk}^{(2)}\right)_{n\times K_2}$ correspondingly. In both matrices, each row is the expression levels of a gene across different time points while each column stands for all the n genes' expression levels at one specific time point.

First, under each condition (treatment or control) and for each gene i, we can define the relationships between gene i and all the other genes by calculating the distance $d(i,j)$ between their expression patterns. Several metrics can be used to describe this distance including the Euclidean distance, Pearson correlation coefficient and etc.. Then for gene i, we obtain two distance vectors $d^{(1)} = (d_1^{(1)}, \cdots, d_{i-1}^{(1)}, d_{i+1}^{(1)}, \cdots, d_n^{(1)})$ and $d^{(2)} = (d_1^{(2)}, \cdots, d_{i-1}^{(2)}, d_{i+1}^{(2)}, \cdots, d_n^{(2)})$, where $d_j^{(1)}$ and $d_j^{(2)}$ are the distances between the expression patterns of gene i and gene j under the control condition and the treatment condition respectively.

Second, for each distance vector under each condition, the ranks of all the genes $j \neq i$ are calculated and denoted by $r^{(1)} = (r_1^{(1)}, \cdots, r_{i-1}^{(1)}, r_{i+1}^{(1)}, \cdots, r_n^{(1)})$ and $r^{(2)} = (r_1^{(2)}, \cdots, r_{i-1}^{(2)}, r_{i+1}^{(2)}, \cdots,$
$r_n^{(2)})$, where r_j^1 and r_j^2 are the rank of $d_j^{(1)}$ in $d^{(1)}$ and $d_j^{(2)}$ in $d^{(2)}$ respectively. Then the change of the relationships between gene i and gene j under the two conditions can be described as

$$\triangle r_j = |r_j^{(1)} - r_j^{(2)}|$$

where $j = 1, 2, \cdots, n$ and $j \neq i$.

Thirdly, we define a "neighborhood" for gene i because the change of gene i under the two conditions should not be described by the change in the relationships between it and all the other genes. Two types of genes are included in the "neighborhood". The first type includes those genes that have very similar expression profiles with gene i because these genes tend to be functionally associated with gene i. However, if we only

consider this type of genes, when all the neighbors of gene i are perturbed by the treatment to the same level, we would not see significant change in the "neighborhood" of gene i although gene i does change under the two conditions. To make up this problem, we include the second type of genes into the "neighborhood" of gene i which have very large distance with gene i under either condition. These distant genes usually consists of genes from various function categories and may have no biological association with gene i. When all the neighbor genes are perturbed at the same level, these distant genes will have large change in their relationships with gene i because most of them may not be perturbed together with gene i or may be perturbed in very different way with gene i. With all these considerations, we have the following three definitions of "neighborhood":

1. q-proximal neighborhood: $G_1^{(k)}(q) = \{j : d_j^{(k)} \leq d^{(k)}(q)\}$,
2. q-distal neighborhood: $G_2^{(k)}(q) = \{j : d_j^{(k)} \geq d^{(k)}(1-q)\}$,
3. q-two-end neighborhood: $G_3^{(k)}(q) = G_1^{(k)}(q) \bigcup G_2^{(k)}(q)$,

where $d^{(k)}(q)$ is the q-th lower quantile of the distance vector $d^{(k)}$ and $k = 1, 2$ for the two conditions. So q described how many genes are included in the "neighborhood" of one gene. Details about how to determine q can be found in "Discussion".

Finally, given the value of q and following one definition of "neighborhood", the Mean of Absolute Rank Difference(MARD) for gene i is defined as

$$M_i(q) = \frac{\sum_{j \in G(q)} \triangle r_j}{\#G(q)}$$

where $G(q) = G_l^{(1)}(q) \bigcup G_l^{(2)}(q)$ is the union of the two sets of neighborhood genes of gene i under control and treatment condition, $l = 1, 2, 3$ corresponding to the three definitions of "neighborhood" and $\#G(q)$ stands for the total number of genes inside

$G(q)$. So for any given gene i and value of q, we can calculate a MARD for each of the three definitions of "neighborhood".

Having the MARD values of all the genes, we can rank the genes in descending order of their MARD values. The larger the MARD value of a gene is, the larger change the gene has under the treatment and control conditions.

In the next two sections, we will test our approach using two treatment-control time-course microarray data sets. In the first data set, time courses of gene expression in response to Ca^{2+} were measured with and without the FK506 treatment in budding yeast [YSG$^+$02]. In both time courses, gene expression levels were measured at four well matched time points after Ca^{2+} addition: 15, 30, 45 and 60 min. Therefore we refer to this data set as the aligned time course data set. The other data set provides the gene expression profiles across the cell cycle of wild-type budding yeast [SSZ$^+$98] and the $\triangle fkh1 \triangle fkh2$ double mutant [ZSV$^+$00]. The two time courses were measured independently by two research groups and different sampling schemes were used. Therefore it's difficult to directly compare the two time courses. We refer to this data set as the unaligned time-course.

3.4 Evaluation of MARD on aligned time course data

3.4.1 Ca^{2+} Response w/o FK506 Inhibition Data

Calcineurin is a Ca^{2+}/ calmodulin-dependent protein phosphatase. It is activated by specific environmental conditions, including exposure to Ca^{2+} or Na^+, and then induces gene expression by regulating the activity of the transcription factor Crz1p/Tcn1p. The effects of Ca^{2+} and Na^+ can be counteracted by FK506, which is an inhibitor of the calcineurin protein, thereby shutting down the entire signaling pathway (see Figure 3.1). To screen for calcineurin-dependent genes regulated by Ca^{2+}, Yoshimoto et al. [YSG$^+$02]

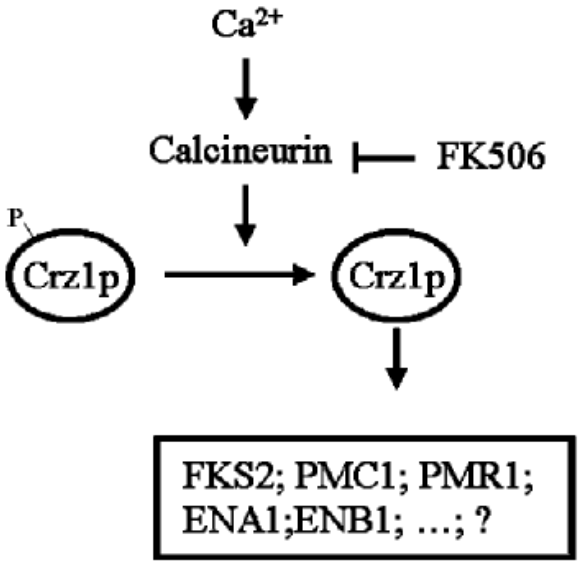

Figure 3.1: Calcineurin/Crz1p signaling pathway in S. cerevisiae.

performed four groups of cDNA microarray: (1) Ca^{2+} time course, (2) Ca^{2+} + FK506 time course, (3) Ca^{2+} +FK506/ Ca^{2+}, and (4) $\triangle crz1$/CRZ1: Ca^{2+}. In experiment (1) and (2), yeast samples were collected at t =15, 30, 45 and 60 min after being exposed to Ca^{2+} and Ca^{2+}+FK506 separately, and were compared with sample collected at t =0. In experiment (3), direct comparison was made between samples collected from the FK506-treated and control samples at 15 and 30 min after Ca^{2+} addition. In experiment (4), direct comparison was made between samples collected from wild type and $\triangle crz1$ strain at 15 and 30 min after Ca^{2+} addition. The authors identified 153 calcineurin-dependent genes activated by Ca^{2+} based on microarray data from all the four experiments.

Our aim is to identify the genes significantly perturbed by the inhibition of calcineurin with FK506. Since FK506 blocks the calcineurin/Crz1p signaling pathway, we would expect that the genes directly related to this pathway are more severely perturbed than other genes. According to our approach, the degree of perturbation of a gene is

measured by its neighborhood-change between the treatment and control time-course experiments. We only use the data from experiment (1) and experiment (2) and regard Ca^{2+}+FK506 time course as treatment and Ca^{2+} time course as control. Totally there are about 6,000 genes whose expression levels were measured in the two time courses. We filter out genes with missing values in either time course after which 5052 genes left. Since genes with constant expression levels across the time points in both time courses are of no interest, we remove 20% constantly expressed genes with the smallest variation across all the time points in the two time courses. For the remaining 4042 genes, we apply the MARD analysis (two-end neighborhood with an informative fraction q=1% in this paper, see "Discussion" for determination of q). Note here only genes activated by Ca^{2+} are of interest, we use ratios rather than log transformed ratios as the expression measurements to lower the MARD value repressed genes. Detail explanation will be given in "Discussion".

3.4.2 Identification of Perturbed Genes

We calculated the MARD values for all the 4042 genes and the distribution of them is shown in Figure 3.2. Biologically speaking, after the inhibition of calcineurin by FK506, we would expect dramatic neighborhood changes for genes that are directly related to Calcineurin/Crz1p signaling pathway. Therefore these genes are expected to have high MARD values. On the other hand, house-keeping genes, which are essential for cell survival, tend to be less severely affected by any perturbation, since significant change in the activities of these genes may be lethal to Yeast. Consequently, these house-keeping genes should have lower MARD values. As shown in Fig. **??**A, the histogram of MARD shows a notable heavy tail on the right-hand side and a small peak on the left-hand side, which seem to be the calcineurin/ Crz1p pathway related genes and house-keeping genes, respectively. We investigate those genes with small MARD

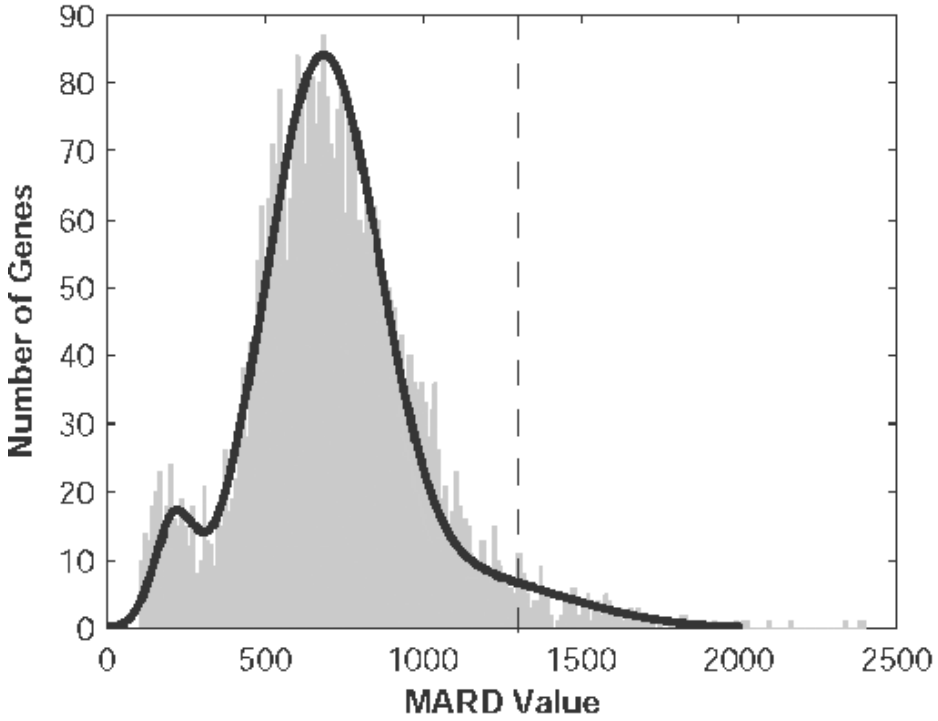

Figure 3.2: Distribution of the MARD values (informative fraction q=1%) of the 4042 genes in the Ca^{2+} Response w/o FK506 Inhibition Data. Threshold at the vertical dash line result in 142 genes.

values (on the left hand), it turns out that most of them are housekeeping genes, such as ribosomal protein genes. Certainly, those genes with large MARD values (on the right side) are more of interest. We validate their association with calcineurin/ Crz1p pathway through comparing with previous studies [YSG$^+$02, Cye03].

Table 3.1 lists the top 40 substantially perturbed genes in the Ca^{2+} response w/o FK506 inhibition data. As shown, most of these genes identified by MARD analysis have also been reported as calcineurin dependent genes by Yoshimoto et al. [YSG$^+$02]. For example, PMC1, the vacuolar Ca^{2+} ATPase involved in depleting cytosol of Ca2+ ions, identified by both MARD analysis and Yoshimoto et al. Previous study shows that PMC1 prevents growth inhibition by activation of calcineurin in the presence of elevated concentrations of Ca^{2+} [CF94]. Without FK506 treatment, expression of Pmc1 gene

Table 3.1: The top 40 substantially perturbed genes in response to FK506 treatment. The JBC column indicates whether a gene was also reported by Yoshimoto et al.

Rank	Gene	Name	Function	JBC
1	YMR316C-A		unknown	Yes
2	YLR414C		unknown	Yes
3	YLR194C		unknown	Yes
4	YER184C		unknown	Yes
5	YPL149W	ATG5	autophagy-related protein	Yes
6	YOR385W		unknown	Yes
7	YBR005W	RCR1	involved in chitin deposition in the cell wall	Yes
8	YNL192W	CHS1	chitin synthase activity	Yes
9	YOL014W		unknown	Yes
10	YGR268C	HUA1	actin patch assembly	Yes
11	YGR144W	THI4	thiamine biosynthesis	Yes
12	YMR316W	DIA1	unknown	Yes
13	YOR209C	NPT1	nicotinate phosphoribosyltransferase activity	Yes
14	YNR059W	MNT4	alpha-1,3-mannosyltransferase activity	Yes
15	YNR010W	CSE2	RNA polymerase II transcription mediator activity	Yes
16	YOL158C	ENB1	ferric-enterobactin transporter activity	Yes
17	YDL234C	GYP7	Rab GTPase activator activity	Yes
18	YGL006W	PMC1	calcium-transporting ATPase activity	Yes
19	YMR096W	SNZ1	protein binding	Yes
20	YJL171C		unknown	Yes
21	YAR027W	UIP3	unknown	Yes
22	YNL044W	YIP3	unknown	Yes
23	YDL009C		unknown	Yes
24	YNL020C	ARK1	protein serine/threonine kinase activity	Yes
25	YDL172C		unknown	Yes
26	YDR482C	CWC21	unknown	Yes
27	YDL012C		unknown	Yes
28	YBR054W	YRO2	unknown	No
29	YPL067C		unknown	No
30	YDL241W		unknown	No
31	YCR011C	ADP1	ATPase activity, coupled to transmembrane movement of substances	No
32	YKL001C	MET14	adenylylsulfate kiunknownse activity	Yes
33	YBR016W		unknown	No
34	YOL016C	CMK2	calcium- and calmodulin-dependent protein kinase activity	Yes
35	YML125C		unknown	Yes
36	YPR170C		unknown	Yes
37	YBR162W-A	YSY6	unknown	No
38	YLR120C	YPS1	aspartic-type endopeptidase activity	Yes
39	YGL165C		unknown	Yes
40	YMR018W		unknown	No

are up-regulated by at most 7-fold in response to Ca^{2+}. However, when the activity of calcineurin is inhibited by FK506, the gene expression of Pmc1 becomes insensitive

to high concentration of Ca^{2+}. Therefore, our result implies that the transcriptional regulation of Pmc1 is dependent on the activity of the Ca^{2+}/Calcineurin pathway, which suggests a positive feedback in this pathway.

3.4.3 Consistency with Previous Study

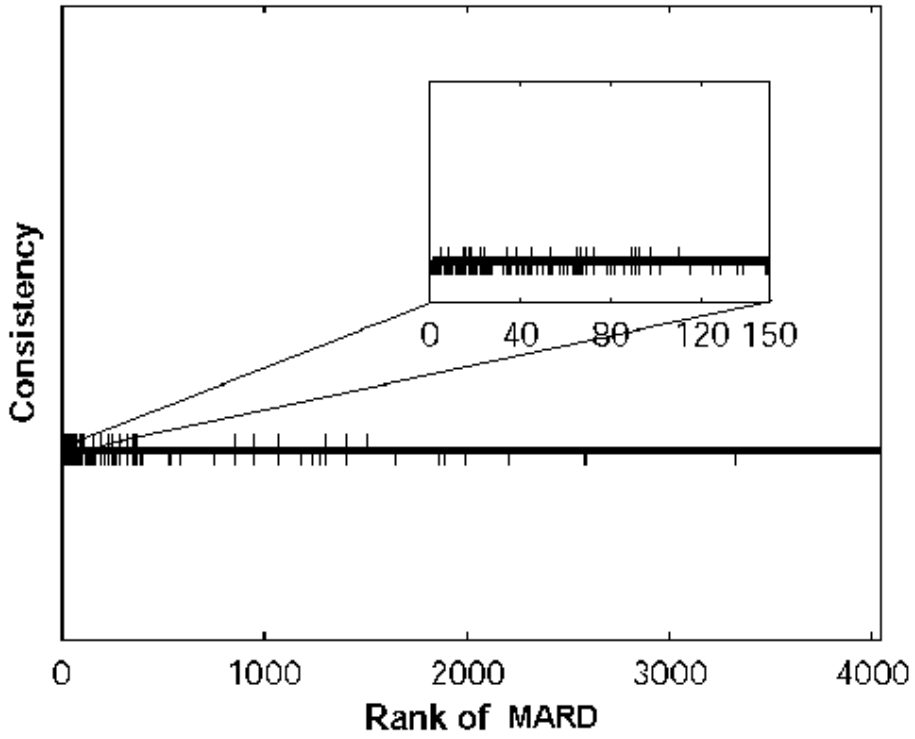

Figure 3.3: The ranks of MARD values for genes identified by previous studies in aligned data. Bars below the thick line are genes identified by Yoshimoto et al. [YSG+02]; Bars above the line are genes with known functions [Cye03].

We checked the consistency of our identified genes with those identified as differentially expressed in Yoshimoto et al [YSG+02]. Yoshimoto et al applied a two-step analysis to identify calcineurin dependent genes activated by Ca^{2+}. First, they selected 934 Ca^{2+}-activated genes that were induced more than 2-fold at either 15 or 30 min after

Ca^{2+} addition in experiment(1). Second, they assessed the extent to which the expression of each of these genes was reduced by calcineurin inhibition by direct comparison of FK506-treated and non FK506-treated cells exposed to Ca^{2+} in experiment(3). Genes identified by this analysis are based on direct ratio measurements("Ca^{2+} addition 15 or 30 min" versus "Ca^{2+} addition 0 min" and "FK506-treated" versus "non FK506-treated") in both steps and thereby are of high confidence. However, some calcineurin dependent Ca^{2+} activated genes may be missing, because: (1) They didn't take into account the genes that activated by Ca^{2+} only at 45 or 60 min (2) The arbitrarily determined 2-fold threshold may filter out some interested genes. Our MARD analysis aims to find the genes that were significantly perturbed in terms of neighborhood by FK506 treatment. We only consider the two time courses in experiments (1) and (2).

Despite the differences between our method and the approach in Yoshimoto et al, the two sets of identified genes are highly consistent with each other. Yoshimoto et al identified 153 calcineurin dependent Ca^{2+} activated genes, among which 111 are present in our data set (4042 genes included in total). To make a fair comparison, we select the top 111 genes with the highest MARD values as listed in supplementary Table. 1. Among these 111 genes, 63 genes are also identified by Yoshimoto et al with a p-value of 5.7×10^{-77}. The consistency of our result with that of Yoshimoto et al is better illustrated in Figure 3.3. Most of the genes contain the Crz1p binding motif in their promoter regions, suggesting that they were directly regulated by Crz1p. As can be seen from Figure 3.3, genes with higher MARD values are more likely to be reported as calcineurin dependent Ca^{2+} activated genes in [YSG$^+$02]. Specifically, all the top 27 genes with highest MARD values are among the 153 genes identified by Yoshimoto et al. More interestingly, we found that crz1 itself is significantly perturbed by FK506 according to our result (with rank of 95) while it is not identified as calcineurin-dependent gene by Yoshimoto et al. It turns out that crz1 gene encodes an auto-regulated transcription

factor, i.e., it regulates the transcription of itself (personal communication with Martha Cyert, Stanford University).

We also apply two-way ANOVA and EDGE analysis for the data [PYL+03, SXL+05]. The two-way ANOVA analysis results in 167 genes whose expressions are significantly different between the FK506 treated and non-FK506 treated time courses with a significance level $\alpha = 0.01$. Among these genes only 6 fall into the 153 genes identified by Yoshimoto et al. If we reduce the significance level to 0.05, 783 genes are identified, among which 54 are also within the 153 genes. However, the EDGE program results in no differentially expressed genes between the two time courses with a false discovery rate less than 10%. This may be caused by the lack of replicates or the small number of time points in the experiment.

3.4.4 Consistency with Direct Comparison

Because the sampling time points are well matched between the treatment and control in this data set, it is possible to directly calculate the gene expression profile changes between treatment and control. Here, we would expect the neighborhood change of a gene to be consistent with its expression pattern change for the following reasons: (1) the biology system is robust [LLYLT04, ASBL99], only a small fraction of genes have significant expression changes in response to a nonlethal perturbation; (2) we use Euclidean distance to measure the neighborhood of genes. On the other hand, since our approach explicitly uses more information about the gene-gene relationship than direct comparison of gene expression patterns, some differences are also expected. The change of gene expression patterns in treatment and control is defined as the normalized Euclidean distance:

$$L(Y_g^{(1)}, Y_g^{(2)}) = \frac{\| Y_g^{(1)} - Y_g^{(2)} \|}{\| Y_g^{(1)} \| + \| Y_g^{(2)} \|} \tag{3.1}$$

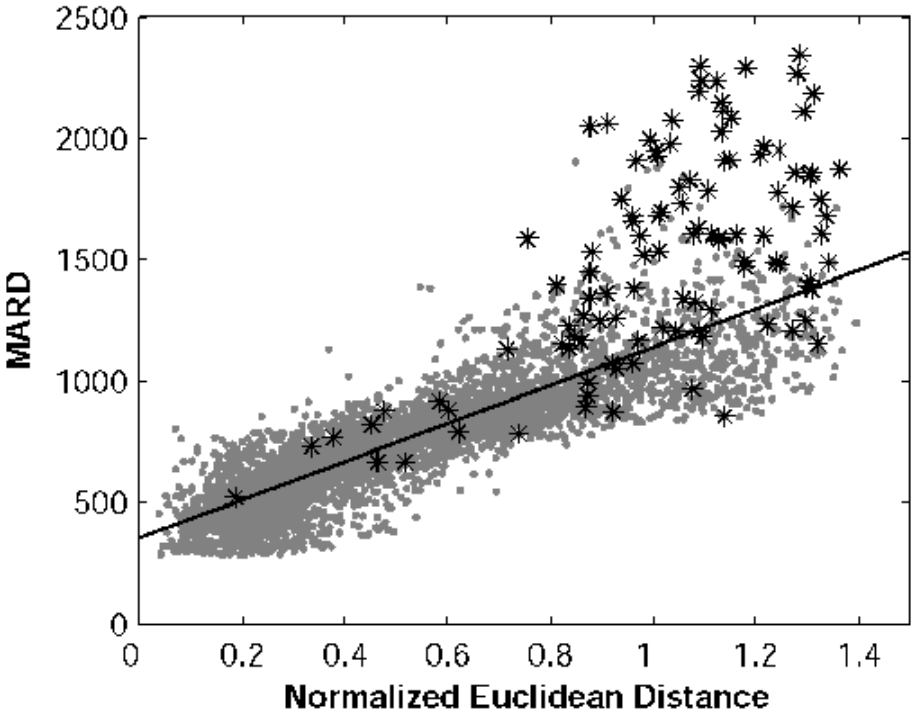

Figure 3.4: Consistency of MARD value with normalized Euclidean distance. The identified genes in [YSG+02] are marked as black stars.

where $Y_g^{(1)}$, $Y_g^{(2)}$ are the two time courses of gene g under control and treatment conditions respectively and $\|\cdot\|$ is the L_2 norm in this study.

We plot the MARD value of each gene versus the expression pattern change for each gene in Figure 3.4. As we can see from the plot, genes with higher MARD values tend to have larger expression pattern change in treatment versus control time course. The correlation coefficient between the MARD values and the normalized Euclidean distances of all the genes is 0.844. Furthermore, most of the genes identified by Yoshimoto et al. have higher MARD values than the normalized Euclidean distances. This indicates that the MARD-score based analysis has a higher discriminant power.

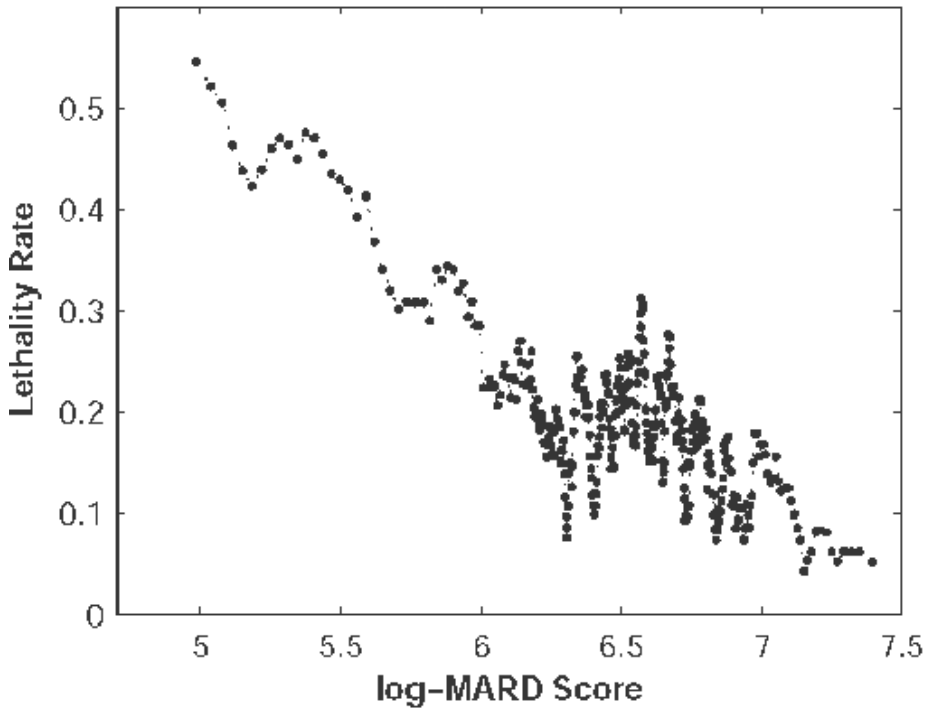

Figure 3.5: Relationship between MARD values (q=1%) and lethality in aligned data.

3.4.5 Essentiality and MARD

Due to the robustness of a biological system, we would expect small neighborhood changes in response to a perturbation for genes that are essential for cell survival. Therefore, we studied the relationship between the MARD value and essentiality of genes. Systematic gene deletion experiments have been performed in yeast [WSA+99]. In total, 5860 yeast genes are deleted and 1117 (19%) of them are identified as essential genes which means that single deletion of these 1117 genes is lethal for cells grow in YPD medium.

We rank the MARD values for all the 4042 genes and calculate the lethality rate using genes ranked from i to $i + 100$ for different $i = 1, 10, 20, \cdots$. The lethality rate is defined as the fraction of essential genes in the gene set. We plot the MARD values against the resulting lethality rate for each gene set in Figure 3.5. As shown in the figure,

the lethality rate decreases from 56% to 5% with the increase of MARD values. This is reasonable because the lethality rate describes how essential the genes in the gene set are to the organism. If most of the genes inside a gene set are essential, the perturbation by the treatment on them should be relatively small because significant perturbation on them may be lethal to the organism. Now we have smaller MARD values for more essential gene set. This means that our MARD statistics is a good measure of the effect of the treatment on each gene. Since our method actually measures the change in the neighborhood of each gene, this also support our rationale that genes, which are more severely affected by a treatment, tend to have larger neighborhood changes and thereby a higher MARD values. In addition, these results also show that gene relationship network is robust, because essential genes that play important roles tend to be less affected by a treatment.

3.5 Evaluation of MARD on un-aligned time course data

3.5.1 The wt/$\triangle fkh1 \triangle fkh2$ cell cycle data

Fkh1 and Fkh2 are two yeast transcription factors involved in cell cycle regulation. Deletion of each of them may cause mis-regulation of some genes, especially cell-cycle related genes. Spellman et al. performed a time-course experiment to identify cell-cycle regulated genes in wild type yeast [SSZ$^+$98]. Zhu et al. performed another time-course experiment in which fkh1 and fkh2 were knocked-out [ZSV$^+$00]. Two clusters of genes (CLB2 and SIC1) that show different expression patterns in the $\triangle fkh1 \triangle fkh2$ mutant were identified as Fkh1 or Fkh2 dependent genes by Zhu et al. Since the two time course data sets have different sampling schemes, the expression patterns of genes

in them can not be directly compared. Bar-Joseph et al. identified 30 cell-cycle genes and 22 non-cycling genes as differentially expressed by representing expression patterns of genes by function curves and comparing directly the function curves [BJGS+03].

3.5.2 Identification of perturbed genes

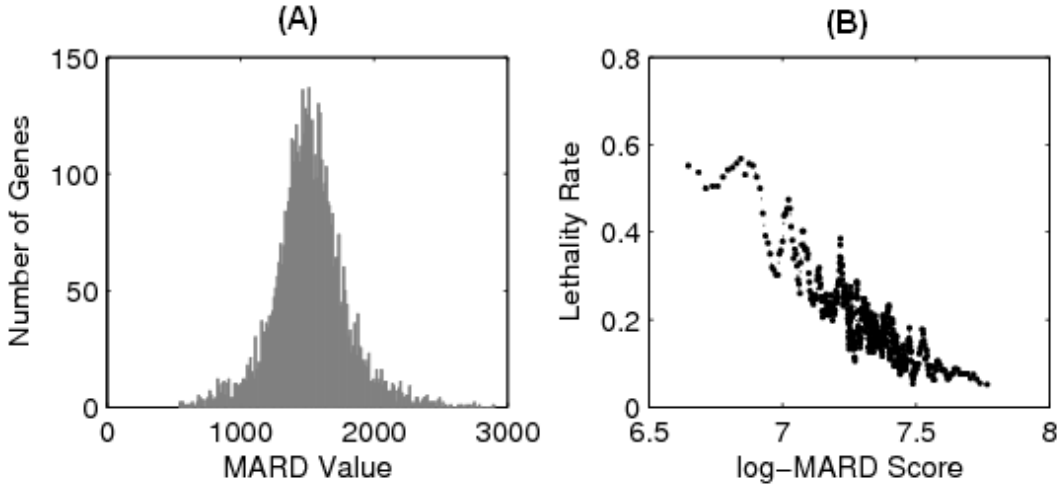

Figure 3.6: MARD analysis of the un-aligned data (the wt/$\triangle fkh1 \triangle fkh2$ cell cycle data). (A) distribution of the MARD value (informative fraction q=1%) of the 5525 genes. (B) Relationship between MARD value and lethality.

After filtering out the genes with more than one missing values, we calculate MARD values (two-end neighborhood with an informative fraction q=1%) for the remaining 5525 genes to identify genes significantly perturbed by $\triangle fkh1 \triangle fkh2$ knockout. In this data set we use log transformed ratios as the expression measurements instead of using the ratios directly. The reason for doing this can be found in Discussion. The distribution of MARD values of all the 5525 genes is shown in Figure 3.6A. We selected the top 100 genes with the highest MARD values which are listed in supplementary Table. 2. Among these 100 genes, 41 genes are cell-cycle related genes according to the result from Spellman et al. ($p - Value = 4.9 \times 10^{-14}$). While comparing these 100 genes

with the results by Bar-Joseph et al. [BJGS+03], 13 genes show up in their top 30 cycling genes and 9 genes show up in their top 22 non-cycling genes ($p-Value = 5.7 \times 10^{-11}$). Finally when comparing our results with that by Zhu et al. [ZSV+00], we find that none of the up-regulated genes and 7 ($p - Value = 7.2 \times 10^{-7}$) of the down-regulated genes identified by them are in our top 100 genes. Again a negative correlation between MARD value and essentiality is observed which is shown in Figure 3.6B.

$\triangle fkh1 \triangle fkh2$ double mutation have global effects on cell growth. With this double mutation, the cells show pseudohyphal and invasive growth, unusual cell morphology, and slow growth rates [ZSV+00]. Consistent with these phenotypes, many of the top 100 genes identified by our approach are involved in cell cycle, cell wall organization, amino acid synthesis or pseudohyphal growth. For example, MEP1 (with rank of 76) is an ammonium permease that regulates pseudohyphal differentiation in response to ammonium limitation [LH98]. TEC1 (with rank of 43) is a transcription factor which is involved in pseudohyphal growth [KWT+02]. This gene is also identified by Zhu et al. but not by Bar et al. [ZSV+00]. PCL2 (CLN4, rank 24) is a G1 cyclin which associates with Pho85p cyclin-dependent kinase (CDK) to contribute entry into the mitotic cell cycle and is essential for cell morphogenesis [MMO+94, MA04]. We also checked the genome-wide binding data [LRR+02] that described the association of Fkh1p and Fkh2p with genes expressed in G1 and S phases, and found 7 genes bound by Fkh1p ($p-Value = 0.014$) and 15 genes bound by Fkh2p ($p-Value = 4.3 \times 10^{-8}$) in the top 100 perturbed genes. Table 3.5.2 lists the top 60 genes that appear to be substantially perturbed in the $fkh1 \triangle fkh2 \triangle$ double mutant. The Zhu, Bar column indicate that whether a gene is identified by Zhu et al and Bar et al, respectively. FKH1 and FKh2 column indicate whether a gene is bound by the two transcription factors. As shown, most of them are cell cycle related genes.

Table 3.2: The top 60 genes that are substantially perturbed in the $fkh1\triangle fkh2\triangle$ double mutant.

Rank	Gene	Name	Zhu	Bar	FKH1	FKH2	Phase	Function
1	YGR108W	CLB1	yes	yes			G2/M	B-type cyclin involved in cell cycle progression
2	YNL058C			yes			G2/M	Potential Cdc28p substrate
3	YLL023C							Unknown
4	YEL040W	UTR2		yes			M/G1	Function in cell wall maintenance
5	YPL221W						G1	Unknown
6	YDL215C	GDH2	yes	yes		yes		NAD(+)-dependent glutamate dehydrogenase
7	YDR098C	GRX3		yes				Glutathione-dependent oxidoreductase
8	YJR148W	BAT2					G1	Cytosolic branched-chain amino acid aminotransferase
9	YLL028W	TPO1					G2/M	Polyamine transporter
10	YDL038C	PRM7						Unknown
11	YDR380W	ARO10					G2/M	Phenylpyruvate decarboxylase
12	YER037W	PHM8		yes				Unknown
13	YJL088W	ARG3						Ornithine carbamoyltransferase
14	YOL152W	FRE7		yes				Unknown
15	YLL061W	MMP1					S	S-methylmethionine permease
16	YIL169C							Unknown
17	YKL109W	HAP4						Subunit of Hap2p3p4p5p CCAAT-binding complex
18	YLR190W	MMR1					G2/M	Phosphorylated protein of the mitochondrial outer membrane
19	YPL057C	SUR1					G1	Probable catalytic subunit of a mannosylinositol phosphorylceramide synthase
20	YIL160C	POT1		yes				3-ketoacyl-CoA thiolase
21	YOL155C							Unknown
22	YHR152W	SPO12					G2/M	Unknown
23	YKL161C							Mpk1-like protein kinase
24	YGL184C	STR3		yes			S	Cystathionine beta-lyase
25	YOL058W	ARG1					S/G2	Arginosuccinate synthetase
26	YNL057W						G2/M	Unknown
27	YMR032W	HOF1					G2/M	Bud neck-localized, SH3 domain-containing protein required for cytokinesis
28	YPL265W	DIP5		yes			G2/M	Dicarboxylic amino acid permease
29	YDL039C	PRM7					G2/M	Hypothetical Pheromone-regulated protein
30	YOR202W	HIS3						Imidazoleglycerol-phosphate dehydratase

Table 3.2: Continued

Rank	Gene	Name	Zhu	Bar	FKH1	FKH2	Phase	Function
31	YLR194C			yes			M/G1	Unknown
32	YGR044C	RME1					G1	Zinc finger protein involved in control of meiosis
33	YEL024W	RIP1						Ubiquinol-cytochrome-c reductase
34	YDL127W	PCL2					G1	G1 cyclin, associates with Pho85p cyclin-dependent kinase
35	YNL009W	IDP3						Peroxisomal NADP-dependent isocitrate dehydrogenase
36	YMR246W	FAA4					M/G1	Long chain fatty acyl-CoA synthetase
37	YGL055W	OLE1		yes			M/G1	Fatty acid desaturase
38	YHR018C	ARG4						Argininosuccinate lyase
39	YPR002W	PDH1						Mitochondrial protein that participates in respiration
40	YER052C	HOM3						Aspartate kinase
41	YGL259W	YPS5						Protein with similarity to GPI-anchored aspartic proteases such as Yap1p and Yap3p
42	YNL078W	NIS1		yes			M/G1	Protein localized in the bud neck at G2/M phase
43	YBR083W	TEC1	yes				M/G1	Transcription factor required for full Ty1 epxression
44	YJR109C	CPA2				yes		Large subunit of carbamoyl phosphate synthetase
45	YKL096W	CWP1		yes			S/G2	Cell wall mannoprotein
46	YLR058C	SHM2					S/G2	serine hydroxymethyltransferase
47	YLR438W	CAR2					G2/M	L-ornithine transaminase (OTAse)
48	YKR075C							Unknown
49	YGL028C	SCW11	yes	yes	yes	yes	G1	Cell wall protein with similarity to glucanases
50	YIL162W	SUC2					G2/M	Invertase, sucrose hydrolyzing enzyme
51	YJR004C	SAG1					M/G1	Alpha-agglutinin of alpha-cells
52	YPL141C				yes	yes	S/G2	Unknown
53	YOL136C	PFK27						6-phosphofructo-2-kinase
54	YDR343C	NA		yes				Unknown
55	YLL012W	YEH1					G1	Steryl ester hydrolase
56	YOR236W	DFR1						Dihydrofolate reductase
57	YER124C	DSE1	yes	yes	yes	yes	G1	Daughter cell-specific protein, may participate in pathways regulating cell wall metabolism
58	YEL039C	CYC7						Cytochrome c isoform 2
59	YPL147W	PXA1						Subunit of a heterodimeric peroxisomal ATP-binding cassette transporter complex
60	YFL052W							Unknown

3.6 Discussions and Conclusions

3.6.1 Measurement selection

As mentioned, the measurement for gene expression is ratio in the first data set (Ca^{2+} response w/o FK506 inhibition), while the measurement is log transformed ratio in the second data set ($\triangle fkh1 \triangle fkh2/wtcellcycle$). In the first data, we want to identify calcineurin-dependent Ca^{2+} activated genes as done by Yoshimoto et al. Although there do exist some genes that are repressed by Ca^{2+}, we are more interested in Ca^{2+} activated genes as what Yoshimoto et al. did in their paper. So in order to make our results comparable with Yoshimoto's results, we reduce the influences of Ca^{2+} repressed genes by using ratio rather than log ratio as the measurement for gene expression. In such situation, the expression ratios for Ca^{2+} repressed genes are limited to $[0, 1]$, while expression ratios for Ca^{2+} activated genes are always greater than 1. Since we use Euclidean distance in calculating the change in neighborhood, the genes identified by MARD analysis tend to be genes that activated by Ca^{2+} in either FK506 treated or non-treated time courses, and most Ca^{2+} repressed genes are ignored. We note that this is a special case, in most cases we want to treat gene activation and repression equivalently and therefore log ratio should be used as the gene expression measurement.

3.6.2 Neighborhood selection

To identify genes that are differentially expressed between treatment and control time courses, we construct gene relationship networks for the the two time courses, respectively. The genes substantially affected by the treatment are expected to show dramatic changes in its neighbor genes. Essentially, here the neighbor genes refer in particular to those genes that have small Euclidean distances with a specific gene, namely, proximal neighbor genes. However if only proximal neighborhood changes are considered, one

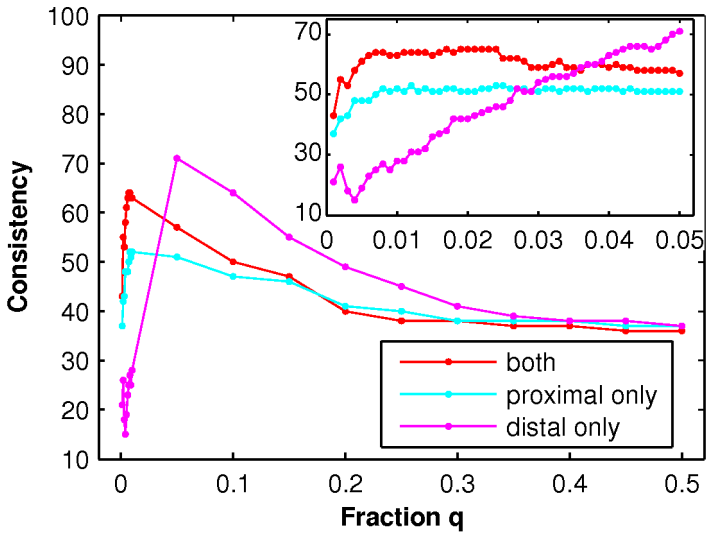

Figure 3.7: Effect of different neighborhood definitions and informative fraction q (proximal only: proximal neighborhood; distal only: distal neighborhood; both: two-end neighborhood). Zoom-in of the examined informative fraction q at interval [0.1%, 5%] is shown as an insert.

may fail to identify some co-affected genes because most of their proximal neighbor genes are similarly affected by the treatment and therefore there's no notable changes in the neighborhood. To make up this problem, we take advantage of those distant neighbor genes which have the largest Euclidean distances with the specific gene. We note that there is no underlying biological relationship between these distant neighbor genes. However, the distant neighbors of a gene tend to be from various function categories and widely distributed in the relationship network. So the change in the relationships between these distant neighbor genes and the specific gene may imply the global position change of the gene in the whole relationship network, which can not be captured by proximal neighborhood change. As shown in Figure 3.7, we studied the effectiveness of the three neighborhood definitions (proximal only, distal only and two-end neighborhood definition) while setting informative fraction q to range from 0.1% to 50%. MARD

analysis is performed on the Ca^{2+} response time courses with and without FK506 treatment. For each setting of the informative fraction q and neighborhood definition, the number of those genes that are among the top 142 genes in our result and also among those calcineurin dependent Ca^{2+} regulated genes reported by Yoshimoto et al. is calculated to measure the effectiveness of each neighborhood definition. The result shows that the distal neighborhood definition can achieve more effectiveness if a large q ($> 4\%$) is used. But when q is small, the effectiveness of the distal neighborhood definition is much worse than the other two definitions. In comparison, the proximal neighborhood or two-end neighborhood can achieve good effectiveness across a wide range of q. According to our experience of MARD analysis in various data sets, including two data sets not reported in this article, we suggest using both-end neighborhood definition. Our general strategy of selecting the fraction value q is as follows: first, we try MARD analysis for q in a range, say $[0.008, 0.05]$ as used in the above cases; second, we look for a stable set of genes that is invariant across the range of q values; third, we validate the function of these genes by scientific facts reported in the literature; fourth, we make further hypotheses based on the computational results. This strategy works well in the examples we have analyzed so far. We hope this bioinformatic methodology will benefit other researchers. We note that the informative fraction q for proximal and distal neighborhood do not necessarily need to be equal in the two-end neighborhood. Further improvement is expected by setting different values for them.

3.6.3 Metric selection

The relationship between genes can also be measured by other metrics besides the Euclidean distance. For example, Pearson correlation is often used to measure the similarity between expression profiles of genes, based on which gene co-expression

networks are constructed and used to predict gene functions, infer transcriptional regulatory networks, and so on [vNSH03, SSR+03]. In addition, comparing correlations between genes across experiments has been proposed to further improve these studies [ZKH+05]. Generally, the correlation can be applied to infer the gene relationship network in MARD analysis. But in practice, there are some disadvantages for using correlation as the metric. First, most of the time course data contain only a small number (< 10) of time points, therefore it is inappropriate to use correlation to measure the gene relationship. Second, several works have shown that co-expression network is scale free in topology [vNSH04, BO04], and the number of nodes with a given degree follows a power law distribution. In contrast to "random" networks, scale-free networks are highly non-uniform. In the gene co-expression networks, the hub genes have many co-expressed neighbors, while most other genes have only a few neighbors. This feature may be taken into account when correlation is used for the MARD analysis and some revisions may be required.

3.6.4 Robustness of MARD analysis

Like many other methods, MARD analysis is also sensitive to noise in the data set to some extent. For example, if an artificial high ratio is introduced by noise at one time point in either treatment or control time course, the corresponding gene may result in a high neighborhood change. Actually this is one of the main challenges of time course analysis. In general, it is hard to discriminate real gene expression change from noise effect because gene may be differentially expressed in only one time point in a time course. The noise effect can be reduced by average the replicates for each time points if replicate experiments are performed. Another way to reduce the noise effect is to increase the number of sampling time points so that a more accurate gene relationship network can be estimated from the time courses. We investigate the influence of time

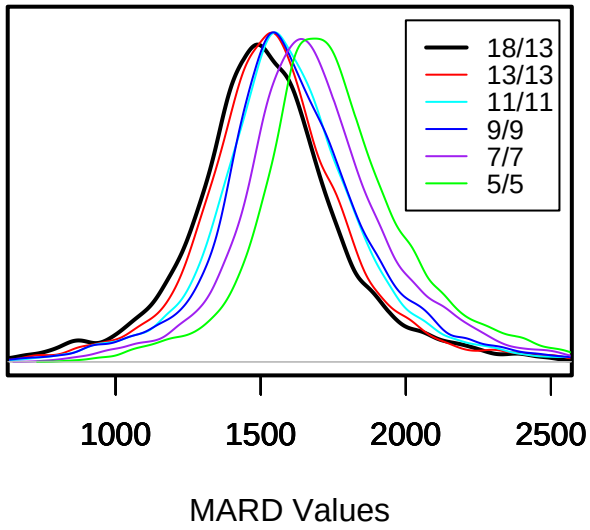

Figure 3.8: Distribution of MARD values from sampled data sets where a subset of time points in treatment and control time course are used.

points in the second data set, where 18 and 13 time points are sampled in the wild type and $\triangle fkh1 \triangle fkh2$ cell cycle experiment, respectively. We randomly choose 13 time points from the wild type time course, and perform MARD analysis for the resulting wild type and $\triangle fkh1 \triangle fkh2$ time courses, both of which have 13 time points. Then we randomly remove one time point from both wild type and $\triangle fkh1 \triangle fkh2$ time courses each time and applied the MARD analysis to the new data set. The rank of MARD value for each gene doesn't have significant change if similar number of sample time points are used in the two time courses. For example, all the C_{18}^{13} sampled 13/13 data sets (13 time points for either time course, C_{18}^{13} possible samplings in total) have a similar result with the original 18/13 data set. The Spearman correlations are generally greater than 0.9 between MARD values of the sampled data sets and those of the original data set. If only the top rank genes are considered, the results are even more consistent with each

other. On the other hand, a clear shift of the distribution of MARD values was observed when fewer time points are used, as shown in Figure 3.8. This indicates that estimation of the gene relationship network is more likely to be influenced by noise when fewer time points are used. We believe that MARD analysis is capable of giving more reliable result with the improvement of microarray technology.

3.6.5 Significance level of MARD values

As so far, all the analysis and results show that the MARD statistics does reflect the degree of perturbation of genes by a treatment. A higher MARD value implies a more severe perturbation. However it is difficult to assign a significance level to an observed MARD value because MARD values for all the genes are strongly dependent with each other. For example, if a gene is substantially affected by a treatment, the MARD values of its neighbor genes will also tend to be large. In addition, it is hard to perform permutation analysis for time courses as used in SAM [TTC01]. In a static microarray experiment, one permutates samples to ”balance” the case and control data sets and thereby estimate the false discovery rate based on the ”balanced” data sets. But in time courses data, different time points provide different aspects of gene expression. Therefore it is inappropriate to permute the time points to calculate the significance level of MARD for each gene.

3.6.6 Conclusion

We have developed a new method to identify differentially expressed genes between treatment and control time courses. Rather than comparing gene expression patterns in the two time courses directly, we construct gene relationship networks for each of the time courses and then measure the neighborhood change of each gene in the two

networks. The genes that are substantially affected by the treatment, i.e. differentially expressed genes, are those that have a remarkable neighborhood changes.

We applied our method to both aligned and un-aligned time course data sets. The results in the aligned data set show that (1) Genes with high MARD values exhibit different expression levels between treatment and control time course in all or a subset of time points; (2) The genes identified by our method are consistent with previous studies, where additional well-designed experiments are performed to ensure the accuracy of the result; (3) We also found some genes that are related to the pathway of interest but failed to be identified by previous approaches. Our method avoids direct comparison of expression pattern of genes between time courses, therefore it is insensitive to sampling effect. We do not require equal or "aligned" sampling time points in the treatment and control time courses. So our method can be used to compare time courses from different sources as shown in the un-aligned wt/$\triangle fkh1 \triangle fkh2$ cell cycle data set. In addition, the MARD value can roughly reflect the importance of a gene in the cell system. Genes with small MARD values tend to be house-keeping genes, most of which are essential for cell survival.

3.7 Application of MARD on *S.pombe* stress response data

In this section, we apply the MARD analysis on *S.pombe* stress response data. To study the transcriptional response of fission yeast to environmental stress, Chen et al. performed microarray experiment to characterize changes in expression profiles of all known fission yeast genes in response to five stress conditions: oxidative stress caused by hydrogen peroxide, heavy metal stress caused by cadmium, heat shock caused by temperature increase to 39°C, osmotic stress caused by sorbitol, and DNA damage

caused by the alkylating agent methylmethane sulfonate [CTM$^+$03]. Under each stress condition, a short time course microarray experiment is performed, including three time points at 0, 15 and 60 min, in wild type, $sty1\triangle$, and $atf1\triangle$ fission yeast. We combine the expression profiles at 15 and 60 min in the five stress condition into a long time course with 10 time points. The combination results in three time courses, corresponding to wild type, $sty1\triangle$, and $atf1\triangle$ fission yeast, respectively. Pairwise comparison of the three time courses with MARD analysis is performed to understand the function of Sty1 and Atf1 in stress response.

3.7.1 Transcriptional responses of fission yeast to stress

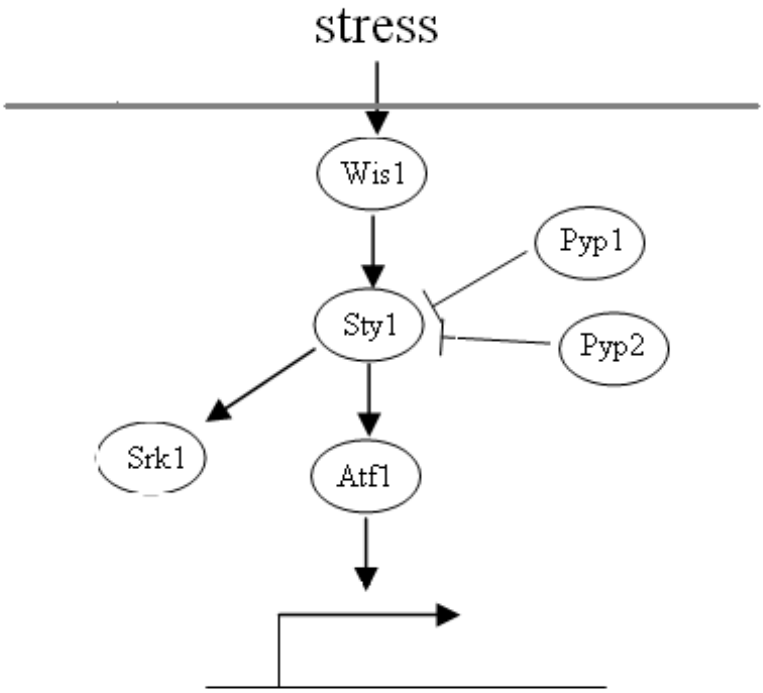

Figure 3.9: Sty1 stress response pathway in fission yeast.

Exposure to low level of stress often results in a transient resistance to higher level of the same stress, as well as to the other types of stress. The cross protection is short-lived and requires new protein synthesis, indicating that changes in gene expression are critical. In *S.pombe* this phenomenon is a consequence of a common stress response pathway, the Sty1 MAPK pathway, which regulates the responses to different stresses [CTM$^+$03]. It is known that in both budding and fission yeast, a similar core group of genes response to all or most stresses. These genes are mainly regulated by stress-specific mechanisms in budding yeast, whereas in fission yeast they response to different stresses through the common Sty1 MAPK pathway. As shown in Figure 3.9, Sty1, a mitogen-activated protein kinase, is activated by WIS1 kinase in response to stress, which then stimulates transcriptional responses through a number of bZip transcription factor, including Atf1, Pcr1, and Pap1 [GDSP98, DSH$^+$04, DSWN$^+$05]. Among these transcription factor, Atf1 is the most well studied. It is constitutively localized in the nuclear and activated by Sty1 kinase through phosphorylation. Two phosphatases, Pyp1 and Pyp2, act as negative regulator of the pathway by inactivating the Sty1 kinase through dephosphorylation [NS99].

3.7.2 Results and conclusions

We perform MARD analysis to three time courses in a pairwise manner, which results in three groups of MARD values, corresponding to $sty1\triangle/wt$, $atf1\triangle/wt$ and $sty1\triangle/atf1\triangle$, respectively. The histograms of the MARD value in the three time course comparison are shown in Figure 3.10. The histograms in $sty1\triangle/wt$, $atf1\triangle/wt$ have long tails on the right hand, whereas the tail of the histogram in $sty1\triangle/atf1\triangle$ is much shorter. This phenomenon can be explained by the fact that Sty1 kinase regulate the gene expression mainly through the transcription factor Atf1 in response to stress [GDSP98]. The high similarity of the MARD values in $sty1\triangle/wt$ to those in

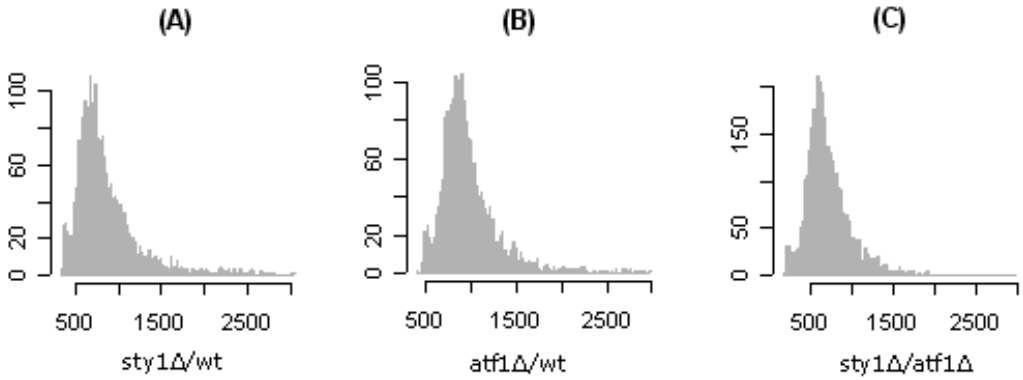

Figure 3.10: Histograms of MRAD values in (A) $sty1\triangle/wt$, (B) $atf1\triangle/wt$ and (C) $sty1\triangle/atf1\triangle$.

$atf1\triangle/wt$ is also shown in their scatter plot (Figure 3.11). As shown, most genes follow into a high dense region in the middle part of the figure. These genes are not perturbed by the deletion of sty1 or atf1 gene. The small group of genes in the bottom-left corner are likely to house-keeping genes whose expression keep constant in different conditions. Those genes that close to the top-right are of interest, which are substantially affected by both sty1 and atf1. The genes that are substantially affected only by sty1 or atf1 are also what we are interested. For example, genes affected by sty1 but not by atf1 imply that transcription responses of these genes are regulated by Sty1 kinase but independent of Atf1, which may through another transcription factor such as PCR1 or PAP1.

In Table 3.3, we list the ranks of MARD values for 49 genes in three time course comparisons: $sty1\triangle/wt$, $atf1\triangle/wt$, and $sty1\triangle/atf1\triangle$. In general, if a gene has high rank (small number)in both $sty1\triangle/wt$ and $atf1\triangle/wt$, it tends to have low rank (large number) in $sty1\triangle/atf1\triangle$ (see the pyp1 gene); if a gene has high in either $sty1\triangle/wt$ or $atf1\triangle/wt$ but low rank in the other, it is likely to have a high rank in $sty1\triangle/atf1\triangle$ (see the pyp2 gene). However, there are some special cases, such as the sou1 gene. It

Table 3.3: Ranks of the MARD values for 49 genes in fission yeast.

ORF	Name	sty∆/wt	atf∆/wt	sty∆/atf∆	Function
SPAC8E11.10	sou1	1	87	22	sou1, short chain dehydrogese
SPCC285.05		2	391	150	conserved protein
SPBC660.05		3	1299	1	hypothetical protein
SPAC19D5.01	pyp2	4	3025	8	pyp2, tyrosine phosphatase Pyp2
SPAC26F1.10C	pyp1	5	2	2027	tyrosine phosphatase Pyp1
SPAC1399.02		6	158	259	membrane transporter
SPAC3H1.06C		7	100	320	membrane transporter (predicted)
SPAC2H10.01		8	164	2216	transcription factor
SPAC4A8.03C	ptc4	9	204	60	protein phosphatase 2C Ptc4
SPACUNK4.17		10	135	42	dehydrogese (predicted)
SPCC1393.12		11	7	1414	sequence orphan
SPAC1F8.06	fta5	12	844	18	fta5, Sim4 and Mal2 associated protein 5, sma5
SPBC713.11C		13	1	944	UPF0057 family
SPCC1322.08	srk1	14	1983	2	srk1, MAPK-activated protein kise Srk1, mkp1
SPAC22A12.06C		15	96	255	serine hydrolase
SPAC4H3.03C		16	1245	16	glucan 1,4-alpha-glucosidase (predicted)
SPAC11E3.14		17	3154	10	conserved fungal protein
SPAC1F3.09		18	29	706	conserved eukaryotic protein
SPCC4B3.10C	ipk1	19	48	1085	inositol 1,3,4,5,6-pentakisphosphate (IP5) kinase
SPCC757.03C		20	133	2057	
SPCC1020.09		21	50	1446	WD repeat protein
SPCC794.08		22	10	2769	hypothetical protein
SPBC31A8.01C	cwh1	23	22	1475	reticulon-like protein
SPBC713.02C	ubpD;ubp21	24	427	1350	ubiquitin C-terminal hydrolase Ubp21
SPBPB2B2.02		25			hypothetical protein
SPBPB2B2.13		26	8	1590	galactokise (predicted)
SPBC660.06		27	1468	30	hypothetical protein
SPBC1289.14		28			adducin N-termil domain protein, SPBC8E4.10c
SPAC32A11.02C		29	1605	112	
SPAC15E1.02C		30	62	574	
SPAC22A12.17C		31	682	95	short chain dehydrogenase (predicted)
SPAC13G7.13C		32	743	648	R-binding protein
SPBC8E4.05C		33		284	3-carboxy-cis,cis-muconate cycloisomerase
SPCC965.14C		34	39	2005	cytosine deaminase (predicted)
SPAC26F1.14C		35	17	1043	apoptosis-inducing factor homolog Aif1
SPBPB2B2.01		36	43		amino acid permease family
SPBC12D12.02C	cdm1	37			D polymerase delta subunit Cdm1
SPCC1223.13		38	276	233	D binding protein (inferred from context)
SPAC5H10.02C		39	1308	2208	
SPAC1786.01C		40	172	54	triacylglycerol lipase
SPBC336.12C	cdc10	41	37	188	MBF transcription factor complex subunit Cdc10
SPCC285.09C	cgs2;pde1	42	2790	113	cAMP-specific phosphodiesterase Cgs2
SPAC23H3.15C		43	73	607	
SPAC11G7.01		44	290	488	glycoprotein (predicted)
SPBC3H7.05C		45	231	145	
SPAC637.13C		46	654	40	cytoskeletal signaling protein
SPBC36.02C		47	5	1881	membrane transporter
SPBC409.08		48	525	1643	membrane transporter
SPAC6F6.17	rif1	49	1831	49	rif1, telomere length regulator protein Rif1, tap1

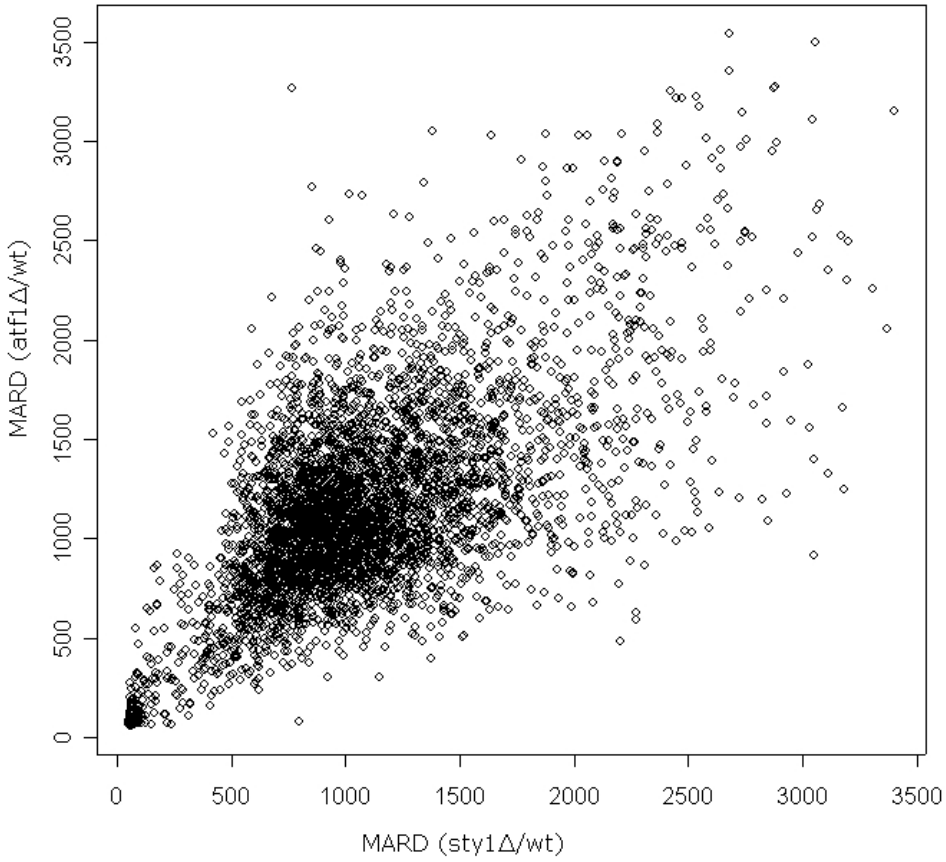

Figure 3.11: Scatter plot of MARD values in $sty1\triangle/wt$ versus those in $atf1\triangle/wt$.

has high rank in all the three comparisons, which indicates that expression of sou1 is affected by both sty1 and atf1 by in a different manner.

Figure 3.12 shows the expression values of the genes that have high ranked MARD values in $sty1\triangle/wt$. The left 10 bars correspond to the time course in wild type and the right 10 bars correspond to that in $sty1\triangle$. From to left to right, these 10 bars represent the log expression values in H_2O_2(15min), H_2O_2(60min), Cd(15min), Cd(60min), Heat(15min), Heat(60min), Sorb(15min), Sorb(60min), MMS(15min), and

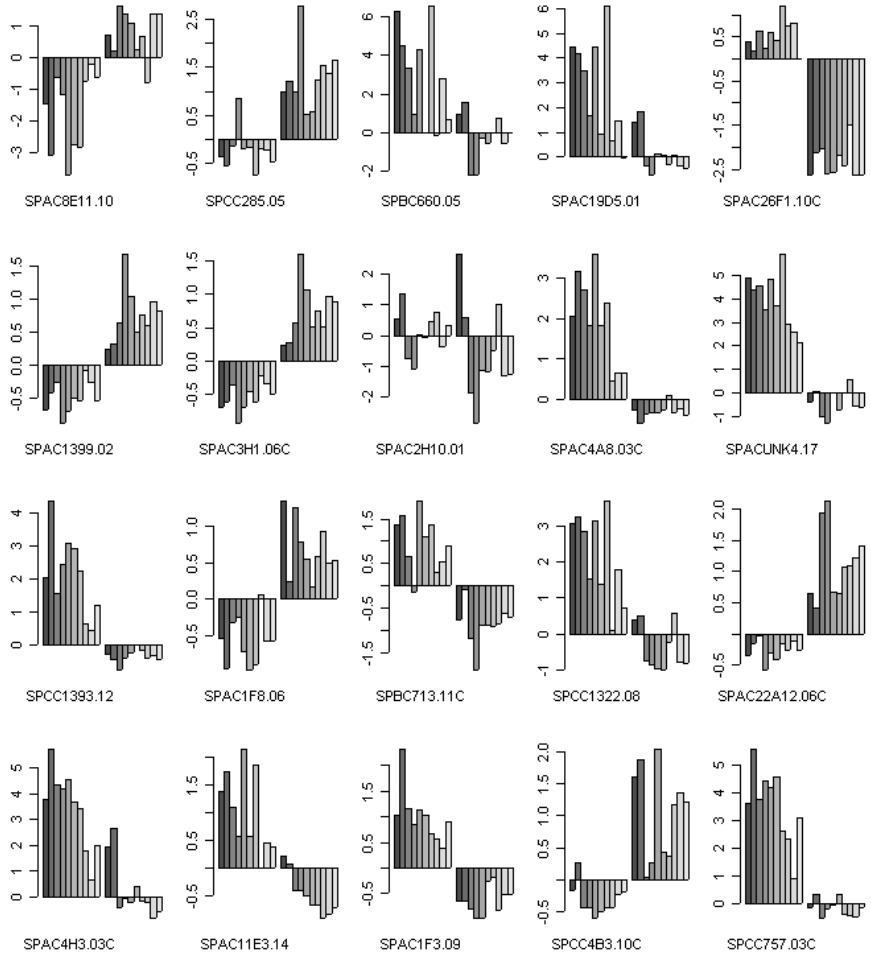

Figure 3.12: Expression values of 20 genes in wild type and $sty1\triangle$ fission yeast. These genes have the top 20 genes identified by MARD analysis in $sty1\triangle/wt$ time course comparison.

MMS(60min), respectively. As can be seen, these genes that identified by MARD analysis do exhibit differentially expressed patterns between the wild type and $sty1\triangle$ time course. Similar results have been obtained in MARD analysis for the other two time course comparisons: $atf1\triangle/wt$ and $sty1\triangle/atf1\triangle$.

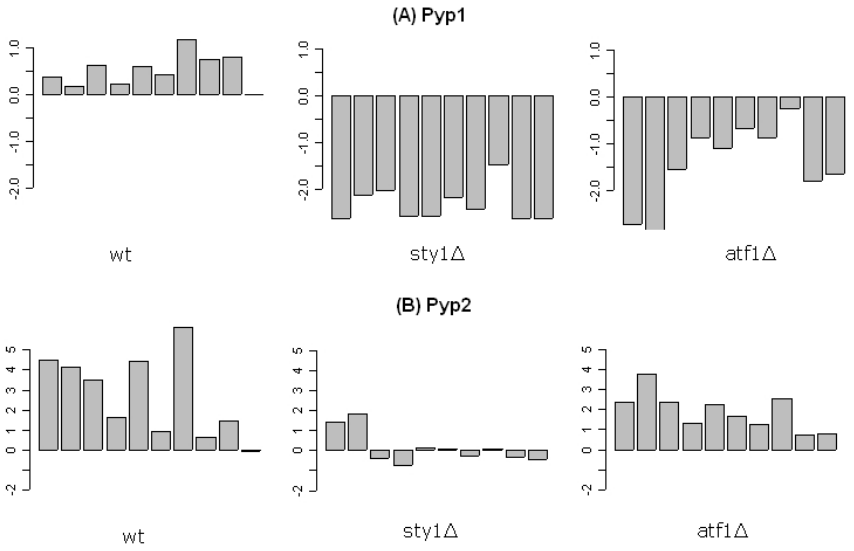

Figure 3.13: Expression values of pyp1 and pyp2 in time course corresponding to wild type, $sty1\triangle$ and $atf1\triangle$, respectively.

Expressions of pyp1 and pyp2 are regulated by different mechanisms

It is known that two tyrosine-specific phosphatases, Pyp1 and Pyp2, negatively regulate the activity of Sty1 kinase via direct dephosphorylation of the tyrosine residue phosphorylated by the Wis1 kinase. Among the two phosphatases, Pyp1 accounts for the major cellular activity that dephosphrylates Sty1 kinase and Pyp2 plays a minor role [NS99]. Phosphatase activities of Pyp1 and Pyp2 are inhibited in stress conditions. MARD analysis indicates that pyp1 and pyp2 are under quite different transcriptional regulation in response to stresses. Expression of pyp1 is substantially affected by both Sty1 kinase and the transcription factor Atf1, with a rank of 5 and 2 in $sty1\triangle/wt$ and $atf1\triangle/wt$ time course comparison (in total, there are about 4410 genes), respectively. Whereas, expression of pyp2 is substantially by the Sty1 kinase with a rank 4 in $sty1\triangle/wt$ comparison, but not by Atf1 with a rank of 3025 in $atf1\triangle/wt$ comparison. Figure 3.13 shows the expression values of pyp1 and pyp2 in the wild type, $sty1\triangle$ and $atf1\triangle$ time course. The stress conditions are arranged from left to right in the same order as above

described. Based on our MARD analysis and the expression patterns of pyp1 and pyp2 in the three time courses, we may figure out the following hypothesis. First, in wild type fission yeast, activation of the Sty1p kinase pathway by stress leads to up-regulation of both pyp1 and pyp2. Second, although Pyp1 plays the major role in dephosphorylation of Sty1 kinase, expression of pyp2 increases (over 30-fold) much more than that of pyp1 (less than 2-fold)as a result of the stress response. That is, pyp2 plays the key role in the negative feedback loop. Third, expression increase of pyp1 depends on both sty1 and atf1; deletion of either of them cause significant down-regulation of pyp1 by up to 7-fold. Fourth, expression increase of pyp2 requires activity of Sty1 kinase but is independent of Atf1. This may suggest up-regulation of pyp2 by Sty1 kinase is via another transcription factor other than ATf1. Taking together, we may construct a regulatory model for pyp1 and pyp2 as shown in Figure 3.14.

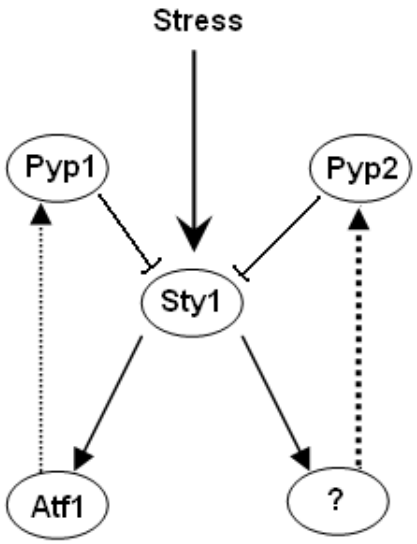

Figure 3.14: Function and regulation of pyp1 and pyp2 in the Sty1 stress response pathway in fission yeast.

This model provides a good example that shows how delicate the natural designs could be. In response to stress conditions, a group of genes are induced in fission yeast.

Among them, there are some genes that negatively regulate the stress response pathway, which ensures that the pathway can be rapidly shut down as soon as the stress has been removed. To achieve this purpose effectively, both pyp1 and pyp2 are used. Pyp1 phosphatase plays the major role in dephosphorylation of Sty1 kinase but expression of pyp2 increases more in the negative feedback loop. Moreover, up-regulation of pyp1 and pyp2, though both depending on Sty1, are via different transcription factors, which enhances the robustness of the feedback loop.

Expression of pcr1 is affected by sty1 but not by atf1

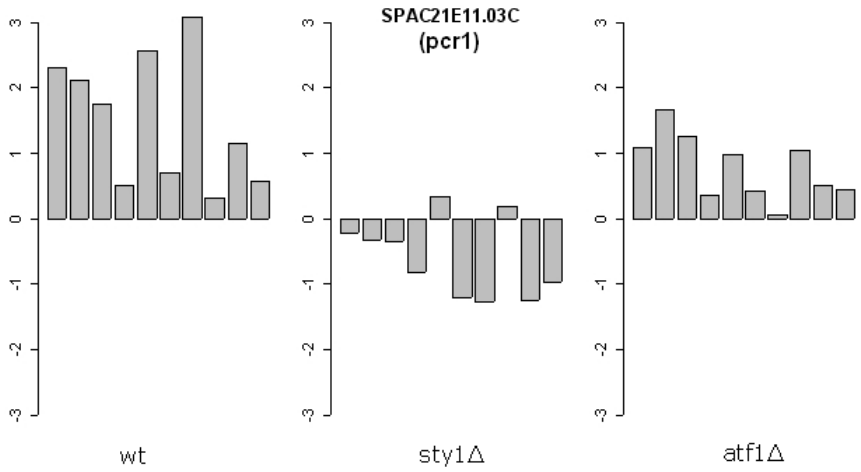

Figure 3.15: Expression values of pcr1 in time course corresponding to wild type, $sty1\triangle$ and $atf1\triangle$, respectively.

Pcr1 is another bZip transcription factor that act downstream of Sty1 kinase in fission yeast stress responses. MARD analysis suggests that expression of pcr1 is dependent on Sty1, but not on Atf1. The ranks of MARD values for pcr1 in $sty1\triangle/wt$, $atf\triangle/wt$, and $sty1\triangle/atf\triangle$ are 103, 1599 and 7, respectively. Expression values of pcr1 in those three time courses are shown in Figure 3.15. These results indicate that Pcr1 acts downstream of Sty1 kinase in parallel with Atf1.

Expression of srk1 is affected by sty1 but not by atf1

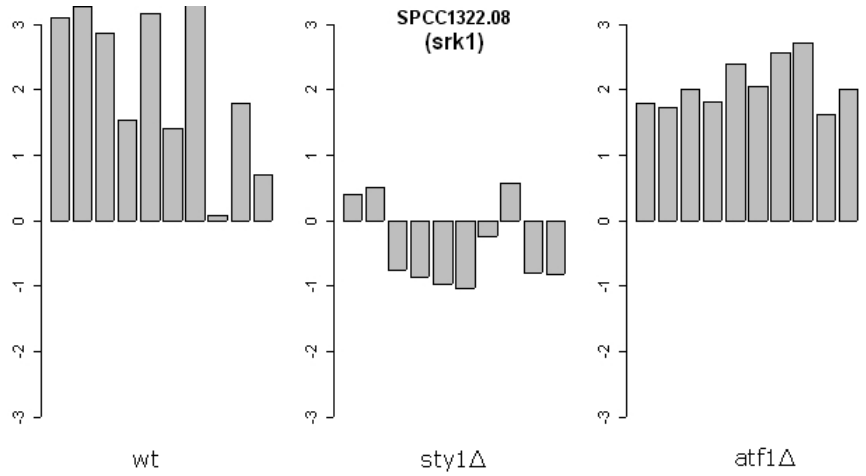

Figure 3.16: Expression values of srk1 in time course corresponding to wild type, $sty1\triangle$ and $atf1\triangle$, respectively.

Srk1 is also a kinase that is involved in fission yeast stress responses. Previous studies have shown that it presents in a complex with the Sty1 kinase and is directly phosphorylated by Sty1 [STC^{+}02]. Our MARD analysis results are consistent with these studies. AS shown in Figure 3.16, expression of srk1 requires the activity of Sty1 but in a Atf1 independent manner. The ranks of MARD values for srk1 in in $sty1\triangle/wt$, $atf\triangle/wt$, and $sty1\triangle/atf\triangle$ are 14, 1983 and 2, respectively.

Expression of ptc4 is affected by both sty1 and atf1

Other than Pyp1 and Pyp2, type 2C serine/threonine phosphatase (PP2C) also involved in dephosphorylation of hence inactivation of Sty1 kinase [NS99]. Interestingly, MARD analysis indicates that expression of ptc4, the gene that encodes PP2C, is also up-regulated like pyp1 and pyp2 as a result of stress response in fission yeast. The up-regulation of ptc4 depends on both sty1 and atf1; the ranks of MARD values are 9 and 60 in $sty1\triangle/wt$ and $atf1\triangle/wt$, respectively (see Figure 3.17). Therefore, the

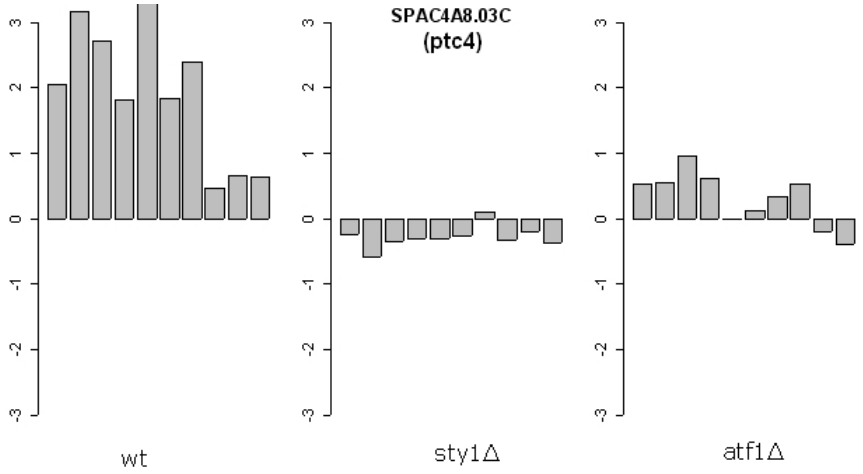

Figure 3.17: Expression values of ptc4 in time course corresponding to wild type, $sty1\triangle$ and $atf1\triangle$, respectively.

up-regulation of ptc4 may reveals another negative feedback loop in fission yeast stress response. Moreover, other than stress response, PP2C is also involved in many other pathways fission yeast. These findings imply a possible mechanism that connect stress responses to other pathways.

MBF may link stress response with cell cycle regulation

In fission yeast, activation of the stress response pathway leads to a inhibition of entry into mitosis. The stress response pathway also promotes commitment to mitosis in unperturbed cell cycles to allow cells to match their rate of division with nutrient availability [CTM$^+$03, SP95]. The nature of the stress response pathway in cell cycle control is not fully understood. Recently, several possible mechanisms have been proposed. López-Avilés et al. proposed that stress activated Srk1 kinase blocks mitotic entry by phosphorylating and inhibiting Cdc25 [LAGG$^+$05]. Petersen et al. suggested that Polo kinase linked the stress pathway to cell cycle control and tip growth [Pet]. Our analysis implies that MBF complex may also involved in the mechanism that links stress response to cell cycle control in fission yeast.

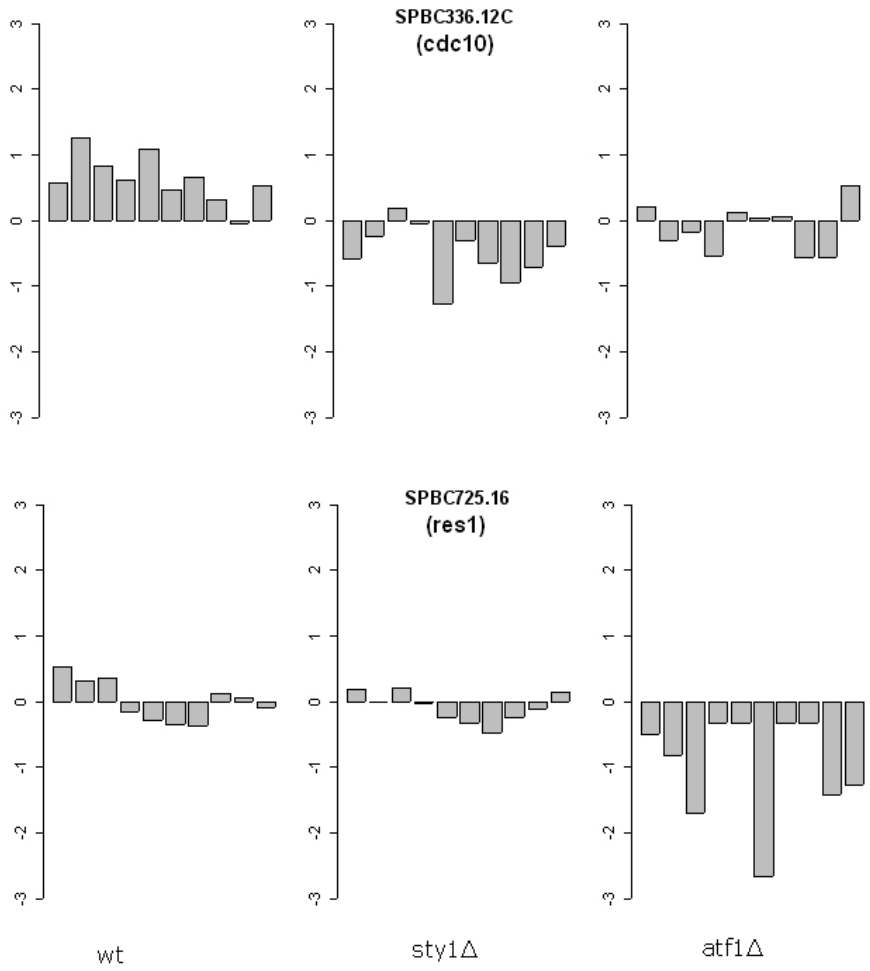

Figure 3.18: Expression values of cdc10 and res1 in time course corresponding to wild type, $sty1\triangle$ and $atf1\triangle$, respectively.

In eukaryotic cells, a key regulatory step of the cell cycle entry occurs at late G1, which has been termed "Start" in yeast. Mitotic entry through Start requires the activity of one or more cyclin-dependent kinases (CDKs) and also the transcription activation of specific genes encoding products for S phase [WSDJ99]. In fission yeast, transcriptional activation at Start is mediated by MBF complex. The fission yeast MBF complex contains Cdc10p and at least two additional proteins, Res1p and Res2p, which bind to Cdc10 at their C-termini. It has suggested that cdc10 plays both positive and negative

roles in cell cycle gene expression [MKCF95]. Res1 and Res2 are highly related but functionally non-identical. The $res1\triangle$ cells have deficiency in mitotic cycle and a cold- and heat-sensitive phenotype resulting in a G1 arrest [TOO+92].

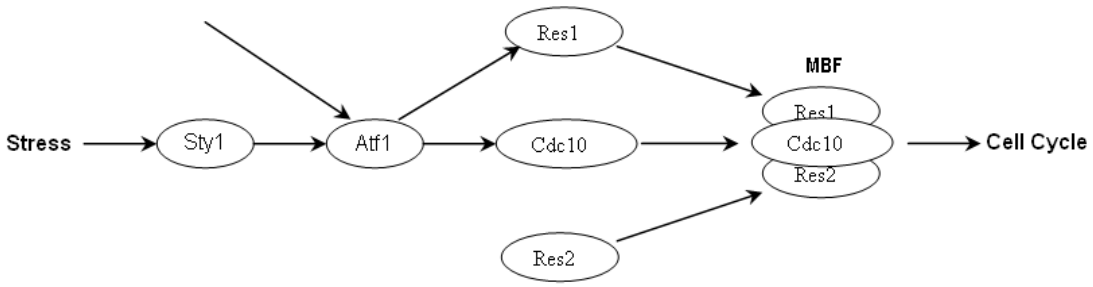

Figure 3.19: A possible mechanism that link stress response to cell cycle control by MBF.

MARD analysis indicates that expression of cdc10 depends on both Sty1 and Atf1; the ranks of MARD values are 41 and 37 in $sty1\triangle/wt$ and $atf1\triangle/wt$, respectively. Expression of res1 also depends on Atf1 activity but in a Sty1-independent manner with a rank of 122 for the MARD value in $atf1\triangle/wt$. Whereas expression of res2 is not affected by both Sty1 and Atf1. Figure 3.19 shows the expression values of cdc10 and res1 in wild type, $sty1\triangle$ and $atf1\triangle$ time course. These results implies another possible mechanism mediated by the MBF complex that links stress response to cell cycle control. Note that the gene expression changes of cdc10 and res1 in response to stress conditions are not large in magnitude, but their differential expressions in various time courses are detected using MARD analysis. In some sense, this reveals that MARD is more sensitive than the point-wise comparison method.

In conclusion, application of MARD analysis on the fission yeast stress response data set reveals possible transcription regulatory mechanisms for many genes. Some of

them are already known thanks to previous studies. The others are still lack of experimental evidences or literature supports. Therefore, the hypothesis based on MARD analysis may provide us some hints and directions for future biological studies.

Chapter 4

Integrative analysis of long-lived yeast mutants

Microarray technology provides a powerful tool for biological studies. It measures expressions of thousands of genes from samples at the same time. Despite its great success, microarray technology has its own inherent limitations. For example, case and control design is frequently used in microarray experiment to compare gene expression profiles in different samples. To understand the underlying mechanisms of certain disease, typically we collect samples from both patients and non-patients, and then we perform microarray experiment to measure gene expressions in the samples. This is what is so called case and control experiment design. For data sets from this kind of experiment design, differentially expressed genes between cases (samples from patients) and controls (samples from non-patients) are often identified, which provide us a list of candidate genes that are potentially related to the disease. The limitations in this kind of studies are: first, it only gives us some genes that change expressions in diseased people compared to healthy people, but diseases are often associated with changes in certain pathways. So a differentially expressed gene list is not enough to infer the mechanism of the disease and analysis in higher levels, i.g. in pathway level, is required. Secondly, the differentially expressed genes are more likely to be a consequence or side-effect of the disease rather than the causal genes which are of more interest. Thirdly, the differentially expressed genes are those genes that are significantly changed in diseased

people. But the strength of relevance is not fully determined by magnitude of expression change. A small magnitude of expression changes for important genes may cause substantial effect.

In this chapter, we perform analysis for the microarray data from our yeast ageing project [Pro]. Collaborating with the laboratory lead by professor Long, we measured gene expression profiles of four yeast knockout strains, as well as wild type yeast. These knockout strains result in significant life span extension with respect to wild types. The goal of the project is to investigate the mechanisms of longevity in these strains. To overcome the problems mentioned above, we emphasize the importance of investigating expression changes in high levels, for example in pathway level. Moreover, we will show how high level analysis is facilitated by integrating large scale public data sets from different sources, such as Gene Ontology (GO), ChIP-chip, gene localization, and so on. In this chapter, we will first introduce the background knowledge and theories about ageing; then we will describe the integrative analysis for our microarray data set and show how this analysis exposes a common mechanism of longevity in four long-lived yeast knockout strains.

4.1 Introduction to ageing

Ageing occurs in organisms ranging from yeast to humans. It describes all the changes that occur in the molecules, organelles, cells, tissues, and organs of an organism.

4.1.1 Theories of Ageing

A number of theories have been proposed to explain the mechanism of ageing. In the following section, we will briefly introduce several of them. Note that these theories provide different but overlapping viewpoints with each other.

The free radical theory

The free radical theory of ageing was first proposed by Harman in the 1950s [Har56]. In this theory, he suggested that aging was a consequence of free radical damage. Later Harman extended the idea to implicate mitochondrial production of ROS in the 1970s [Har72]. According to this theory,ageing of organisms are caused by accumulation of free radiacl damages in protein, lipid and nucleic acids (DNA, RNA) across time. Free radical attack on protein, lipid and nucleic acids leads to a reduction in their respective function, thereby decreasing cell function, then organ function, and finally, organismal function. In biochemistry, the free radicals of interest are often referred to as reactive oxygen species (ROS). ROS are generated in multiple compartments and by multiple enzymes in the cell. These enzymes that contribute to the generation of ROS include plasma membrane proteins, such as NADPH oxidases; enzymes that involved in lipid metabolism within the peroxisomes; as well as various cytosolic enzymes, such as cyclooxygenases. Although all these sources contribute to the overall intracellular ROS generation, the majority of them are produced in mitochondria, as by-products of oxidative phosphorylation.

The disposable soma theory.

The disposable soma theory of ageing was first introduced by Weismann and later developed by Kirkwood et al. [Kir88, Kir92, Kir02]. The basic idea of the theory is that cell maintenance, such as DNA repair, protein turnover, and antioxidant defenses, requires caloric energy. Competition of this with metabolic demands for energy have forced natural selection into an optimization process which compromises between longevity and growth or Reproduction. In most time, using extra energy to increase reproductive capacity will be more beneficial from an evolutionary standpoint, because it will enhance the fitness of that individual. Therefore, organisms have evolved in such a way

that the amount of energy invested in maintaining the soma is sufficient to keep the animal alive long enough to reproduce but less than that would be required to keep it alive indefinitely. Consistent with this theory, trade-offs, such as decreased fertility or growth, are observed in most but not all long-lived mutant organisms. However this theory conflicts with the fact that caloric restriction (CR) extends life span in many species.

The accumulated mutation theory

The theory was proposed by Medawar in 1952. The centrol idea of this theory is that the force of natural selection decreases with age increasing [Par01]. For a deleterious mutation that manifests itself at a young age, there will be strong selection pressure to eliminate it. But mutations that cause deleterious effect in later life of an organism, can be passed from generation to the next and may accumulate in the genome due to the weakness of selection force.

The antagonistic pleiotropy theory

The theory was proposed by Williams in 1957. It suggests that genes exist which have beneficial effects early in life but harmful effects later in life. If these genes confer increased reproductive success early in life, they would be selected despite the fact that they may cause a decline in vitality late in life. According to this theory, we can deduce that mutations resulting in life span extension would cause defects in growth or fertility. However this is not always true. For example, some daf-2 mutants in C.elegans, survive for more than twice as long as wild type but grow and reproduce normally [LF03].

The programmed ageing theory

The programmed and altruistic ageing theory claims that ageing is programmed so that organisms age and die to benefit related individuals or their group [LMS05]. According

to the theory, an ageing and death program that benefits closely related organisms can be explained by kin selection; and death for the benefit of unrelated organisms can be explained by group selection. Theoretically, ageing could provide long term benefits at the group or population level that include population stabilization, enhanced genetic diversity, a shortening of the effective generation cycle and acceleration of the pace of adaptation [FBV+04]. Local extinctions from overpopulation might facilitate a kind of population-level selection that is strong and rapid enough to offset the individual costs of programmed ageing.

4.1.2 Ageing in yeast

Replicative ageing

Replicative life span is defined as the total number of daughter cells generated by a mother cell. For budding yeast, the mother cells reproduce asymmetrically by originating buds, which finally separate from the mother cells and grow into daughter cells. The daughter cells are smaller than mother cells and can be easily recognized. The mother cells become old and stop producing new buds after a certain number of divisions. But the daughter cells do not inherit the senescence from the mother cells and have the potential to live a full life span. To measure the yeast replicative lifespan, cells are initially spread at low density onto growth medium agar and incubated to allow bud emergence. Newly born daughter cells are micromanipulated to fresh areas of the plate. The lifespan is determined by counting and removing the buds that they produce, until they don't bud any more.

The most commonly accepted explanation for the replicative ageing of yeast is the accumulation of extrachromosomal ribosomal DNA circles (ERC) in old mother cells [SMG97]. ERCs are self-replicating units produced in the nucleolus by rDNA

homologous recombination. Because they segregate in a highly biased manner to mother cells during cell division, they are accumulated in mother cells in proportion to the number of cell divisions. The segregation bias also assures that daughter cells are ERCs free and therefore live a full life span. After a certain of cell divisions, the old mother cells contains too many ERCs, which may interfere with cell growth by titrating essential replication and transcription factors and result in replicative ageing. In consistent with this model, yeast proteins Sir2 has been suggested to slow the replicative ageing of yeast by repressing mitotic and meiotic recombination between rDNA repeats and thereby preventing the formation of ERCs [PDG99, MMG99]. Over-expression of Sir2 extends the replicative life span whereas the deletion of SIR2 gene decreases replicative longevity [MMG99]. On the other hand, mutation of another protein Fob1, which increase the accumulation of ERCs by facilitating rDNA recombination, extend the replicative life span [DPK$^+$99].

To date, about 50 genes have been found to regulate replicative ageing. These genes involve in different but interrelated biological processes, such as stress response, genome stability, telomere function, energy metabolism, mitochondrial segregation and so on. This reflects the complexity nature of mechanisms underlying yeast replicative ageing.

Chronological ageing

The other system to measure yeast longevity has been developed by Longo's laboratory [LGV96]. It measures the capacity of a population of non-dividng yeast to maintain viability over time. Yeast can enter different non-dividing phases which depend on the type and the level of nutrient available in the medium. In SDC medium, which contains a limited amount of nutrients, yeast cells grow rapidly and then survive at high metabolic rates for about six days. If yeast cells growing in SDC are switched to water between

day 1 and 5, metabolic rates decrease and survival is extended by 2-3 fold. Incubation of yeast in water can be viewed as a form caloric restriction. In SDC medium, yeast cells enter a high-metabolic post-diauxic phase; whereas in water they enter a hypo-metabolic stationary phase [Lon03]. Its seems that survival is regulated by analogous pathway and mechanism in SDC median and in water, even though the cells are in different phases, because long-lived mutants isolated by incubation in SDC also have a longer chronological life span when incubated in water. Therefore the chronological life span can be measured in either system.

Replicative ageing model of yeast may be useful for understanding the aging of dividing cells of high eukaryotes, while the chronological ageing model of yeast may be informative of events in post-mitotic cells. Moreover, the chronological life span models ageing of yeast in natural environment because it measures the survival of yeast population in a non-dividing states. Despite of this difference, replicative ageing and chronological ageing are highly related with each other. First, both forms of ageing are characterised by a progressive deterioration in replicative potential that culminates in a post-mitotic phenotype that may be termed senescence. Both forms of post-mitotic cells exhibit surface wrinkling and an increased cell size [BS96, MKHS03]. Second, chronologically aged cells exhibit impaired replicative longevities and vice versa [ASGG99]. Third, most of the genes that effect on chronological ageing also involved replicative ageing. For example, deletion of Sch9 gene leads to extension of both replicative life span and chronological life. However, these two forms of ageing may have different metabolism. First, some genes have converse effects on them. Deletion of Ras2 causes extension of chronological life span but reduces the replicative life span [Lon03]. In additional, It has been demonstrated that Sir2 activity correlates with yeast replicative life span: SIR2 deletion strains are short lived, whereas strains that overexpress SIR2

are long lived [MMG99]. Sir2 promotes replicative longevity by repressing the recombination of repetitive ribosomal DNA (rDNA) and the subsequent formation of extrachromosomal rDNA circles (ERCs). Sir2 decreases chronological longevity, perhaps by promoting DNA damage, inhibiting stress resistance, and/or inhibiting the activation of the alcohol dehydrogenase, Adh2 [FGB$^+$05].

4.1.3 Sch9, Ras2, Tor1, Sir2 and ageing

Conservation and ageing

Mutations in genes that affect a wide range of biological processes have been found to change life spans in model organisms. The biological processes include endocrine signaling, stress response, metabolism, and telomere function, et al [Ken05]. In *Saccharomyces cerevisiae*, about 50 genes have been identified as ageing related genes [KKFK04a, KK05]. Despite their effect on ageing, these genes have different functions. For example, Phb1 and Phb2 encode subunits of prohibitin complex, which is involved in mitochondrial segregation [PJB$^+$02]; Dna2, Ctf4 and Rad27 encode proteins that play roles in maintaining genome stability [HBC$^+$02]; Lag1 is involved in ceramide biosynthesis [DCF$^+$94]; Sod1, Sod2, Msn2, and Msn4 are stress response genes; Hex2 encodes the hexokinase isoenzyme 2 that catalyzes phosphorylation of glucose [KK05]. In addition, the effects of these genes on ageing are also dependent on the genetic background of yeast strain and the type and level of nutrients in the medium. Taken all these into consideration, it's very difficult to explain ageing using a universal model. To understand the nature of yeast ageing, we must use some strategies to simplify the problem.

One strategy is to take advantage of the conservation of genes that affect ageing across different species. It seems that some genes are associated with ageing only in

certain organisms. These genes obscure the common mechanism of ageing and complicate the ageing studies. To overcome this problem, we may focus on processes or genes that are found to be associated with ageing in different organisms. One may argue that the causes of aging are not likely to be conserved from one organism to another, because they are not adaptive. This claim is true. Ageing factors may not subject to natural selection directly for the lack of selection force. However, the regulation of life span in response to environmental conditions is adaptive and therefore very likely to be conserved.

Calorie restriction and ageing

Following this strategy, we may center our attention to calorie restriction (CR) and three relevant gene or gene groups. CR has been shown to achieve extension of life span in a broad spectrum of organisms ranging from yeast to mammalian [LDG00, IAdC$^+$04]. The animals under CR are characterized by lower body temperature, lower blood glucose and insulin level, and reduced body fat and weight [KG03]. The CR animals also appear to be more resistant to external stresses, including heat and oxidative stress [SW96]. Evolutionarily, CR may represent adaptation to scarcity in a boom and bust cycle. Any organism that could slow ageing and reproduction in times of scarcity and remain able to reproduce when food reappeared would enjoy a competitive advantage over neighbors that could not [HA89, GP05].

Sir2

Despite the controversy and uncertainty, three genes or gene groups may be relevant to CR. The first gene is Sir2 [RH04, Gua05]. Previous studies have shown that the life span of short-lived strain lacking Sir2 can not be extended by CR, which imply that Sir2

is required for life span extension by CR [LDG00]. Other evidences have been presented to show that CR and Sir2 act in different genetic pathways to promote longevity and that Sir2 is not required for full life span extension in response to CR [KKFK04b]. Later Lamming et al. reported that CR fails to extend life span in a strain lacking both SIR2 and HST2 at 0.5% glucose. They concluded that CR extends life span by reducing rDNA recombination and ERC formation in a SIR2- and HST2-dependent fashion [LLEM$^+$05, LLEM$^+$06]. But another group obtained a conflicting result using the same genetic background. Their result shows that CR is still able to extend life span both in yeast strains that lack Sir2, Hst2, and Fob1 and in yeast that also lack Hst1 [KSH$^+$06]. Although the relationship between Sir2 and CR may continue to be debated, it is generally accepted that Sir2 plays some roles in ageing in different organisms. An extra copy of Sir2 extends yeast replicative longevity by 40% by reducing both rDNA recombination and the accumulation of extrachromosomal DNA circles (ERCs) [MMG99]. Conversely, the deletion of SIR2 dramatically decreases replicative life span [MMG99]. Fabrizio et al. indicated that the effects of Sir2 on chronological life span are opposite to replicative life span. They suggested that the lack of Sir2 along with calorie restriction and/or mutations in the yeast AKT homolog, Sch9, or Ras pathways causes a dramatic chronological life-span extension [FGB$^+$05]. In C. elegans, dosage of the SIR2 ortholog, sir-2.1, increases the mean life span by up to 50% [HATG01]. In flies, the Sir2 ortholog, dSir2, has been reported to extend life span as well [RH04]. In addition, life-span extension by CR is blocked in strains lacking dSir2. These findings suggest that CR works through a Sir2-dependent mechanism in this organism.

RAS/cAMP/PKA pathway

The second group of genes that are relevant to CR includes TOR, PKA, and SCH9. It has been shown that at either 0.5% or 0.05% glucose, CR extends life span of yeast in a

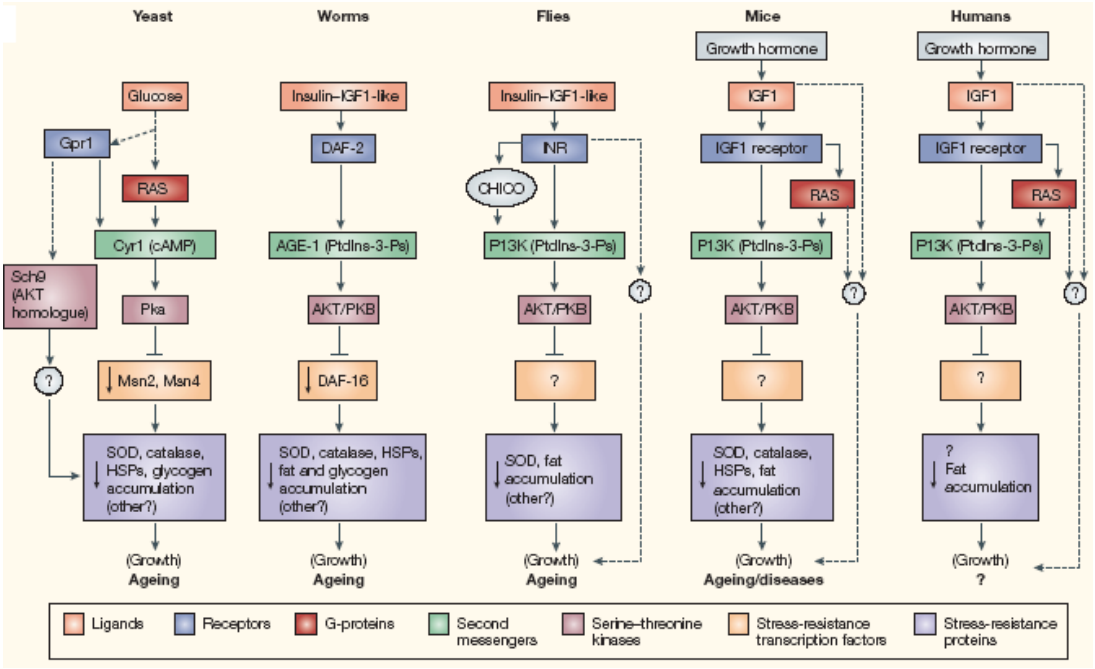

Figure 4.1: Longevity regulatory pathway in five organisms. *The figure is copied from Longo et al.(NATURE REVIEWS GENETICS, Vol. 6, 866-872).*

manner dependent on the nutrient-responsive kinases, TOR, PKA, and SCH9 [FPP+01, KPS+05]. Mutations that result in decreased activity of PKA, Sch9, or TOR increase both replicative and chronological life span. Moreover, the long replicative life span of TOR and Sch9 deletion strains is not further increased by calorie restriction [KPS+05], suggesting that TOR and Sch9 might mediate replicative life-span extension by calorie restriction [KSK05]. Of more importance, mutations that decrease the activity of the orthologous proteins (Tor and Akt) in worms and flies also extend life span, suggesting that these kinases share an evolutionarily conserved role in responding to nutrients and growth factors.

In yeast and higher eukaryotes, TOR, Sch9, and PKA coordinate signals from nutrients and growth factors to regulate ribosome biogenesis, stress response, cell size, autophagy, and other cellular processes. This represents a common longevity pathway

that is conserved from yeast to mouse as shown in Figure. 4.1. As have been mentioned, down-regulation of the pathway results in longevity in organism from yeast to mouse. For example, mutations that decrease the level or activity of IGF1 in mouse extend lifespan by up to 65% compared with that of the wild type. Conservation of the longevity regulatory pathways suggests that at least part of these pathways evolved from a common set of starvation-response genes in the ancestral organisms.

Stress response genes

The third group of genes includes some genes downstream of the longevity regulatory pathway, such as Msn2, Msn4, and SOD genes. The effect of these stress responses genes have been studied in different organisms. In yeast, over-expression of Sod1 and Sod2 increases the life span by about 30%. In C.elegans, life span of wild type worms can be extended if they're treated with small synthetic SOD/catalase mimetics [MRM$^+$00]. In M.drosophila, over-expression of Sod1 increases survival rate by up to 40% [PED$^+$98]. Although stress response genes may play important roles and sometimes are essential for longevity (i.g. Sod2 is required for life span extension in Sch9$\triangle$) [FLM$^+$03], they are likely to act as the effectors of the above mentioned longevity pathway.

In our ageing project, we focus our attention on four genes: Sch9, Tor1, Ras2, and Sir2. Ras2 encodes a GTP-binding protein, which stimulates the production of cAMP by adenylate cyclase and therefore is a positive regulator of the PKA activity. We produced four long-lived yeast knock-out strains: sch9$\triangle$, tor1$\triangle$, ras2$\triangle$, and sch9sir2$\triangle$. Then we measured the gene expression profiles in these strains. By integrating other public available data sources, we performed computational analysis on the microarray data. We hope our work can shed light on the underlying mechanisms of longevity in these strains.

4.2 Materials and methods

4.2.1 DNA Microarray hybridization and data processing

Yeast were grown in SDC medium containing 2% glucose and supplemented with amino acids, adenine, and uracil for two and half days. Then cells were collected from wild type, $sch9\triangle$, $ras2\triangle$, $tor1\triangle$ and $sch9sir2\triangle$ strains. They were used to extract total RNA according to the acid phenol method.

Total RNA from independent cultures of each strain was used as a template to synthesize complementary RNA (cRNA) and the cRNA was hybridized to Affymetrix GeneChip Yeast2.0 Array. For each of the five strains, three replicate arrays were generated, of which each corresponds to RNA obtained from independent population. All the replicates for the same strain are highly consistent with Pearson correlation coefficients greater than 0.96. The probe-level data was normalized using "Invariant Set" method. The expression levels of all probe sets were calculated using "Model-Based analysis of Oligonucleotide Arrays"[LW01a] with "PMonly" PM correction. Bioconductor affy package software was used for the analysis(http://www.bioconductor.org/). Note that we did not use the Sub-Sub method to do normalization for the data, because the chips used in this study, Affymetrix Yeast2.0 Array, contains probes from two yeast species: *S.cerevisiae* and *S.pombe*, which make it inappropriate to apply our normalization method.

The Yeast2.0 Array contains probe sets for both *S.cerevisiae* and *S.pombe*. Only probe sets from *S.cerevisiae* are used in later analysis. Gene expression change were calculated between two strains using pairwise comparison. So 3×3 comparisons result in 9 ratios, which are averaged to get the mean fold change (FC). Fold changes of all the *S.cerevisiae* probe sets were calculated for comparisons: $sch9\triangle/wt$, $ras2\triangle/wt$, $tor1\triangle/wt$, $sch9sir2\triangle/wt$, and $sch9sir2\triangle/sch9\triangle$. Most genes correspond to only

one probe set in Affymetrix GeneChip Yeast2.0 Array and the average fold changes were used for genes with multiple probe sets.

4.2.2 Gene Ontology analysis

GO information was downloaded from "ftp://genome-ftp.stanford.edu/pub/go/ontology/" based on data at July 29, 2005. Yeast gene annotation data was downloaded from "ftp://genome-ftp.stanford.edu/pub/go/gene_ association/". The data structure for gene ontology (GO) is directed acyclic graph (DAG). Each node in the DAG is a set of genes with given annotations. Nodes that are closer to the terminal have more detailed annotation and thereby are more informative. To avoid redundancy and overlapping between GO nodes, we identified 44 cellular components, 53 molecular functions and 109 biological processes informative nodes from the GO DAG. Terminal informative nodes are defined as those nodes that are closest to the terminal and have at least 30 genes. The GO categories that associate with terminal informative node is defined as terminal informative GO categories (TIGO).

In general, to test whether a priorly defined set of genes S is significantly affected in a mutation strain(e.g., $sch9\triangle$), we applied a similar method as the Gene Set Enrichment Analysis[STM$^+$05]. We rank the log transformed fold changes of all genes in $sch9\triangle/wt$, which results in a ranked list G. If S is not significantly affected, we would expect that the members of S are randomly distributed throughout G. Otherwise we claim that S is significantly affected. If most members of S are found at the top of list G, we define it as positively affected gene set. Conversely, if most members of S are found at the bottom of list G, we define it as negatively affected gene set. In practice, we simply compare the fold changes of genes in S with those in $G - S$ using Wilcoxin rank test. Here the gene set S can be a GO category, genes related to pathway, genes

bound by a transcription factor, or genes localized in the same organelle. To guarantee a reliable result, we only apply this test to gene sets with at least 10 members.

Based on the method described above, we calculated p-values for all the defined TIGO categories. Then we performed multiple testing correction using method introduced by Storey et al. [ST03]. The q-values were computed using ”qvalue” package provided in R software (http://www.r-project.org/). We compared our results with those obtained by running the web program: GOstat (http://gostat.wehi.edu.au/) [BS04]. They are in good consistency. By taking only the terminal informative GO categories rather than all the GO nodes, our method avoids the redundancy problem and the results are easier to be interpreted.

4.2.3 Pathway analysis

To understand the mechanisms of ageing, it is helpful to find out which pathways are changed in the long-lived mutants, which motivates us to identify the significantly affected pathways. We downloaded the pathway data set from KEGG database: http://www.genome.jp/kegg/. The data set includes 102 *S.cerevisiae* pathways in total. To identify significantly affected pathways in each strain, p-values and q-values were calculated for each pathway using methods described above. Here, all the genes belong to a pathway forms a gene set.

4.2.4 Cellular organelle analysis

Here we regard genes with the same cellular localization as a gene set and performed the analysis described above. The cellular localization data was downloaded from http://yeastgfp.ucsf.edu/. In this data set, ,75% proteins were classified into 22 distinct subcellular localization categories, including mitochondria, nucleus, nucleolus, vacuole,

vacuole membrane, budding neck, etc. It is known that some organelles, such as mitochondria, play a central role in ageing of yeast. We hope that the cellular organelle analysis would provide some information about yeast ageing in the sub-cellular (organelle) level.

4.2.5 ChIP-Chip based transcription factor analysis

It is often difficult to determine whether the activity of a transcription factor is changed or not according to its expression level in a microarray data. Because, first the activity of a transcription factor is often regulated in protein level, i.g. by phosphorylation or translocation, rather than in mRNA level; Secondly, transcription factors are often expressed in a low level, which makes it difficult to detect the expression changes of them due to the high noise in microarray data. As such, we have to apply an indirect strategy to identify the affected transcription factors by investigating the target genes that are regulated by the transcription factor. In yeast, large scale studies have been performed to identify the interactions between transcription factors and genes by using ChIP-Chip experiment. Here we use this valuable data source to infer significantly affected transcription factors in the four long-lived mutants. We downloaded the Chip-Chip data set from http://web.wi.mit.edu/young/regulatory_code/. It contains gene binding information for 203 transcription factor (TF), where each TF-gene association was assigned a p-value. We set the threshold to be 0.001, which corresponds to a false positive of about 4% and a false negative of about 25%[HGL$^+$04]. Again the target genes for each TF were regarded as a gene set and the significantly affected TFs are identified using the methods described above. Our method is relatively robust to the noise in Chip-Chip data, because when we lower the threshold for TF-gene association down to 0.01, we obtain very similar results.

4.2.6 Motif enrichment analysis

The previously described transcription factor analysis has a major limitation, because the target genes regulated by a TF is determined based on ChIP-Chip experiment. The cells for ChIP-Chip experiment were cultured in YPD medium at log phase, whereas the cells in our microarray experiment were grown in SDC medium and collected at day 2.5. As we know, the gene set regulated by a TF could be different in various conditions. To overcome this problem, we perform motif enrichment analysis which takes advantage of the sequence information in the regulatory region of all genes. Basically, we analyze the enrichment of a motif in the up- or down-regulated genes. Those enriched motifs are likely to be the regulatory binding sites of TFs that cause the up- or down-regulation.

We use AlignACE with 12 bp motifs, and search up to 800 bp upstream of each gene in *S. cerevisiae*. After removing the redundancy, 666 motifs were obtained, including 51 motifs with known binding transcription factors. For each gene, the upstream motifs, motif orientations and scores were recorded. Refer to Beer et al. [BT04] for details and the motif data is available at http://genomics.princeton.edu/tavazoie/Supplementary%20Data.htm.

To identify the enriched motifs in a given gene set of size K, we used hypergeometric test. Suppose there are totally M genes with a given motif, and the rest N genes don't have this motif (in total there are M+N genes). Let X be the number of genes in the gene set that contain the motif, then $X \sim hyper(M, N, x)$. That is:

$$p(M, N, x) = \frac{\binom{M}{x}\binom{N}{K-x}}{\binom{M+N}{M}}$$

The P value for an observed x is $Pr(X \geq x|M, N)$, namely, the probability of observing at least x genes with the interested motif by chance. We calculated the p-values of all the 666 motifs in the up-regulated gene set and down-regulated gene set. Then multiple testing correction was performed and q-value was computed for each motif using the "qvalue" package provided in R software. An arbitrary threshold of 2 (fold change) was used to determine up- or down-regulation.

4.3 Results

4.3.1 Similarity of gene expression profiles in the long-lived mutants

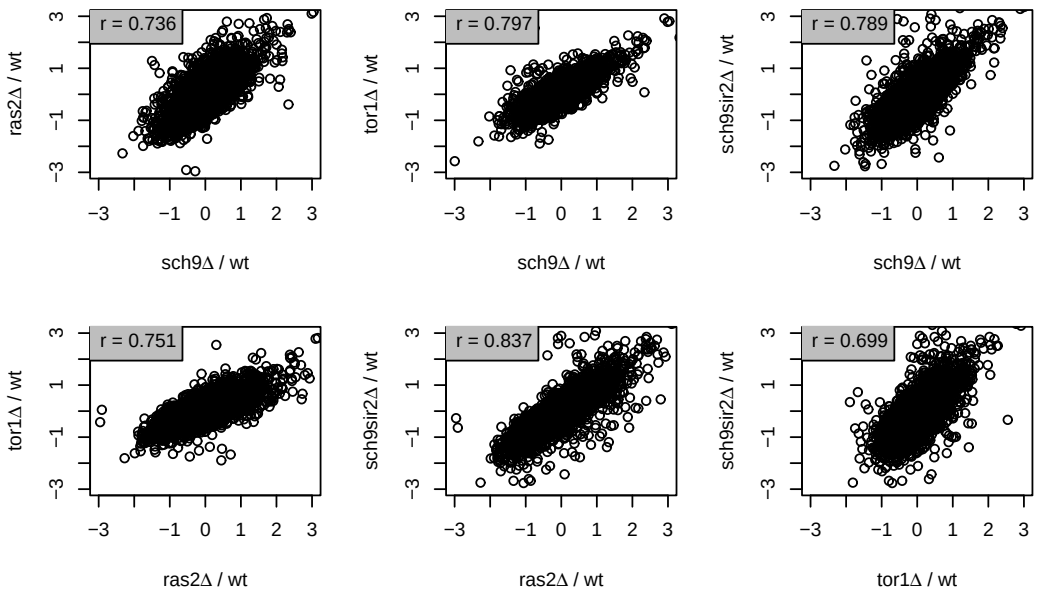

Figure 4.2: Similarity of gene expression profiles in the four long-lived mutants: $Sch9\triangle$, $Ras2\triangle$, $Tor1\triangle$ and $Sch9Sir2\triangle$.

We collected the RNA samples at day 2.5 from *S.cerevisiae* wild type and four long-lived mutant strains: $sch9\triangle$, $ras2\triangle$, $tor1\triangle$ and $sch9sir2\triangle$. The expression levels for 5841 genes were measured using Affymetrix GeneChip Yeast2.0 arrays. Then we

calculate the log transformed fold changes (log ratio) for all genes in the four mutant strains versus in wild type, respectively. These calculations result in an expression profile for each mutant, which reflects the gene expression change in the corresponding mutant with respect to the wild type strain. As shown in Figure 4.2, the expression profiles of the four long-lived mutants are highly similar to each other. This may imply that deletion of $sch9$, $ras2$, $tor1$ or $sch9sir2$ influence a common set of downstream genes. Previous study has proposed that protein Sch9, Tor1 and Ras2 control a molecular switch that regulates the response to nutrient availability [REM$^+$05]. Therefore, the similarity of expression profiles of them is more or less expected.

4.3.2 Differentially expressed genes in the long-lived mutants

By arbitrarily setting 2 as the threshold for up- and down-regulated genes, we found 147, 324, 132, 304 up-regulated genes (log ratio $\geq$ 2) and 130, 364, 60, 425 down-regulated genes (log ratio $\leq$ 2) in $sch9\triangle$, $ras2\triangle$, $tor1\triangle$ and $sch9sir2\triangle$ strain with respect to wild type, respectively. Among these genes, 65 up-regulated genes and 24 down-regulated genes are shared by all these four mutants (see Table 4.1 and Table 4.2).

It was suggested in previous studies that Ras2, Tor1 and Sch9 are all involved in low nutrient response and adaptation, in which the PKA kinase plays an important role [REM$^+$05, ZMC05, PDC$^+$03, Lon03, CCL$^+$99]. Consistently, we find many of the common up-regulated genes in the long-lived mutants are functionally related to this pathway. In all the four long-lived mutants, expressions of Hxt2 and Hxt4 are up-regulated. It is known that these two genes encode high-affinity glucose transporters of the major facilitator superfamily whose expression is induced by low levels of glucose and repressed by high levels of glucose [OJ95, OJ99]. The increase of their expression may facilitate the transport of glucose into yeast cells from medium. Another gene,

Table 4.1: The common up-regulated genes in the four mutants.

Name	Function	sch9Δ/wt	ras2Δ/wt	tor1Δ/wt	sch9sir2Δ/wt
ANT1	Peroxisomal adenine nucleotide transporter	2.13	2.73	2.01	2.30
ARN2	Transporter	2.14	2.37	2.51	2.76
ATG19	Involved in the cytoplasm-to-vacuole targeting pathway	4.51	2.68	3.96	3.41
BIO4	Dethiobiotin synthetase	2.25	2.25	2.04	2.32
ECM4	Non-essential protein of unknown function, similar to Ygr154cp	2.48	2.25	2.33	2.15
ERG25	C-4 methyl sterol oxidase,	2.50	2.12	2.54	2.27
ERG5	C-22 sterol desaturase	5.09	5.44	4.27	5.18
FMP45	Required for viability in stationary phase	4.92	5.20	4.50	6.81
GND2	6-phosphogluconate dehydrogenase	2.49	3.11	2.25	3.00
GPD1	NAD-dependent glycerol-3-phosphate dehydrogenase	2.49	2.98	2.48	2.83
GPM2	Homolog of Gpm1p phosphoglycerate mutase	2.81	2.52	4.15	2.63
GRE1	Stress induced , regulated by the HOG pathway	5.27	5.22	4.16	5.84
HBT1	Substrate of the Hub1p ubiquitin-like protein	5.08	5.69	3.47	6.19
HMS1	C2H2 zinc-finger protein	4.03	2.49	2.84	5.71
HXT2	High-affinity glucose transporter of the major facilitator superfamily	4.44	2.97	4.70	5.68
HXT4	High-affinity glucose transporter of the major facilitator superfamily	2.72	4.34	2.18	3.55
IML1	Unknown	2.27	2.19	2.23	2.63
JLP1	Fe(II)-dependent sulfonatealpha-ketoglutarate dioxygenase	2.57	25.95	2.78	3.00
MCH2	Protein with similarity to mammalian monocarboxylate permeases	3.49	5.18	2.86	5.17
MFA2	Mating pheromone a-factor	2.23	2.58	2.64	2.17
MIG2	Involved in repression of SUC2 by high levels of glucose	4.96	6.46	3.41	6.71
PDC6	Minor isoform of pyruvate decarboxylase	7.45	10.63	7.56	9.73
PHM6	Unknown	2.49	2.23	2.45	5.00
PIR1	O-glycosylated protein required for cell wall stability	3.09	2.61	2.33	3.54
PMA2	Plasma membrane H+-ATPase, isoform of Pma1p	7.94	8.58	6.88	10.22
POT1	3-ketoacyl-CoA thiolase	3.25	5.80	2.47	5.43
PPH22	Catalytic subunit of protein phosphatase 2A	2.16	2.30	2.02	2.22
RPI1	Putative transcriptional regulator	3.28	4.49	2.65	4.45
RPL17A /B	Protein component of the large (60S) ribosomal subuni	2.57	2.03	2.81	2.29
RPL18A/B	Protein component of the large (60S) ribosomal subunit	2.60	2.55	2.71	2.53
RPL22A	Protein component of the large (60S) ribosomal subunit	2.67	2.21	2.87	2.74
RPL22B	Protein component of the large (60S) ribosomal subunit	2.59	2.46	2.25	2.59
RPL26B	Protein component of the large (60S) ribosomal subunit	2.24	2.21	2.30	2.21
RPS0A	Protein component of the small (40S) ribosomal subunit	2.49	2.13	2.33	2.44
RTN2	Unknown	3.78	4.19	3.04	4.79
SPS100	Required for spore wall maturation, expressed during sporulation	10.03	14.17	4.53	16.79
STB2	Part of a large protein complex with Sin3p and Stb1p	2.13	2.13	2.20	2.53
THI4	Thiamine biosynthesis and mitochondrial genome stability	3.69	6.17	4.80	5.71
TIS11	mRNA-binding protein expressed during iron starvation	3.46	2.77	2.56	3.46
TKL2	Transketolase, similar to Tkl1p	2.55	3.37	2.42	3.60
TPK1	Subunit of cytoplasmic cAMP-dependent protein kinase	2.47	2.47	2.37	2.72
XYL2	Xylitol dehydrogenase	3.27	3.48	2.35	6.47
YAL037C-A	Unknown	3.32	7.63	2.39	4.27
YBL039W-A	Unknown	2.72	2.97	2.91	2.95
YBR047W	Mitochondria protein	4.48	2.42	2.93	2.33
YBR071W	Unknown	4.45	3.28	3.31	3.40
YDL057W	Unknown	3.16	3.49	2.82	3.50
YDL218W	Unknown	26.63	42.20	11.11	53.01
YDR034W-B	Unknown	2.91	3.77	3.06	3.11
YDR222W	Unknown	3.95	5.42	2.23	4.84
YGR026W	Unknown	3.17	4.05	3.08	4.00
YGR154C	Omega-class glutathione transferase	2.19	3.96	2.14	2.60
YHL035C	Unknown	2.65	2.01	2.08	2.29
YHR140W	Unknown	3.14	7.34	2.75	4.77
YJR024C	Unknown	2.04	2.44	2.18	2.56

Table 4.1: Continued

Name	Function	sch9Δ/wt	ras2Δ/wt	tor1Δ/wt	sch9sir2Δ/wt
YKL050C	Unknown	3.44	4.54	2.29	5.29
YKL107W	Unknown	3.22	3.40	2.89	5.14
YLL056C	Unknown	2.57	5.39	2.36	3.09
YLR099W-A	Unknown	2.45	2.19	2.33	2.29
YLR164W	Homologous to TIM18p	4.31	5.27	4.09	7.23
YLR281C	Unknown	2.81	3.91	2.59	3.72
YLR346C	Unknown	2.61	2.26	2.74	2.15
YMR175W-A	Unknown	8.22	8.97	7.01	10.42
YMR251W	Omega class glutathione transferase	3.23	2.88	2.29	4.17
YOR186W	Unknown	2.90	5.68	3.64	5.28

Rpi1, is also up-regulated in all the mutants. The protein RPI1 is an inhibitor of the Ras-cAMP pathway, whose over-expression suppresses the heat shock sensitivity of Ras2 over-expression in wild type. Thus, up-regulation of Rpi1 may enhance the stress resistance in these mutants which may contribute to the life span extension.

Interestingly, we found that the expression level of Tpk1, which encodes one of the subunits of cAMP dependent kinase PKA, is decreased in all the mutants. This is out of our expectation, because RAS2 positively regulates the PKA kinase activity and we may expect a down-regulation of the genes that encode subunit of protein kinase PKA. The contradiction may reflect the gap between gene expressions and protein activities. Deletion of Ras2 gene causes repression of the PKA pathway by regulating the protein activities. Whereas in the gene expression level, there may be a negative feedback which increases the expression of genes encoding subunits of PKA kinase. On one hand, the regulation in protein activities is more sensitive than that in gene expression levels. On the other hand, negative feedback is often used to ensure that a pathway can be shut down when the signals are removed. The regulation the PKA pathway by Ras2 may be more complicated and elaborated than what we have thought. Additionally, the high similarities between the expression profiles in the four mutants are interesting. Ras2, Tor1 and Sch9 may regulate the nutrient responses through a common or at least a related mechanism.

Table 4.2: The common down-regulated genes in the four mutants.

Name	Function	sch9△/wt	ras2△/wt	tor1△/wt	sch9sir2△/wt
ADK2	Mitochondrial adenylate kinase	-2.32	-3.56	-2.56	-2.84
AEP2	Mitochondrial protein	-2.01	-2.58	-2.24	-2.83
CWP1	Cell wall mannoprotein	-3.68	-2.65	-3.00	-2.26
GIN4	Protein kinase involved in bud growth and assembly of the septin ring	-3.23	-2.67	-2.57	-3.93
IMG2	Mitochondrial ribosomal protein of the small subunit	-2.42	-3.28	-2.02	-3.42
IXR1	Binds DNA containing intrastrand cross-links formed by cisplatin	-2.07	-3.13	-2.07	-2.27
KTR5	Putative mannosyltransferase involved in protein glycosylation	-3.36	-3.96	-3.07	-3.53
MRP13	Mitochondrial ribosomal protein of the small subunit	-2.11	-2.69	-2.11	-2.84
MRPL17	Mitochondrial ribosomal protein of the large subunit	-2.08	-2.18	-2.01	-2.45
MRPL35	Mitochondrial ribosomal protein of the large subunit	-2.35	-3.50	-2.37	-2.64
MRPL7	Mitochondrial ribosomal protein of the large subunit	-2.22	-2.88	-2.13	-2.53
MRPL9	Mitochondrial ribosomal protein of the large subunit	-2.22	-2.92	-2.00	-2.39
MRPS8	Mitochondrial ribosomal protein of the small subunit	-2.99	-3.27	-2.08	-3.95
OMS1	Protein integral to the mitochondrial membrane	-2.80	-3.37	-2.23	-3.19
RSM19	Mitochondrial ribosomal protein of the small subunit	-2.06	-3.12	-2.16	-2.56
TAH1	HSP90 cofactor	-5.03	-4.83	-3.51	-6.73
TIM9	Mitochondrial intermembrane space protein	-2.52	-3.42	-2.07	-3.87
YBL005W-A	TyB Gag-Pol protein	-2.47	-2.69	-2.80	-4.59
YDR444W	Unknown	-2.39	-2.45	-2.03	-2.48
YKL105C	Unknown	-2.39	-3.53	-2.91	-2.75
YKR016W	The authentic, localized to the mitochondria	-3.00	-3.46	-2.23	-3.59
YLR012C	Unknown	-7.98	-11.33	-5.94	-13.06
YMR010W	Unknown	-2.07	-2.25	-2.45	-2.24
YNL295W	Unknown	-2.04	-3.33	-2.32	-2.72

In the common down-regulated genes (see Table 4.2), 13 out of these 24 genes encode mitochondria proteins that include 8 mitochondria ribosomal proteins (RP). It is known that mitochondria plays an important role in ageing. Down-regulation of mitochondria genes implies that the dependence of cells on mitochondria is reduced in the long-lived mutants. In contrast to the down-regulation of mitochondria RP genes, the cytosolic RP genes tend to be up-regulated. In the 65 common up-regulated genes, 6 are cytosolic RP genes. It has been suggested that TOR1 regulates RP gene expression via PKA pathway and inhibition of TOR1 protein by rapamycin causes repression of RP gene expression [MSH04]. It is interesting to see that expressions of cytosolic RP genes are up-regulated in $Tor1\triangle$ mutant. It is possible that TOR1 may not be essential for expression of RP genes and the up-regulation of RP genes in $Tor1\triangle$ is caused by proteins that have redundant functions with TOR1. Alternatively, it could be simply a consequence of the delay of ageing in the long-lived mutants. The expression of RP

genes decreases after cells enter the stationary phase during ageing. Since the ageing is delayed in the long-lived mutants, the cells from the mutants are "younger" than those from the wild type collected at the same day. In comparison with the "elder" wild type cells, the expression of RP genes in "younger" mutants are less reduced and therefore appear to be up-regulated.

The overlapping between the differentially expressed genes in the four long-lived mutants is shown as venn diagrams in Figure 4.3. As can be seen, both the up-regulated and down-regulated genes in these mutants are highly overlapped.

4.3.3 Significantly affected GO categories in the long-lived mutants

To understand the mechanisms of longevity, we would like to know which function categories are changed in the long-lived mutants. The Gene Ontology (GO) project has developed three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes (P), cellular components (C) and molecular functions (F) in a species-independent manner [Ont]. A cellular component is just that, a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer). A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. Molecular function describes activities, such as catalytic or binding activities, at the molecular level. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

The GO uses a directed acyclic graph (DAG) to represent the hierarchical structure of gene function categorization. In the DAG, each node includes a set of gene with the same function terms. The terms that associated with a node closer to the terminals

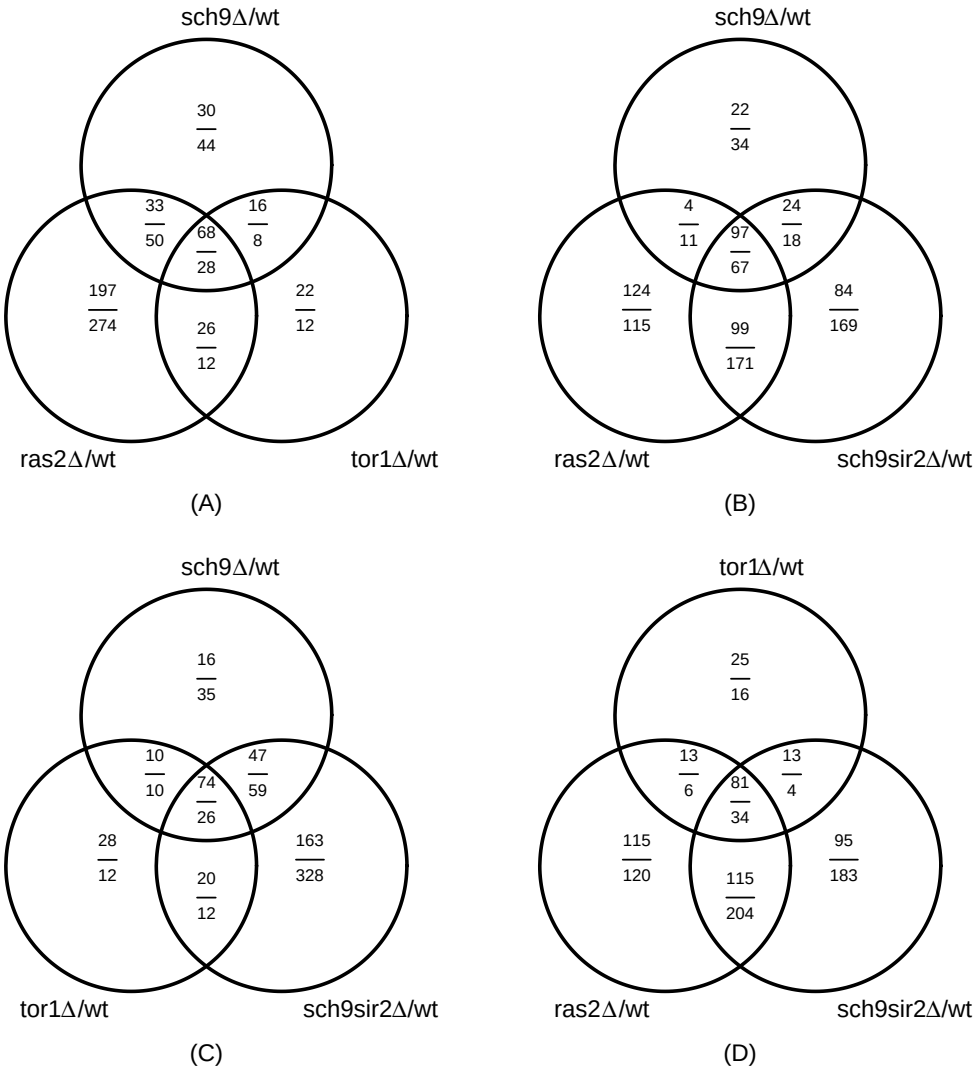

Figure 4.3: Overlap of up-regulated (numbers over the line) and down-regulated (numbers below the line) genes in the four long-lived mutants: $Sch9\triangle$, $Ras2\triangle$, $Tor1\triangle$ and $Sch9Sir2\triangle$.

are more informative. To avoid redundancy and low informativeness, we selected only those GO categories that are closest to the terminal in the DAG and contain at least 30 genes. We denote these GO categories as terminal informative GO categories (TIGO). Totally, we selected 44 cellular component TIGOs, 53 molecular function TIGOs and 109 biological process TIGOs from *S.cerevisiae*. From them, we identified some TIGOs

that are positively affected (see Table 4.3.3) or negatively affected (see Table 4.3.3) in at least one of the long-lived mutants. Totally there are 7 cellular component TIGOs, 4 molecular function TIGOs, 8 biological processing TIGOs that are positively affected and 19 cellular component TIGOs, 12 molecular function TIGOs, 28 biological processing TIGOs that are negatively affected in at least one the comparisons: $sch9\triangle/wt$, $ras2\triangle/wt$, $tor1\triangle/wt$, and $sch9sir2\triangle/sch9\triangle$.

Table 4.3: Positively affected TIGO categories in the four mutants.

Positively affected TIGO categories				sch9Δ/wt3		rasΔ/wt3		torΔ/wt3		sch9sir2Δ/wt3		sch9sir2Δ/sch9Δ	
Category	GO ID	Gene Num	Annotation	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
C	GO:0005842	93	cytosolic large ribosomal subunit	**0**	**0**	**1.64E-12**	**2.37E-10**	**0**	**0**	**1.58E-12**	**2.37E-10**	**1.00**	**0.87**
C	GO:0005843	63	cytosolic small ribosomal subunit	**0**	**0**	**7.49E-09**	**6.49E-07**	**0**	**0**	**1.38E-09**	**1.50E-07**	**0.99**	**0.87**
C	GO:0009277	111	cell wall	0.89	0.87	**5.65E-06**	**0.00029**	**6.19E-06**	**0.00030**	0.0072	0.074	**6.90E-07**	**4.59E-05**
C	GO:0005811	33	lipid particle	0.052	0.26	0.052	0.26	0.069	0.30	0.0029	0.038	**0.00042**	**0.0091**
C	GO:0005789	115	endoplasmic reticulum membrane	0.024	0.16	**6.46E-07**	**4.59E-05**	0.0055	0.061	0.0096	0.088	0.014	0.11
C	GO:0031226	31	intrinsic to plasma membrane	0.0086	0.082	**7.05E-05**	**0.0023**	0.029	0.18	0.0066	0.071	0.057	0.27
C	GO:0005794	130	Golgi apparatus	**0.00043**	**0.0092**	0.33	0.87	0.030	0.18	0.30	0.85	1.00	0.87
F	GO:0015082	35	di-, tri-valent inorganic cation transporter activity	0.023	0.15	**0.00012**	**0.0035**	0.013	0.10	**0.00023**	**0.0055**	0.0020	0.028
F	GO:0046873	38	metal ion transporter activity	0.014	0.11	**0.00014**	**0.0038**	0.023	0.15	**0.00016**	**0.0043**	0.0026	0.035
F	GO:0000030	42	mannosyltransferase activity	0.011	0.094	**9.70E-05**	**0.0029**	0.062	0.28	**0.00029**	**0.0067**	0.0060	0.065
F	GO:0003735	229	structural constituent of ribosome	**1.82E-05**	**0.00069**	0.54	0.87	**1.41E-05**	**0.00056**	0.26	0.79	1.00	0.87
P	GO:0008643	32	carbohydrate transport	0.14	0.52	**1.89E-06**	**0.00011**	0.0047	0.054	**1.45E-06**	**9.00E-05**	**2.56E-06**	**0.00014**
P	GO:0046474	34	glycerophospholipid biosynthesis	0.99	0.87	0.042	0.22	0.86	0.87	0.58	0.87	**2.24E-05**	**0.00081**
P	GO:0030384	31	phosphoinositide metabolism	0.56	0.87	**0.00046**	**0.0095**	0.58	0.87	0.036	0.20	**0.00013**	**0.0036**
P	GO:0016125	37	sterol metabolism	0.0057	0.062	**7.51E-05**	**0.0023**	0.0075	0.076	**0.00042**	**0.0091**	0.0053	0.060
P	GO:0006487	43	protein amino acid N-linked glycosylation	0.0069	0.072	**4.99E-05**	**0.0017**	0.031	0.19	0.0013	0.020	0.0084	0.081
P	GO:0046365	33	monosaccharide catabolism	0.0013	0.020	**8.81E-06**	**0.00038**	**2.94E-05**	**0.0010**	**0.00017**	**0.0043**	0.012	0.10
P	GO:0042364	36	water-soluble vitamin biosynthesis	0.00087	0.015	**0.00023**	**0.0055**	0.013	0.10	0.00064	0.012	0.023	0.15
P	GO:0007047	140	cell wall organization and biogenesis	**0.00024**	**0.0055**	0.0047	0.054	0.035	0.20	**0.00048**	**0.0096**	0.12	0.47

Table 4.4: Negatively affected TIGO categories in the four mutants.

Negatively affected TIGO categories

Category	GO ID	Gene Num	Annotation	sch9Δ/wt3		rasΔ/wt3		torΔ/wt3		sch9sir2Δ/wt3		sch9sir2Δ/sch9Δ	
				p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
C	GO:0005762	43	mitochondrial large ribosomal subunit	**1.56E-19**	**3.32E-17**	**1.34E-20**	**4.29E-18**	**1.13E-20**	**4.29E-18**	**8.20E-17**	**1.05E-14**	**6.78E-05**	**0.0007**
C	GO:0005763	34	mitochondrial small ribosomal subunit	**6.94E-13**	**4.93E-11**	**4.83E-14**	**4.41E-12**	**3.17E-13**	**2.54E-11**	**2.06E-12**	**1.25E-10**	**5.06E-05**	**0.00056**
C	GO:0016591	74	DNA-directed RNA polymerase II\, holoenzyme	**1.61E-05**	**0.00023**	**4.97E-10**	**2.27E-08**	**9.05E-05**	**0.00086**	**2.24E-10**	**1.10E-08**	**4.20E-11**	**2.24E-09**
C	GO:0005743	158	mitochondrial inner membrane	**2.64E-16**	**2.82E-14**	**3.14E-09**	**1.34E-07**	**3.56E-17**	**5.70E-15**	**6.31E-08**	**2.02E-06**	0.89	0.64
C	GO:0000502	46	proteasome complex	**0.00039**	**0.0026**	**1.35E-08**	**4.79E-07**	0.0045	0.017	**5.46E-06**	**9.33E-05**	**4.40E-05**	**0.00051**
C	GO:0016585	71	chromatin remodeling complex	0.066	0.12	**9.96E-08**	**3.04E-06**	0.0034	0.014	**0.00023**	**0.0017**	**2.65E-05**	**0.00035**
C	GO:0000123	39	histone acetyltransferase complex	0.019	0.048	**1.52E-05**	**0.00022**	0.0034	0.014	**7.85E-05**	**0.00078**	**9.53E-05**	**0.00089**
C	GO:0005684	37	major (U2-dependent) spliceosome	0.023	0.054	**7.28E-05**	**0.00074**	0.15	0.21	**0.00010**	**0.00093**	**6.80E-05**	**0.00070**
C	GO:0000778	47	condensed nuclear chromosome kinetochore	0.022	0.054	**0.00047**	**0.0029**	0.080	0.13	**2.74E-05**	**0.00035**	**7.40E-07**	**2.06E-05**
C	GO:0005643	50	nuclear pore	0.87	0.14	**0.0014**	**0.0073**	0.0086	0.028	0.026	0.061	0.046	0.088
C	GO:0000790	36	nuclear chromatin	0.39	0.43	**0.0018**	**0.0085**	0.16	0.23	0.045	0.088	0.013	0.038
C	GO:0005934	50	bud tip	0.44	0.47	**0.0021**	**0.0099**	0.19	0.26	0.037	0.078	0.0031	0.013
C	GO:0000131	35	incipient bud site	0.63	0.61	0.0022	0.010	0.044	0.087	0.074	0.13	**0.0019**	**0.0091**
C	GO:0005875	35	microtubule associated complex	0.043	0.086	0.0025	0.011	0.37	0.42	**0.00021**	**0.0016**	**5.54E-06**	**9.33E-05**
C	GO:0046540	30	U4/U6 x U5 tri-snRNP complex	0.18	0.25	0.014	0.041	0.64	0.61	0.0024	0.011	**9.57E-05**	**0.00089**
C	GO:0030880	32	RNA polymerase complex	0.0068	0.023	0.057	0.11	0.016	0.044	**0.00032**	**0.0022**	**0.00042**	**0.0026**
C	GO:0005768	70	endosome	0.99	0.64	0.098	0.16	1.00	0.64	0.35	0.41	**0.0011**	**0.0060**
C	GO:0005794	130	Golgi apparatus	1.00	0.64	0.67	0.63	0.97	0.64	0.70	0.64	**0.00087**	**0.0049**
C	GO:0005842	93	cytosolic large ribosomal subunit	1.00	0.64	1.00	0.64	1.00	0.64	1.00	0.64	**0.00048**	**0.0029**
F	GO:0016251	62	general RNA polymerase II transcription factor activity	0.019	0.048	**1.38E-06**	**3.04E-05**	0.022	0.053	**4.51E-05**	**0.00052**	**2.85E-06**	**5.63E-05**
F	GO:0003924	59	GTPase activity	0.037	0.078	**3.66E-05**	**0.00046**	0.065	0.11	**2.90E-06**	**5.63E-05**	**8.05E-07**	**2.15E-05**
F	GO:0008080	37	N-acetyltransferase activity	0.0069	0.023	**0.00032**	**0.0022**	0.0064	0.022	**0.00013**	**0.0010**	**0.00066**	**0.0039**
F	GO:0004540	100	ribonuclease activity	0.057	0.11	**0.00045**	**0.0028**	0.0030	0.013	0.022	0.053	0.053	0.099
F	GO:0005478	30	intracellular transporter activity	0.11	0.17	**0.0012**	**0.0064**	0.25	0.32	**0.00040**	**0.0026**	**7.17E-06**	**0.00011**
F	GO:0004527	33	exonuclease activity	0.10	0.16	**0.0012**	**0.0064**	**0.0018**	**0.0089**	0.036	0.076	0.056	0.10
F	GO:0004519	56	endonuclease activity	0.053	0.098	**0.0017**	**0.0084**	0.049	0.093	0.018	0.047	0.041	0.084
F	GO:0008565	50	protein transporter activity	**0.00027**	**0.0019**	0.015	0.041	**7.99E-05**	**0.00079**	0.0055	0.020	0.27	0.34
F	GO:0005083	58	small GTPase regulator activity	0.98	0.64	0.019	0.048	0.29	0.36	0.23	0.30	**0.00036**	**0.0024**
F	GO:0008026	37	ATP-dependent helicase activity	0.067	0.12	0.032	0.071	**0.0014**	**0.0073**	0.044	0.087	0.093	0.15
F	GO:0003724	41	RNA helicase activity	0.030	0.067	0.041	0.083	**0.00044**	**0.0028**	0.018	0.047	0.086	0.14
F	GO:0003899	33	DNA-directed RNA polymerase activity	0.0045	0.017	0.043	0.085	0.0096	0.030	**0.00031**	**0.0022**	**0.00076**	**0.0044**

Table 4.4: Continued

Category	GO ID	Gene Num	Annotation	sch9Δ/wt3		rasΔ/wt3		torΔ/wt3		sch9sir2Δ/wt3		sch9sir2Δ/sch9Δ	
				p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
P	GO:0009060	82	aerobic respiration	**2.66E-08**	**8.96E-07**	**1.32E-06**	**3.01E-05**	**4.73E-09**	**1.78E-07**	**0.00013**	**0.0011**	0.86	0.64
P	GO:0007005	95	mitochondrion organization and biogenesis	**6.62E-05**	**0.00070**	**4.51E-06**	**8.02E-05**	**0.00013**	**0.0011**	**4.23E-05**	**0.00051**	0.017	0.044
P	GO:0000398	97	nuclear mRNA splicing\, via spliceosome	0.010	0.031	**6.16E-06**	**0.00010**	0.066	0.12	**4.13E-06**	**7.56E-05**	**1.29E-06**	**3.01E-05**
P	GO:0016570	59	histone modification	**0.0016**	**0.0078**	**7.30E-06**	**0.00011**	**0.00022**	**0.0016**	**4.30E-05**	**0.00051**	**0.00080**	**0.0046**
P	GO:0006367	44	transcription initiation from RNA polymerase II promoter	0.032	0.071	**9.23E-06**	**0.00014**	0.0086	0.028	**0.00054**	**0.0033**	**6.20E-05**	**0.00067**
P	GO:0045944	48	positive regulation of transcription from RNA Pol II promoter	0.11	0.17	**4.17E-05**	**0.00051**	**0.00016**	**0.0013**	0.0047	0.018	0.039	0.081
P	GO:0006473	38	protein amino acid acetylation	0.011	0.035	**0.00012**	**0.0010**	0.0062	0.022	**0.00071**	**0.0042**	**0.0013**	**0.0070**
P	GO:0006406	60	mRNA-nucleus export	0.045	0.088	**0.00014**	**0.0011**	**0.00037**	**0.0024**	**0.00034**	**0.0023**	**0.00011**	**0.00098**
P	GO:0006119	46	oxidative phosphorylation	**7.03E-07**	**2.04E-05**	**0.00016**	**0.0012**	**9.01E-07**	**2.31E-05**	**0.00083**	**0.0048**	0.81	0.64
P	GO:0006626	47	protein-mitochondrial targeting	**8.33E-06**	**0.00013**	**0.00040**	**0.0026**	**1.46E-06**	**3.11E-05**	**0.00086**	**0.0049**	0.38	0.43
P	GO:0006402	55	mRNA catabolism	0.070	0.12	**0.00096**	**0.0053**	0.0031	0.013	0.019	0.048	0.022	0.054
P	GO:0006289	31	nucleotide-excision repair	0.12	0.18	**0.0017**	**0.0085**	0.034	0.074	0.019	0.049	0.033	0.072
P	GO:0001403	30	invasive growth	0.54	0.54	0.0022	0.010	0.016	0.043	0.0073	0.024	**0.00025**	**0.0019**
P	GO:0006365	67	35S primary transcript processing	**0.0019**	**0.0092**	0.0040	0.016	**3.84E-06**	**7.23E-05**	**5.10E-05**	**0.00056**	**0.0015**	**0.0077**
P	GO:0006118	31	electron transport	**0.00012**	**0.0010**	0.0043	0.017	**0.00010**	**0.00093**	0.12	0.18	1.00	0.64
P	GO:0006906	30	vesicle fusion	0.19	0.26	0.0053	0.020	0.30	0.37	0.0029	0.013	**2.70E-05**	**0.00035**
P	GO:0006383	38	transcription from RNA polymerase III promoter	0.022	0.053	0.0062	0.022	0.024	0.057	**0.0016**	**0.0079**	**0.0016**	**0.0079**
P	GO:0006611	44	protein-nucleus export	0.041	0.083	0.0068	0.023	**0.00062**	**0.0037**	0.021	0.051	0.084	0.14
P	GO:0016044	31	membrane organization and biogenesis	**0.0021**	**0.0098**	0.0097	0.031	**0.0014**	**0.0072**	0.012	0.035	0.28	0.35
P	GO:0043414	35	biopolymer methylation	**0.0011**	**0.0059**	0.013	0.038	0.0029	0.013	**0.00012**	**0.0010**	0.0038	0.015
P	GO:0006413	46	translational initiation	0.020	0.051	0.036	0.076	**0.0011**	**0.0059**	0.022	0.053	0.11	0.17
P	GO:0030490	46	processing of 20S pre-rRNA	**0.00035**	**0.0023**	0.042	0.084	**1.32E-06**	**3.01E-05**	**0.00011**	**0.00096**	0.016	0.043
P	GO:0007166	40	cell surface receptor linked signal transduction	0.99	0.64	0.12	0.18	0.63	0.61	0.028	0.064	**0.00023**	**0.0017**
P	GO:0009064	43	glutamine family amino acid metabolism	0.048	0.092	0.13	0.19	**8.42E-05**	**0.00082**	0.026	0.061	0.29	0.36
P	GO:0000154	89	rRNA modification	0.050	0.093	0.19	0.26	**0.0011**	**0.0059**	0.019	0.048	0.064	0.11
P	GO:0000749	53	response to pheromone during cellular fusion	0.90	0.64	0.23	0.30	0.80	0.64	0.0055	0.020	**1.78E-05**	**0.00025**
P	GO:0008652	99	amino acid biosynthesis	0.66	0.62	0.41	0.45	**2.00E-06**	**4.13E-05**	**1.80E-05**	**0.00025**	**4.05E-09**	**1.62E-07**
P	GO:0000096	32	sulfur amino acid metabolism	0.80	0.64	1.00	0.64	0.86	0.64	0.0072	0.024	**0.00022**	**0.0017**

Positively affected TIGO categories

Our TIGO analysis shows that the GO categories that are associated with cytosolic large ribosomal subunit (GO:0005842) and cytosolic small ribosomal subunit (GO:0005843) are positively affected in all the long-lived mutants. Namely, genes that belong to these categories tend to be up-regulated in the mutants with respect to the wild type. The GO category associated with monosaccharide catabolism is also positively affected. Genes in this GO category are involved in chemical reactions and pathways that result in the breakdown of monosaccharides, polyhydric alcohols containing either an aldehyde or a keto group. This may reflect the enhancement of cells from the the mutants to consume glucose in the medium. Although most of the TIGOs are similarly affected in the four mutants, some of them are specifically affected in certain mutants. For example, the GO category associated with cell wall is negatively affected (notable but not significant)in the $sch9\triangle$, but positively affected in the all other three mutants.

Negatively affected TIGO categories

More TIGO categories are negatively affected in the long-lived mutants, which include categories that are associated with aerobic respiration (GO:0009060), mitochondria organization and biogenesis (GO:0007005), histone modification (GO:0016570), oxidative phosphorylation (GO:0006119), and so on (see Table 4.3.3. As can be seen, many of the negatively affected TIGOs are related to mitochondria, which may imply a lower mitochondria metabolic rate in these mutants. The rate of energy generation and consumption in the mutants may be reduced, because the aerobic respiration (GO:0009060), the oxidative phosphorylation (GO:0006119), and the electron transport (GO:0006118) categories are all negatively affected. In addition, the GO categories associated with global transcription and translation are also negatively affected, which include GO

categories that are related to transcription initiation from RNA polymerase II promoter (GO:0006367), transcription from RNA polymerase III promoter (GO:0006383), mRNA metabolism (GO:0006402), the mRNA-nucleus export (GO:0006406), processing of 20S pre-rRNA (GO:0030490) and translation initiation (GO:0006413). On the other hand, the category associated with proteasome complex (GO:0000502), which catalyzes protein degradation is also negatively affected. Therefore, we may expect a low rate of gene transcription and protein translation, as well as a low rate of protein degradation in the long-lived mutants. Cells of long-lived mutants survive in an economical style in comparison with the wild type. The low metabolic rates in these mutants remind us about the similar features that appear in yeast cells under calorie restriction (CR). Deletion of Sch9, Ras2 or Tor1 may imitate the responses to CR and thereby extend the life span. Namely, the life span extension under CR may depend on a mechanism in which SCH9, RAS2, and TOR1 are involved.

When we compare $sch9sir2\triangle$ with $sch9\triangle$, the significantly affected TIGO categories are different from those in the four mutants (compared with wild type). As known, deletion of Sch9 increases the life span by three fold. Double deletion of Sch9 and Sir2 extends the life span up to six fold, although single deletion of Sir2 cause no significant change of life span. The difference in significantly affected TIGOs may imply a different mechanism of further life span extension in the double mutant $sch9sir2\triangle$ with respect to single mutant $sch9\triangle$.

4.3.4 Significantly affected pathways in the long-lived mutants

Despite the informativeness of GO categories, they are simply sets of genes with associated functions. For example, the biological process categories in GO are not equivalent to pathways. As a matter of fact, GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway. To investigate the mechanism

Table 4.5: Positively and negatively affected pathways in the long-lived mutants. Significant affected pathways (q-value$\leq$0.01) are shown in bold.

Positively affected pathways

Pathway	sch9△/wt		ras△/wt		tor△/wt		sch9sir2△/wt		sch9sir2△/sch9△	
	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
Starch and sucrose metabolism	0.58	0.82	0.0046	0.045	0.046	0.21	0.0032	0.035	**4.4E-06**	**2.1E-04**
N-Glycan biosynthesis	0.23	0.54	**8.8E-05**	**0.0032**	0.048	0.21	0.0052	0.046	7.4E-04	0.016
Glycolysis / Gluconeogenesis	0.0022	0.025	**8.2E-07**	**4.8E-05**	**1.5E-04**	**0.0050**	**2.1E-05**	**8.7E-04**	9.5E-04	0.018
Galactose metabolism	0.30	0.64	**2.7E-04**	**0.0072**	0.020	0.12	0.035	0.18	0.0014	0.023
Glycan structures - biosynthesis	0.026	0.15	**1.8E-04**	**0.0052**	0.030	0.16	0.0041	0.042	0.015	0.10
Fructose and mannose metabolism	0.0085	0.067	**3.0E-04**	**0.0072**	0.0050	0.045	0.0019	0.024	0.033	0.17
Ribosome	**0**	**0**	**3.7E-15**	**2.7E-13**	**0**	**0**	**2.2E-16**	**2.2E-14**	1.00	0.82

Negatively affected pathways

Pathway	sch9△/wt		ras△/wt		tor△/wt		sch9sir2△/wt		sch9sir2△/sch9△	
	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
Basal transcription factors	**1.8E-04**	**0.0042**	**1.6E-07**	**2.8E-05**	**1.7E-04**	**0.0042**	**1.9E-06**	**1.4E-04**	**1.9E-05**	**8.3E-04**
Citrate cycle (TCA cycle)	**3.3E-04**	**0.0073**	0.0077	0.082	**2.5E-05**	**9.3E-04**	0.012	0.100	0.73	0.86
Oxidative phosphorylation	**6.0E-06**	**3.6E-04**	0.0041	0.054	**2.7E-05**	**9.3E-04**	0.0078	0.082	0.96	0.86
Proteasome	2.1E-03	0.034	**1.8E-07**	**2.8E-05**	0.0057	0.073	**3.6E-05**	**0.0011**	**7.9E-05**	**0.0022**
Ribosome	1.00	0.86	1.00	0.86	1.00	0.86	1.00	0.86	**1.0E-06**	**1.1E-04**
SNARE interactions in vesicular transport	0.14	0.49	0.0035	0.048	0.40	0.78	0.0014	0.026	**8.8E-06**	**4.5E-04**

of longevity of the long-lived mutants, we analyze the difference between them and the wild type from a pathway perspective. We downloaded 102 pathways of *S.cerevisiae* from the KEGG database [oGG] and identified all the significantly changed pathways in the mutants with respect to the wild type. Totally 7 pathways are positively affected and 6 pathways are negatively affected in at least one of the five comparisons: $sch9\triangle/wt$, $ras2\triangle/wt$, $tor1\triangle/wt$, and $sch9sir2\triangle/sch9\triangle$ (see Table 4.5).

Positively affected pathways

As shown in Table 4.5, the Glycolysis/Gluconeogenesis pathway is positively affected in the four long-lived mutants with respect to the wild type. Glycolysis includes 10 reactions occurring in the cytosol that converts glucose into pyruvate. In aerobic organism, glycolysis is the prelude to the citric acid cycle (TCA) and the electron transport chain in oxidative phosphorylation. The glycolytic pathway has a dual role: it degrades glucose to generate ATP, and it provides building blocks for the synthesis of cellular components. Gluconeogenesis is the synthesis of glucose from noncarbohydrate sources, such

as lactate, amino acids, and glycerol. Our results suggest that genes involved in the Glycolysis/Gluconeogenesis pathway tend to be up-regulated in the mutants, which may result in a enhancement of the cells to make use of glucose or other carbon sources.

Two other sugar related pathways, the galactose metabolism and the fructose and mannose metabolism, are also positively affected in the long-lived mutants, though they are affected as significantly as the Glycolysis/Gluconeogenesis pathway. The ribosome pathway (not include mitochondria ribosomal subunits) is also positively affected, which is consistent with our GO analysis.

Negatively affected pathways

In the long-lived mutants, TCA cycle and oxidative phosphorylation pathways are negatively affected. The TCA cycle, also called citric cycle, is the final common pathway for the oxidation of fuel molecules. It also serves as a source of building blocks for biosynthesis. The TCA cycle operates only under aerobic conditions, because it requires a supply of NAD^+ and FAD, which are changed into NADH and $FADH_2$ after accepting electrons. These electron acceptors are regenerated when NADH and $FADH_2$ transfer their electrons to O_2 through the electron transport chain. In oxidative phosphorylation, the electron transport chain is coupled to the synthesis of ATP by a proton gradient across the inner mitochondria membrane. Oxidative phosphorylation is the major source of ATP in aerobic organisms. In yeast, the reaction of the TCA cycle and oxidative phosphorylation occur inside the mitochondria, in contrast to those of glycolysis, which occur in the cytosol. Under aerobic conditions, oxidative phosphorylation is efficient to generate ATPs, but at the same time it produces the reactive oxygen species (ROS) as by products, which is thought to be one of the causes of ageing. The repression of the two pathways in the long-lived mutants may provide us some hints about the mechanism of longevity of them.

The basal transcription factors form a complex that acts as a general transcription machine. Interestingly, we found that the complex is negatively affected, or down-regulated in expression, in all of the long-lived mutants. This is also consistent with the results obtained by GO analysis. The down-regulation of the basal transcription factors may reflect the low metabolic rate in these mutants. The cells live a economical life and thereby only a low basal transcription is required to maintain survival. Also, proteasome, the complex in charge of protein degradation, is negatively affected in the long-lived mutants.

4.3.5 Significantly affected cellular components in the long-lived mutants

Table 4.6: Positively and Negatively affected Cellular organelles. Significant findings (q-value$\leq$0.01) are shown in bold.

Positively affected cellular organelles

Cellular Organelle	sch9Δ/wt		rasΔ/wt		torΔ/wt		sch9sir2Δ/wt		sch9sir2Δ/sch9Δ	
	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
ER	**4.2E-04**	**0.0033**	**0.0E+00**	**0.0E+00**	**3.7E-09**	**8.0E-08**	**3.1E-08**	**5.4E-07**	**5.0E-10**	**2.2E-08**
vacuole	**6.1E-05**	**5.8E-04**	**1.1E-09**	**3.2E-08**	**3.1E-06**	**3.5E-05**	**4.5E-07**	**6.5E-06**	**2.0E-04**	**0.0017**
vacuolar membrane	0.0042	0.022	0.038	0.14	**4.7E-04**	**0.0033**	0.049	0.2	0.34	0.64
actin	**0.0014**	**0.0080**	0.32	0.62	**9.4E-04**	**0.0058**	0.046	0.1	0.66	0.78
endosome	0.040	0.14	0.66	0.78	**7.5E-04**	**0.0050**	0.4	0.7	0.96	0.78
punctate composite	0.083	0.24	0.66	0.78	**0.0016**	**0.0085**	0.4	0.7	0.89	0.78
cytoplasm	**3.3E-06**	**3.5E-05**	0.99	0.78	0.016	0.078	0.8	0.8	1.00	0.78

Negatively affected cellular organelles

Cellular Organelle	sch9Δ/wt		rasΔ/wt		torΔ/wt		sch9sir2Δ/wt		sch9sir2Δ/sch9Δ	
	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
mitochondrion	**3.9E-37**	**1.6E-35**	**8.4E-29**	**1.4E-27**	**1.4E-43**	**1.1E-41**	**3.0E-24**	**3.5E-23**	0.028	0.081
nucleus	**9.3E-07**	**5.4E-06**	**3.5E-31**	**7.0E-30**	**9.1E-12**	**7.4E-11**	**1.5E-25**	**2.1E-24**	**1.2E-31**	**3.1E-30**
nucleolus	**5.1E-09**	**3.2E-08**	**4.6E-06**	**2.5E-05**	**1.1E-10**	**7.2E-10**	**2.1E-15**	**2.1E-14**	**1.2E-11**	**9.2E-11**
cytoplasm	1.00	0.74	0.0080	0.028	0.98	0.74	0.19	0.36	**5.6E-12**	**5.0E-11**
bud neck	0.38	0.59	0.0032	0.014	0.30	0.52	0.0067	0.026	**7.6E-05**	**3.8E-04**
spindle pole	0.01	0.05	0.00	0.01	0.16	0.33	**0.00**	**0.01**	**0.00**	**0.01**

It is well known that some cellular organelles play important roles in biological processes. For example, mitochondria is the organelle where TCA cycle and oxidative phosphorylation occur and is highly associated with aging. This motivates us to think

about such a question: genes localized in which organelles are more likely to be affected in the long-lived mutant strains? Large-scale analysis of protein localization has been performed in *S.cerevisiae*, which enables us to investigate this problem [HFG+03]. As shown in Table 4.6, we identified the significantly affected organelles. Our results indicate that ER-located and vacuole-located proteins are positively affected, while proteins located in mitochondria, nucleus or nucleolus are negatively affected in all of the long-lived mutants.

The endoplasmic reticulum is part of the endomembrane system, which modifies proteins, makes macromolecules, and transfers substances throughout the cell. In budding yeast cells, vacuoles are the storage compartments of amino acids and the detoxification compartments. Under conditions of starvation, proteins are degraded in vacuoles, which is called autophagy. The up-regulations of vacuole-located proteins may implies that autophagy in the cells of these long-lived mutants is enhanced to maintain survival in low nutrient conditions, such as SDC medium.

A dominant role for the mitochondria is the production of ATP as reflected by the large number of proteins in the inner membrane for this task. This is done by oxidizing the major products of glycolysis: pyruvate and NADH that are produced in the cytosol. This process of cellular respiration, also called aerobic respiration, is dependant on the presence of oxygen. When oxygen is limited the glycolytic products will be metabolized by anaerobic respiration, a process that is independent of the mitochondria. The production of ATP from glucose has an approximately 15 fold higher yield during aerobic respiration compared to anaerobic respiration. Our analysis shows that mitochondrial proteins tend to be down-regulated in the transcription level. This may reflect a switch from aerobic respiration to anaerobic respiration for energy. This is consistent with previous results from pathway analysis: the Glycolysis/Gluoconeogenesis pathway is positively affected, whereas the TCA cycle and oxidative phosphorylation

are negatively affected. Additionally, proteins localized in nucleus or nucleolus tend to be down-regulated, which is also a reflection of metabolic rate decrease.

4.3.6 Significantly affected transcription factors in the long-lived mutants

Table 4.7: Positively and Negatively affected transcription factors. Significant findings (q-value$\leq$0.01) are shown in bold.

Positively affected transcriptional factors

Transcription Factor	sch9Δ/wt		rasΔ/wt		torΔ/wt		sch9sir2Δ/wt		sch9sir2Δ/sch9Δ	
	p-value	**q-value**	**p-value**	**q-value**	**p-value**	**q-value**	**p-value**	**q-value**	**p-value**	**q-value**
CIN5	0.0082	0.074	**6.9E-05**	**0.0019**	0.0015	0.021	**7.0E-06**	**3.2E-04**	**2.0E-04**	**0.0043**
FHL1	**0**	**0**	**3.2E-15**	**2.5E-13**	**0**	**0**	**0**	**0**	1.00	0.79
HAP1	0.89	0.79	0.34	0.57	0.92	0.79	0.042	0.18	**1.4E-05**	**5.7E-04**
INO4	1.00	0.79	0.97	0.79	0.92	0.79	0.92	0.79	**1.1E-04**	**0.0027**
MBP1	0.29	0.53	0.051	0.20	**3.0E-05**	**0.0011**	0.099	0.29	0.11	0.31
MET31	0.11	0.32	**6.4E-05**	**0.0019**	0.0053	0.054	0.80	0.79	0.98	0.79
NRG1	0.62	0.79	8.4E-04	0.015	0.077	0.26	5.9E-04	0.011	**1.8E-05**	**6.8E-04**
RAP1	**0**	**0**	**5.9E-10**	**3.4E-08**	**0**	**0**	**4.0E-13**	**2.6E-11**	0.93	0.79
SUM1	0.88	0.79	0.17	0.40	0.045	0.19	0.073	0.25	**4.8E-04**	**0.0093**
SWI4	0.12	0.32	0.030	0.15	**1.2E-04**	**0.0028**	0.050	0.20	0.14	0.35
SWI5	0.0018	0.024	0.024	0.13	**2.4E-04**	**0.0051**	9.9E-04	0.016	0.24	0.49
SWI6	0.28	0.51	0.041	0.18	**3.2E-05**	**0.0011**	0.048	0.19	0.011	0.086
YAP5	0.0012	0.017	0.0035	0.043	**7.9E-05**	**0.0020**	0.0012	0.017	0.32	0.56
YAP6	0.039	0.18	**1.1E-06**	**5.6E-05**	9.6E-04	0.016	**4.9E-05**	**0.0015**	**2.6E-04**	**0.0052**

Negatively affected transcriptional factors

Transcription Factor	sch9Δ/wt		rasΔ/wt		torΔ/wt		sch9sir2Δ/wt		sch9sir2Δ/sch9Δ	
	p-value	**q-value**	**p-value**	**q-value**	**p-value**	**q-value**	**p-value**	**q-value**	**p-value**	**q-value**
ABF1	0.0012	0.048	**1.2E-06**	**1.8E-04**	**2.0E-06**	**1.8E-04**	1.1E-04	0.0063	0.0037	0.10
ARG80	0.0056	0.14	0.0055	0.14	**2.0E-05**	**0.0013**	0.013	0.23	0.31	0.79
GCN4	0.66	0.91	0.046	0.40	**2.9E-08**	**7.8E-06**	9.1E-06	6.9E-04	7.0E-10	3.8E-07
HAP4	**1.8E-06**	**1.8E-04**	6.8E-04	0.033	**1.6E-06**	**1.8E-04**	8.4E-04	0.037	0.68	0.91
INO4	**1.3E-04**	**0.0067**	0.034	0.33	0.076	0.47	0.076	0.47	1.00	0.91

In most cases, an external or internal signal will eventually be transmitted to one or a set of transcription factors, and as a consequence gene expressions are change to respond to the signal. If we can find out the transcription factors that cause the gene expression changes in the long-lived mutants, it will be helpful to infer the underlying mechanism of longevity. Unfortunately, gene expression in microarray data provides limited information to detect the change of transcription factor activity. The reasons are as follows: (1) The expression levels of transcription factors are relatively low. (2)The activities of

transcription factor are prevalently regulated by post-translational modification, e.g. by protein phosphorylation. As such, we apply an indirect strategy to find out the affected transcription factors in these mutants by studying the expression change of genes regulated by those transcription factors. For a given transcription factor, if the expression levels of its target genes are significantly up-regulated in comparison with the whole transcriptome background, we conclude that the activity of this transcription factor is enhanced. Conversely,if the expression levels of its target genes are significantly down-regulated, we assume that the activity of this transcription factor is repressed. To determine the target gene set of a transcription factor, we use the TF-gene binding information provided by the ChIP-Chip data. Large scale ChIP-Chip experiments have carried out to systematically identify the binding sites of 203 transcription factors in *S. cerevisiae* [HGL$^+$04]. Table 4.7 shows the transcription factors that are significantly activated or inactivated in various comparisons. Note that FHL1 and RAP1 are significantly activated in all the 4 mutants: $sch9\triangle$, $ras2\triangle$, $tor1\triangle$ and $sch9sir2\triangle$, relative to wild type. This is consistent with what we expect, because we know that RAP1 and FHL1 are responsible for the regulation of ribosomal protein genes. In addition, we find that SUM1 is significantly activated in $sch9sir2\triangle$ with respect to $sch9\triangle$. Previous studies have shown that SUM1 is a transcriptional repressor required for repression of middle sporulation-specific genes during mitosis; and that a dominant mutation of SUM1 is able to suppress the silencing defects of SIR2 mutations [LR91, XPGD$^+$99, FGB$^+$05]. So the activation of SUM1, in $sch9sir2\triangle$ relative to $sch9\triangle$, may reflect a feedback in response to Sir2 deletion.

Table 4.8: Motifs enriched in up-regulated genes.Significant findings (q-value$\leq$0.01) are shown in bold.

Consensus Sequence	Transcription Factor	sch9Δ/wt		rasΔ/wt		torΔ/wt		sch9sir2Δ/wt		sch9sir2Δ/sch9Δ	
		p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value	p-value	q-value
RRTCACGTG-	CBF1	0.36	0.24	**4.2E-05**	**0.0028**	0.13	0.14	0.10	0.12	0.64	0.33
RTGT-YGGRTG	FHL1	9.3E-04	0.012	0.0042	0.024	**1.7E-06**	**3.2E-04**	**4.2E-04**	**0.0074**	0.42	0.26
AWAGGGAT	GIS1	**5.3E-05**	**0.0028**	**2.8E-04**	**0.0059**	0.058	0.088	**1.3E-05**	**0.0012**	0.16	0.15
AMAA-TGTGG	MET4	0.12	0.13	**8.4E-06**	**9.5E-04**	0.67	0.34	0.056	0.086	0.13	0.14
CGCATMCCCCAC	MIG1	0.022	0.054	**4.9E-04**	**0.0083**	0.036	0.068	8.3E-04	0.011	0.082	0.11
AGGGG	MSN2/4	**1.3E-06**	**3.2E-04**	**1.1E-04**	**0.0038**	**5.6E-05**	**0.0028**	**1.4E-06**	**3.2E-04**	0.16	0.15
GY–TSKCACGTG-G	PHO4	0.0024	0.018	**5.0E-05**	**0.0028**	0.0016	0.016	0.0075	0.030	0.61	0.33
G-RGGGG-GGGG	STRE	0.0014	0.015	0.045	0.077	0.0012	0.014	**4.4E-04**	**0.0076**	0.058	0.088
RYGWCASWAAW	SUM1	0.11	0.13	**2.0E-04**	**0.0051**	0.0050	0.026	0.0057	0.027	0.0011	0.014
ACCYT-AGGTT	ZAP1	0.30	0.21	8.7E-04	0.012	0.56	0.31	**2.5E-04**	**0.0057**	**5.5E-05**	**0.0028**

4.3.7 Significantly enriched motifs in promoter regions of differentially expressed genes

Although the transcription factor analysis based on ChIP-chip data provides us some information about transcriptional regulation in these long-lived mutants, it has the following limitations: (1)The ChIP-chip experiments are performed using yeast cells at log-phase in YPD (Yeast Peptone Dextrose) medium. However, our microarray experiments are carried out using yeast cells collected at day 2.5 with SDC (Synthetic Dextrose Complete) as medium. It is known that some transcription factors regulate different sets of target genes in different cell stages or different conditions. So it may be inappropriate to determine the target gene sets for some transcription factors according to the available Chip-Chip data. (2) The binding information for some transcription factors are missed in the ChIP-Chip data. For example, it has been known that RAS2/CYR1/PKA, TOR1 and SCH9 activate several transcription factors, such as MSN2/4 and GIS1. These transcription factors regulate the expression of STRE/PDS controlled genes, in which many are stress response genes [Lon03, REM$^+$05]. However, the binding information for GIS1 protein is not available in ChIP-chip data.

To overcome these limitations, we apply a systematic in-silicon analysis to identify the motifs that are significantly enriched in the up-regulated or down-regulated gene set for each mutant. Beer et al. identified 666 non-redundant motifs from 800bp upstream

sequences of all genes in *S.cerevisiae* [BT04]. Among these motifs, 51 have known binding transcription factors. To find out the transcription factors that are associated with differentially expressed genes in $sch9\triangle$, $ras2\triangle$, $tor1\triangle$ and $sch9sir2\triangle$ mutants, we analyze the enrichment of motifs in both up-regulated and down-regulated genes in each mutant. Our results show that there is no motif enriched in the down-regulated gene set for all the comparisons: $sch9\triangle/wt$,$ras2\triangle/wt$, $tor1\triangle/wt$, $sch9sir2\triangle/wt$ and $sch9sir2\triangle/sch9\triangle$. Whereas in the up-regulated gene sets, we find some significantly enriched motifs as shown in Table 4.8. It is notable that the motif bound by Gis1 is enriched in $sch9\triangle/wt$, $ras2\triangle/wt$, $sch9sir2\triangle/wt$ comparisons, and MSN2/4a binding motif is enriched in $sch9\triangle/wt$, $ras2\triangle/wt$, $tor1\triangle/wt$ and $sch9sir2\triangle/wt$ comparisons. These results are consistent with previous knowledge about MSN2/4. Both MSN2 and MSN4 are repressed by PKA, which is activated by Ras2 and Tor1 protein. Also we know that GIS1 is activated by glucose-repressible protein kinase RIM15, whose activity is inhibited by both PKA and SCH9 kinase [PDC$^+$03]. Therefore the mechanism of life span extension of $sch9\triangle$, $ras2\triangle$, $tor1\triangle$ and $sch9sir2\triangle$ can at least partially be explained by the activation of MSN2/4 and/or GIS1 in these mutants.

It should be noted that both MSN2 and MSN4 are condition altered transcription factors according to Harbison et al's paper [HGL$^+$04]. Namely, they bind to different sets of target genes in different conditions. Since the medium and cell phase used in our study are different from those used in the ChIP-Chip experiments, we do not find the activation of MSN2 or MSN4 using the ChIP-Chip based method. But we detect the enrichment of MSN2/4 binding site in promoter regions of the up-regulated genes using the motif analysis.

4.4 Conclusions and discussions

4.4.1 Energy switch

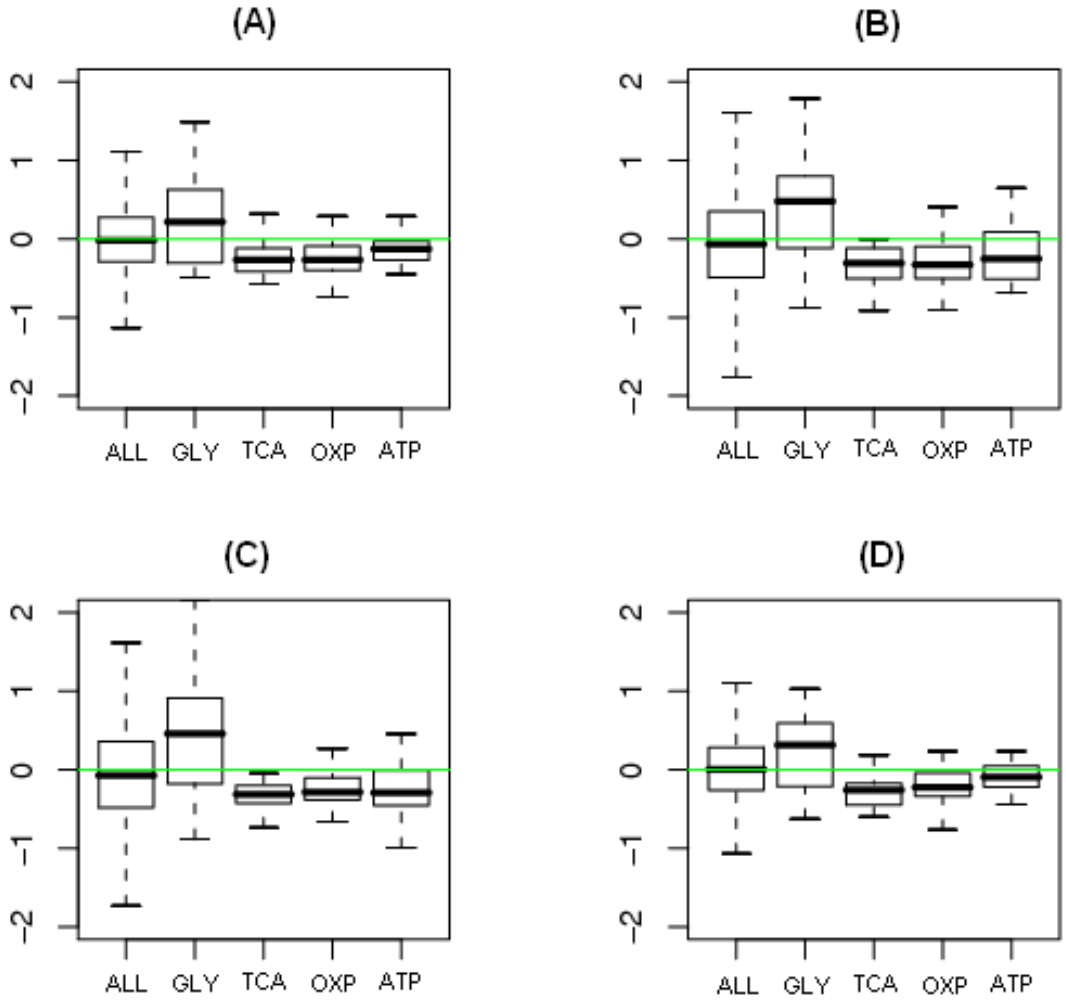

Figure 4.4: Box-plots of log ratios in the long-lived mutants. (A) $sch9\triangle/wt$; (B) $ras2\triangle/wt$; (C) $sch9sir2\triangle/wt$; (D) $tor1\triangle/wt$. ALL- all genes; GLY- Glycolysis/Gluconeogenesis; TCA- citric acid cycle; OXP- oxidative phosphorylation; ATP- ATP generation.

In the long-lived mutants, genes involved in Glycolysis/Gluconeogenesis tend to be up-regulated; and genes that participate in TCA cycle and oxidative phosphorylation

tend to be down-regulated. This interesting phenomenon is observed consistently in all the four long-lived mutants and supported by both GO and pathway analysis as shown in Figure 4.4. On one hand, the up-regulation of Glycolysis/Gluconeogenesis related genes imply that cells from the mutants consume the carbon sources in a more efficient and economical manner compared with the wild type. The change may be achieved through a mechanism similar to the one in CR. In consistent with this hypothesis, a recent study shows that calorie restriction of $tor1\triangle$ or $sch9\triangle$ cells failed to further increase of the life span [KPS$^+$05]. On the other hand, the down-regulation of TCA cycle and oxidative phosphorylation related genes indicates that mutant cells switch to alternative energy pathways, possibly glycolysis, for energy. These pathways depend less on TCA cycle and oxidative phosphorylation and consume less O_2. Consequently, it may also produce less ROS in comparison with the wild type cells.

Rea et al. proposed a metabolic model to describe the "Energy switch" hypothesis for longevity mutants in *C.elegans* [RJ03]. According to their hypothesis, the relative balance between TCA based mitochondrial-dependent metabolism and alternative pathways that do not involve the electron transport chain or are independent of the mitochondria may determine the overall oxidant burden and hence the life span. In *C.elegans*, alternative energy pathways include malate dismutation. Our results indicate that the "energy switch" may be used to explain the long-lived mutations in *S.cerevisiae*. In $sch9\triangle$, $tor1\triangle$, $ras2\triangle$, and $sch9sir2\triangle$, the alternative energy pathway is likely to be the glycolysis pathway that occurs in the cytosol. The energy switch in these long-lived mutants, together with other effects, e.g. low metabolic rate and enhanced stress resistance, may play important roles in life span extension.

4.4.2 Stress resistance

Deletion of RAS2 or SCH9 increases the stress resistance of yeast cells. MSN2/MSN4 and SOD2 are required for longevity extension in $ras2\triangle$, which suggests that these stress response genes play important roles in longevity [LGV96, Lon03, FLM^{+}03]. Similarly, the life span extension by SCH9 deletion requires RIM15 but

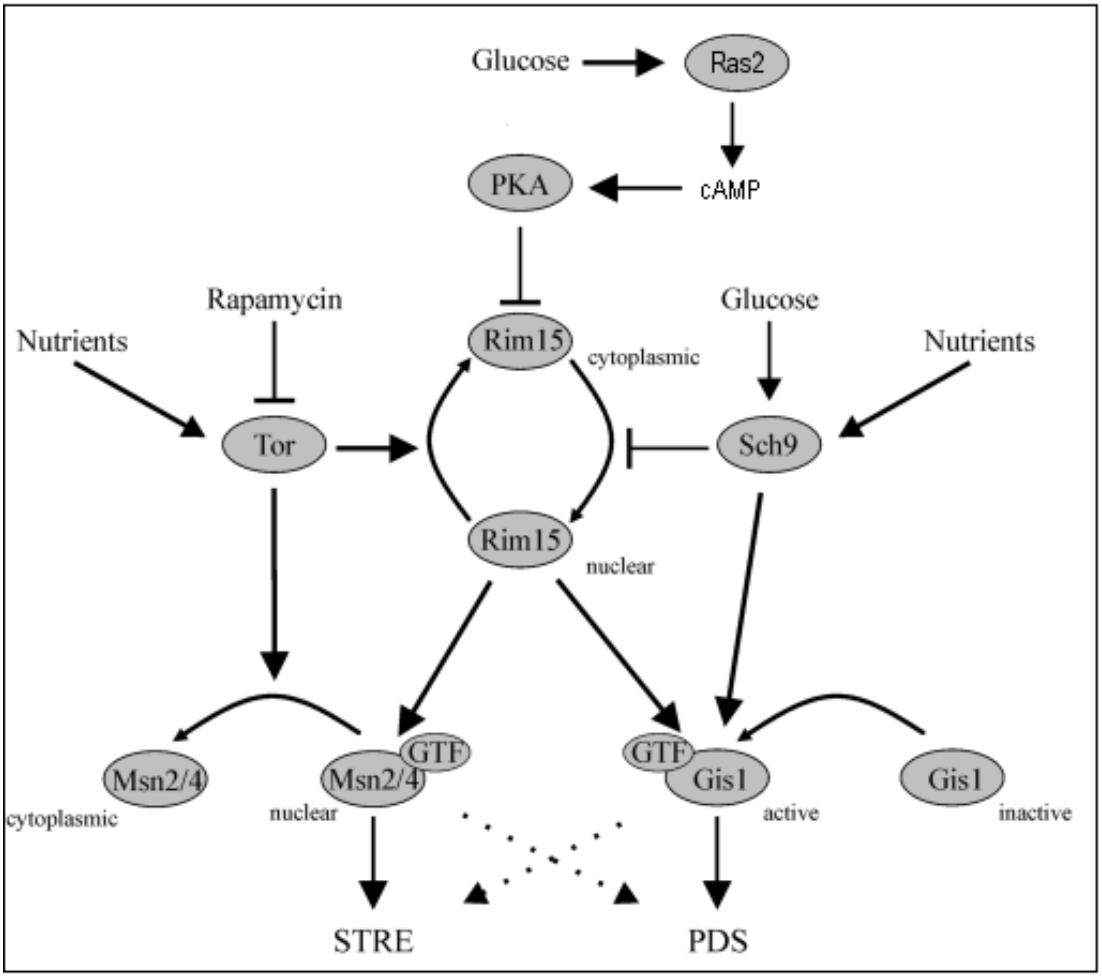

Figure 4.5: Sch9 and TOR signalling are subject to cAMP-gating in yeast. GTF stands for general transcription complex. Arrows and bars refer to positive and negative interactions. Dashed lines refer to potential cross-regulation. *The figure is copied from Roosen et al. (Molecular Microbiology, Vol.55, 862-880) with small revisions.*

not MSN2/MSN4 [FPP$^+$01]. Our motif analysis suggests that two transcription factors, MSN2/MSN4 and GIS1, may function downstream of the pathway to regulate the expression of stress response genes. As shown in Table 4.8, the motifs bound by MSN2/MSN4 are significantly enriched in the up-regulated genes from all the four long-lived mutants; the motifs bound by GIS1 are significantly enriched in the up-regulated genes from all mutants except $tor1\triangle$ (the p-value is 0.058). Our results support the model proposed by Roosen et al. [REM$^+$05]. As shown in Figure 4.5, a main gatekeeper, the protein kinase PKA, switches on or off the activities and signals transmitted through primary pathways such as SCH9 and TOR. SCH9 positively controls PDS-driven gene expression mainly via GIS1 and RIM15. TOR and PKA control STRE-driven gene expression mainly via MSN2/4 and RIM15. But cross-talks exist between the two pathways. Stress response genes, i.g. SOD2, facilitate the removal of the endogenous ROS, and enhance the stability of genome and mitochondria DNA. All these effects play positive roles in longevity.

4.4.3 Mitochondria and ageing

It has been 50 years since Harman first proposed the "free radical theory" of aging[Har56]. According to this theory, reactive oxygen species (ROS) damage macromolecules and thereby accelerate ageing. The majority of cellular ROS (approximately 90%) is generated in mitochondria as a byproduct of oxidative phosphorylation during respiration[BNF05]. A number of mutation affecting respiration have been found to increase life span, and at least some of them may achieve this by decreasing ROS levels[Ken05]. We find that many of the down-regulated genes encode mitochondrial proteins, and the expression levels of genes that encode proteins localized in mitochondria tend to be negatively affected in the long-lived mutants. Consistently, in these mutants, TCA and oxidative phosphorylation are negatively affected, both of which

occur in the mitochondria. As a consequence, respiration is reduced to some extent and thereby less ROS are produced. Our results suggest the importance of mitochondria in yeast ageing.

4.4.4 Low metabolic rate

Both the GO and the pathway analysis imply a low metabolic rate in the long-lived mutants. The basal transcription and translation are reduced by some extent. Cells with mutants may survive in a more economical mode, which consumes less ATP and possibly produces less harmful byproducts, such as ROS. As known, various organisms, including yeast, live with low metabolic rate under CR conditions. The metabolic rate reduction in the long-lived mutants with respect to the wild type again implies the strong association of SCH9, TOR1 and RAS2 with CR.

4.4.5 Future works

In the future, we may study the gene expression in these mutants with a time course experiment design. The time course gene expression data provides more abundant information to infer the regulatory network during ageing process. It may help us to know which phase is important for longevity. In *C.elegants*, treating worms with daf-2 RNAi from the time of hatching extends life span and delays reproduction, but treating them when they are young adults extends life span to the same extent with little or no effect on reproduction [DCK02]. This indicates that genes may function differently in different stages. The time course experiments are able to provide information to achieve more accurate and detail understanding about function of the ageing related genes.

In addition, we can study the interaction effects of ageing related genes. For example, we can survey the double mutants of Fob1 or Sir2 with Sch9, Tor1, Ras2 etc. This

will enable us to know whether these genes function independently or coordinately to change life span.

References

[ABB03] O. Alter, P. O. Brown, and D. Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci U S A*, 100:3351–3356, 2003.

[Aff] Affymetrix. http://www.affymetrix.com/index.affx.

[Aff01] *Affymetrix Microarray Suite User Guide,5th edition*. Affymetrix, Santa Clara, CA, 2001.

[ASBL99] U. Alon, M. G. Surette, N. Barkai, and S. Leibler. Robustness in bacterial chemotaxis. *Nature*, 397:168–171, 1999.

[ASGG99] K. Ashra, D. Sinclair, J. T. Gordon, and L. Guarente. Passage through stationary phase advances replicative ageing in saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, 96:9100–9105, 1999.

[Ast03] M. Astrand. Contrast normalization of oligonucleotide arrays. *J Comput Biol*, 10:95–102, 2003.

[BIApS03] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. p. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193, 2003.

[Bio] Bioconductor. http://www.bioconductor.org/.

[BJ04] Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20:2493–2503, 2004.

[BJGG+03] Z. Bar-Joseph, G. K. GerBER, D. K. Gifford, T. S. Jaakkola, and I. Simon. Continuous representation of time-series gene expression data. *J Comput Biol*, 10:341–356, 2003.

[BJGS+03] Z. Bar-Joseph, G. Gerber, I. Simon, D. K. Gifford, and T. S. Jaakkola. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc Natl Acad Sci U S A*, 100:10146–10151, 2003.

[BNF05] R. S. Balaban, S. Nemoto, and T. Finkel. Mitochondria, oxidants, and aging. *Cell*, 120:483–95, 2005.

[BO04] A. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5:101–113, 2004.

[BS96] M. G. Barker and K. A. Smart. Morphological changes associated with the cellular ageing of a brewing yeast strain. *J. Am. Soc. Brew. Chem.*, 54:121 126, 1996.

[BS04] T. Beissbarth and T.P. Speed. Gostat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20:1464–1465, 2004.

[BT04] M. A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117:185–198, 2004.

[CCL+99] M. E. Cardenas, N. S. Cutler, M. C. Lorenz, C. J. Di Como, , and J. Heitman. The tor signaling cascade regulates gene expression in response to nutrients. *Genes Dev*, 13:3271–3279, 1999.

[CF94] K. W. Cunningham and G. R. Fink. Calcineurin-dependent growth control in saccharomyces cerevisiae mutants lacking pmc1, a homolog of plasma membrane ca2+ atpases. *Journal of Cellular Biochemistry*, 124:351–363, 1994.

[Cle79] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, 74:829–836, 1979.

[CLZ+03] M. Caceres, J. Lachuer, A. Zapala, C. Redmond, L. Kudo H. Geschwind, J. Lockhart M. Preuss, and C. Barlow. Elevated gene expression levels distinguish human from non-human primate brains. *Proc Natl Acad Sci U S A*, 100:13030–13035, 2003.

[CTM+03] D. Chen, W. M. Toone, J. Mata, R. Lyne, G. Burns, K. Kivinen, A. Brazma, N. Jones, and J. Bähler. Global transcriptional responses of fission yeast to environmental stress. *Molecular Biology of the Cell*, 14:214–229, 2003.

[Cye03] M. S. Cyert. Calcineurin signaling in saccharomyces cerevisiae: how yeast go crazy in response to stress. *Biochem. Biophys. Res. Commun*, 311:1143–1150, 2003.

[DCF+94] N. P. D'mello, A. M. Childress, D. S. Franklin, S. P. Kale, C. Pinswasdi, and S. M. Jazwinski SM. Cloning and characterization of lag1, a longevity-assurance gene in yeast. *J Biol Chem*, 269:15451–15459, 1994.

[DCK02] A. Dillin, D. K. Crawford, and C. Kenyon. Timing requirements for insulin/igf-1 signaling in c. elegans. *Science*, 298:830–834, 2002.

[DHHR02] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18:105–110, 2002.

[DIB97] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[DPK+99] P. A. Defossez, R. Prusty, M. Kaeberlein, S. J. Lin, P. Ferrigno, P. A. Silver, R. L. Keil, and L. Guarente. Elimination of replication block protein fob1 extends the life span of yeast mother cells. *Mol Cell*, 3:447–455, 1999.

[DSH+04] M. K. Davidson, H. K. Shandilya, K. Hirota, K. Ohta, and W. P. Wahls. Atf1-pcr1-m26 complex links stress-activated mapk and camp-dependent protein kinase pathways via chromatin remodeling of cgs2+. *Journal of Cellular Biochemistry*, 279:50857–50863, 2004.

[DSWN+05] I. Dunand-Sauthier, C. A. Walker, J. Narasimhan, A. K. Pearce, R. C. Wek, and T. C. Humphrey. Stress-activated protein kinase pathway functions to support protein synthesis and translational adaptation in response to environmental stress in fission yeast. *Eukaryot Cell*, 4:1785–1793, 2005.

[DYCS02] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed. Statistical methods for identifying genes with differential expression in replicated cdna microarray experiments. *Statistica Sinica*, 12:111–139, 2002.

[EKK+02] W. Enard, P. Khaitovich, J. Z. Klose, S. öllner, F. Heissig, P. Giavalisco, S. Nieselt, E. Muchmore, A. Varki, and et al. R. Ravid. Intra- and interspecific varaiation in primate gene expreesion patterns. 296:340–343, 2002.

[ESBB98] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95:14863–14868, 1998.

[FBV+04] P. Fabrizio, L. Battistella, R. Vardavas, C. Gattazzo, L. L. Liou, A. Diaspro, J. W. Dossen, E. B. Gralla, and V. D. Longo. Superoxide is a mediator of an altruistic aging program in saccharomyces cerevisiae. *J. Cell Biol*, 166:1055–1067, 2004.

[FGB+05] P. Fabrizio, C. Gattazzo, L. Battistella, M. Wei, C. Cheng, K. McGrew, and V. D. Longo. Sir2 blocks extreme life-span extension. *Cell*, 123:655–667, 2005.

[FLM+03] P. Fabrizio, L. L. Liou LL, V. N. Moy, A. Diaspro, J. SelverstoneValentine, E. B. Gralla, and V. D. Longo. Sod2 functions downstream of sch9 to extend longevity in yeast. *Genetics*, 163:35–46, 2003.

[FPP+01] P. Fabrizio, F. Pozza, S. D. Pletcher, C. Gendron, , and V. D. Longo. Regulation of longevity and stress resistance by sch9 in yeast. *Science*, 292:288–290, 2001.

[GDSP98] F. Gaits, G. Degols, K. Shiozaki, and P.Russell. Phosphorylation and association with the transcription factor atf1 regulate localization of spc1/sty1 stress-activated kinase in fission yeast. *Genes Dev*, 12:1391–1397, 1998.

[GG03] J. Gu and X. Gu. Induced gene expression in human brain after the split from chimpanzee. *Trends in Genetics*, 19:63–65, 2003.

[GP05] L. Guarente and F. Picard. Calorie restriction- the sir2 connection. *Cell*, 120:473–482, 2005.

[Gua05] L. Guarente. Calorie restriction and sir2 genes–towards a mechanism. *Mech Ageing Dev*, 126:923–928, 2005.

[HA89] D. E. Harrison and J. R. Archer. Natural selection for extended longevity from food restriction. *Growth Dev Aging*, 53:3, 1989.

[Har56] D. Harman. Aging: a theory based on free radical and radiation chemistry. *JOURNAL OF GERONTOLOGY*, 11:298–300, 1956.

[Har72] D. Harman. A biologic clock: the mitochondria? *JOURNAL OF THE AMERICAN GERIATRICS SOCIETY*, 20:145–147, 1972.

[HATG01] H.A. H. A. Tissenbaum and L. Guarente. Increased dosage of a sir-2 gene extends lifespan in caenorhabditis elegans. *Nature*, 410:227–230, 2001.

[HBC+02] L. L. Hoopes, M. Budd, W. Choe, T. Weitao, and J. L. Campbell. Mutations in dna replication genes reduce yeast life span. *Mol Cell Biol*, 22:4136–4146, 2002.

[HFG+03] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J.S. Weissman, and E. K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425:686–691, 2003.

[HGL+04] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, M. Hannett, J. B. Tagne, and D. B. Reynolds et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431:99–104, 2004.

[HHS+02] W. Huber, V. S. Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:96–104, 2002.

[HLM+01] Q. Huang, D. Liu, P. Majewski, L. C. Schulte, J. M. Korn, R. A. Young, E. S. Lander, and N.Hacohen. The plasticity of dendritic cell responses to pathogens and their components. *Science*, 294:870–875, 2001.

[IAdC+04] D. K. Ingram, R. M. Anson, R. de Cabo, J. Mamczarz ADN M. Zhu, J. Mattison, M. A. Lane, and G. S. Roth. Development of calorie restriction mimetics as a prolongevity strategy. *Ann N Y Acad Sci*, 1019:412–423, 2004.

[IBC+03] R. A. Irizarry, B. Hobbs B, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, and T. P. Speed U. Scherf. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003.

[KCM02] T. Kepler, L. Crosby, and K. Morgan. Normalization and analysis of DNA microarray data by self-consistensy and local regression. *Genome Biology*, 3:1–12, 2002.

[Ken05] C. Kenyon. The plasticity of aging: insight from long-lived mutants. *Cell*, 120:449–460, 2005.

[KG03] J. Koubova and L. Guarente. How does calorie restriction work? *Genes Dev*, 17:313–321, 2003.

[Kir88] T. B. Kirkwood. The nature and causes of ageing. *Ciba Found Symp*, 134:193–207, 1988.

[Kir92] T. B. Kirkwood. Comparative life spans of species: why do species have the life spans they do? *Am J Clin Nutr*, 55:1191S–1195S, 1992.

[Kir02] T. B. Kirkwood. Evolution of ageing. *Mech Ageing Dev*, 123:737–745, 2002.

[KK05] M. Kaeberlein and B. K. Kennedy. Large-scale identification in yeast of conserved ageing genes. *Mech Ageing Dev*, 126:17–21, 2005.

[KKFK04a] M. Kaeberlein, K. T. Kirkland, S. Fields, and B. K. Kennedy. Genes determining yeast replicative life span in a long-lived genetic background. *Mech Ageing Dev*, 126:491–504, 2004.

[KKFK04b] M. Kaeberlein, K. T. Kirkland, S. Fields, and B. K. Kennedy. Sir2-independent life span extension. *PLOS BIOLOGY*, 2:e296, 2004.

[KPS+05] M. Kaeberlein, R. W. Powers, K. K. Steffen, E. A. Westman, D. Hu, N. Dang, E. O. Kerr, K. T. Kirkland, S. Fields, and B. K. Kennedy. Regulation of yeast replicative life span by tor and sch9 in response to nutrients. *Science*, 310:1193–1196, 2005.

[KSH+06] M. Kaeberlein, K. K. Steffen, D. Hu, N. Dang, E. O. Kerr, M. Tsuchiya, S. Fields, and B. K. Kennedy. Comment on "hst2 mediates sir2-independent life-span extension by calorie restriction". *Science*, 312:1312, 2006.

[KSK05] B. K. Kennedy, E. D. Smith, and M. Kaeberlein. The enigmatic role of sir2 in aging. *Cell*, 123:548–550, 2005.

[KWT+02] T. Kohler, S. Wesche, N. Taheri, G. H. Braus, and H. U. Mosch. Dual role of the saccharomyces cerevisiae tea/atts family transcription factor tec1p in regulation of gene expression and cellular development. *Eukaryot. Cell*, 1:673–686, 2002.

[LAGG+05] S. López-Avilés, M. Grande, M. González, A. Helgesen, V. Alemany, M. Sanchez-Piris, O. Bachs, J. B. A. Millar, and R. Aligue. Inactivation of the cdc25 phosphatase by the stress-activated srk1 kinase in fission yeast. *Mol Cell*, 17:49–59, 2005.

[LDB+96] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Want, M. Kobayashi, H. Horton, and E. L. Brown. DNA expression monitoring by hybridization of high density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[LDG00] S. J. Lin, P. A. Defossez, and L. Guarente. Requirement of nad and sir2 for life-span extension by calorie restriction restriction in saccharomyces cerevisiae. *Science*, 289:2062–2063, 2000.

[LF03] V. D. Longo and C. E. Finch. Evolutionary medicine: from dwarf model systems to healthy centenarians. *Science*, 299:1342–1346, 2003.

[LGV96] V. D. Longo, E. B. Gralla, and J. S. Valentine. Superoxide dismutase activity is essential for stationary phase survival in saccharomyces cerevisiae mitochondrial production of toxic oxygen species in vivo. *J Biol Chem*, 271:12275–12280, 1996.

[LH98] M. C. Lorenz and J. Heitman. The mep2 ammonium permease regulates pseudohyphal differentiation in saccharomyces cerevisiae. 17:1236–1247, 1998.

[Li03] L. M. Li. Blind inversion needs distribution (BIND): the general notion and case studies. *Science and Statistics: A Festschrift for Terry Speed*, 40:273–293, 2003.

[Li04] L. M. Li. An algorithm for computing exact least trimmed squares estimate of simple linear regression with constraints. *Computational Statistics and Data Analysis*, 48:717–734, 2004.

[LL03] Y. Luan and H. Li. Clustering time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics*, 19:474–482, 2003.

[LLEM+05] D. W. Lamming, M. Latorre-Esteves, O. Medvedik, S. N. Wong, F. A. Tsang, C. Wang, S. J. Lin, and D. A. Sinclair. Hst2 mediates sir2-independent life-span extension by calorie restriction. *Science*, 309:1861–1864, 2005.

[LLEM+06] D. W. Lamming, M. Latorre-Esteves, O. Medvedik, S. N. Wong, F. A. Tsang, C. Wang, S. J. Lin, and D. A. Sinclair. Response to comment on "hst2 mediates sir2-independent life-span extension by calorie restriction". *Science*, 312:1312, 2006.

[LLYLT04] F. Li, T. Long, Q. Ouyang Y. Lu, and C. Tang. The yeast cell-cycle network is robustly designed. *Proc Natl Acad Sci U S A*, 101:4781–4786, 2004.

[LMS05] V. D. Longo, J. Mitteldorf, and V. P. Skulachev. Programmed and altruistic ageing. *Nat Rev Genet*, 6:866–872, 2005.

[Lon03] V. Longo. The ras and sch9 pathways regulate stress resistance and longevity. *Exp Gerontol*, 38:807–811, 2003.

[LR91] P. Laurenson and J. Rine. Sum1-1: a suppressor of silencing defects in saccharomyces cerevisiae. *Genetics*, 129:685–696, 1991.

[LRR+02] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, and et al. I. Simon. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[LW01a] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98:31–36, 2001.

[LW01b] C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error applications. *Genome Biology*, 2:1–11, 2001.

[MA04] J. Moffat and B. Andrews. Late-g1 cyclin-cdk activity is essential for control of cell morphogenesis in budding yeast. *Nat. Cell. Biol*, 6:59–66, 2004.

[MKCF95] C. J. Mclnerny, P. J. Kersey, J. Creanor, and P. A. Fantes. Positive and negative roles for cdc10 in cell cycle gene expression. *Nucleic Acids Research*, 23:4761–4768, 1995.

[MKHS03] D. L. Maskell, A. I. Kennedy, J. A. Hodgson, and K. A. Smart. Chronological and replicative lifespan of polyploid saccharomyces cerevisiae (syn. s. pastorianus). *FEMS Yeast Research*, 3:201 209, 2003.

[MMG99] M. Kaeberlein M, M. McVey, and L. Guarente. The sir2/3/4 complex and sir2 alone promote longevity in saccharomyces cerevisiae by two different mechanisms. *Genes Dev*, 13:2570–2580, 1999.

[MMO+94] V. Measday, L. Moore, J. Ogas, M. Tyers, and B. Andrews. The pcl2 (orfd)-pho85 cyclin-dependent kinase complex: a cell cycle regulator in yeast. *Science*, 266:1391–1395, 1994.

[MRM+00] S. Melov, J. Ravenscroft, S. Malik, M. S. Gill, D. W. Walker, P. E. ClaytonAND B. C. Wallace, B. Malfroy, S. R. Doctrow, and G. J. Lithgow. Extension of life-span with superoxide dismutase/catalase mimetics. *Science*, 289:1567–1569, 2000.

[MSH04] D. E. Martin, A. Soulard, and M. N. Hall. Tor regulates ribosomal protein gene expression via pka and the forkhead transcription factor fhl1. *Cell*, 119:969–979, 2004.

[Net] The Comprehensive R Archive Network. http://cran.r-project.org/index.html.

[NRS+02] G. J. Nau, J. F. L. Richmond, A. Schlessinger, E. G. Jennings, E. S. Lander, and R. A. Young. Human macrophage activation programs induced by bacterial pathogens. *Proc Natl Acad Sci U S A*, 99:1503–1508, 2002.

[NS99] A. N. Nguyen and K. Shiozaki. Heat-shock-induced activation of stress map kinase is regulated by threonine- and tyrosine-specific phosphatases. *Genes Dev*, 13:1653–1663, 1999.

[oGG] Kyoto Encyclopedia of Genes and Genomes. http://www.genome.jp/kegg/.

[OJ95] S. Ozcan and M. Johnston. Three different regulatory mechanisms enable yeast hexose transporter (hxt) genes to be induced by different levels of glucose. *Mol Cell Biol*, 15:1564–1572, 1995.

[OJ99] S. Ozcan and M. Johnston. Function and regulation of yeast hexose transporters. *Microbiol Mol Biol Rev*, 63:554–569, 1999.

[Ont] Gene Ontology. http://www.geneontology.org/.

[Par01] L. Partridge. Evolutionary theories of ageing applied to long-lived organisms. *Exp Gerontol*, 36:641–650, 2001.

[PDC+03] I. Pedruzzi, F. Dubouloz, E. Cameroni, V. Wanke, J. Roosen, J. Winderickx, , and C. De Virgilio. Tor and pka signaling pathways converge on the protein kinase rim15 to control entry into g0. *Mol Cell*, 12:1607–1613, 2003.

[PDG99] P. U. Park, P. A. Defossez, and L. Guarente. Effects of mutations in dna repair genes on formation of ribosomal dna circles and life span in saccharomyces cerevisiae. *Mol Cell Biol*, 19:3848–3856, 1999.

[PED+98] T. L. Parkes, A. J. Elia, D. Dickinson, A. J. Hilliker, J. P. Phillips, and G. L. Boulianne. Extension of drosophila lifespan by overexpression of human sod1 in motorneurons. *Nature Genetics*, 19:171–174, 1998.

[Pet]

[PJB+02] P. W. Piper, G. W. Jones, D. Bringloe, N. Harris, M. MacLean, and M. Mollapour. The shortened replicative life span of prohibitin mutants of yeast appears to be due to defective mitochondrial segregation in old mother cells. *Aging Cell*, 1:149–157, 2002.

[PLL+03] S. D. Peddada, E. K. Lobenhofer, L. Li, C. A. Afshari, C. R. Weinberg, and D. M. Umbach. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. 19:834–841, 2003.

[Pro] Yeast Ageing Project. http://leili-lab.cmb.usc.edu:9673/yeastaging.

[PYL+03] T. Park, S. G. Yi, S. Lee, S. Y. Lee, D. H. Yoo, J. I. Ahn, and Y. S. Lee. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, 19:694–703, 2003.

[REM+05] J. Roosen, K. Engelen, K. Marchal, J. Mathys, G. Griffioen, E. Cameroni, J. M. Thevelein, C. De Virgilio, B. De Moor, and J. Winderickx. Pka and sch9 control a molecular switch important for the proper adaptation to nutrient availability. *Molecular Microbiology*, 55:862–880., 2005.

[RH04] B. Rogina and S. L. Helfand. Sir2 mediates longevity in the fly through a pathway related to calorie restriction. *Proc Natl Acad Sci U S A*, 101:15998–16003, 2004.

[RJ03] S. Rea and T. E. Johnson. A metabolic model for life span determination in caenorhabditis elegans. *Dev. Cell*, d:197203, 2003.

[RL87] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

[RNM+00] C. J. Roberts, B. Nelson, M. J. Marton, R. Stoughton, M. R. Meyer, H. A. Bennett, Y. D. He, H. Dai, W. L. Walker, and T. R. Hughes et al. Signaling and circuitry of multiple mapk pathways revealed by a matrix of global gene expression profiles. *Science*, 287:873–880, 2000.

[SFKR04] E. Segal, N. Friedman, D. Koller, and A. Regev. A module map showing conditional activity of expression modules in cancer. *Nat Genet*, 36:1090–1098, 2004.

[SLEW01] E. E. Schadt, C. Li, B. Ellis, and W. H. Wong. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, 37:120–125, 2001.

[SMG97] D. A. Sinclair, K. Mills, and L. Guarente. Accelerate ageing and nucleolar fragmentation in yeast sgs1 mutants. *Science*, 277:1313–1316, 1997.

[SP95] K. Shiozaki and P.Russell. Cell-cycle control linked to extracellular environment by map kinase pathway in fission yeast. *Nature*, 378:739–743, 1995.

[SSDB95] M. Schena, D. Shalon, R. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.

[SSR+03] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34:166–176, 2003.

[SSZ+98] P. T. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297, 1998.

[ST03] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100:9440–9445, 2003.

[STC+02] D. A. Smith, W. M. Toone, D. Chen, J. Bahler, N. Jones, B. A. Morgan, and J. Quinn. The srk1 protein kinase is a target for the sty1 stress-activated mapk in fission yeast. 277:33411–33421, 2002.

[STM+05] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102:15278–15279, 2005.

[SW96] R. S. Sohal and R. Weindruch. Oxidative stress, caloric restriction, and aging. *Science*, 273:59–63, 1996.

[SXL+05] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A*, 102:12837–12842, 2005.

[TOO+92] K. Tanaka, K. Okizaki, N. Okizaki, T. Ueda, A. Sugiyama, H. Nojima, and H. Okayama. A new cdc gene required for s phase entry of *Schizosaccharomyces pombe* encodes a protein similar to cdc10+ and swi4 gene products. 11:4923–4832, 1992.

[TOR+01] G. C. Tseng, M. Oh, L. Rohlin, J. C. Liao, , and W. H. Wong. Issues in cDNA microarray analysis: quality filtering,channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, 29:2549–2557, 2001.

[TpS06] Y. C. Ta and T. p. Speed. A multivariate empirical bayes statistic for replicated microarray time course data. *Annals of Statistics*, 34:5, 2006.

[TTC01] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98:5116–5121, 2001.

[vHL02] S. van Huffel and P. Lemmerling. *Total least squares and errors-in-variables modeling: Analysis, Algorithms and Applications*. Kuuwer Academic Publishers, Dordrecht, the Netherlands, 2002.

[vKv+03] J. van de Peppel, P. Kemmeren, H. van Bakel, M. Radonjic, D. van Leenen, and F. C. P. Holstege. Monitoring global messenger rna changes in externally controlled microarray experiments. *EMBO reports*, 4:387–393, 2003.

[vNSH03] V. van Noort, B. Snel, and M. A. Huynen. Predicting gene function by conserved co-expression. *Trends Genet*, 19:238–242, 2003.

[vNSH04] V. van Noort, B. Snel, and M. Huynen. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. 5:280–284, 2004.

[WD04] Y. Wang and H. G. Dohlman. Pheromone signaling mechanisms in yeast: a prototypical sex machine. 306:1508–1509, 2004.

[WI04] Z. Wu and R. A. Irizarry. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *RECOMB*, pages 98–106, 2004.

[WJJ+02] C. Workman, J. Jensen, H. Jarmer, R. Berka, L. Gautier, B. Nielsen, H. Saxild, C. Nielson, S. Brunak, and S. Knudsen. new nonlinear normalization method for reducing variability in DNA microarray experiments. *Genome Biology*, 3:1–16, 2002.

[WKS04] S. Wichert, K.Fokianos, and K. Strimmer. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20:5–20, 2004.

[WSA+99] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, and H. Bussey et al. Functional characterization of the saccharomyces cerevisiae genome by gene deletion and parallel analysis. *Scicence*, 285:901–906, 1999.

[WSDJ99] S. Whitehall, P. Stacey, K. Dawson, and N. Jones. Cell cycle-regulated transcription in fission yeast: Cdc10-res protein interactions during the cell cycle and domains required for regulated transcription. *Mol Biol Cell*, 10:3705–3715, 1999.

[XOZ02] X. L. Xu, J. M. Olson, and L. P. Zhao. A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible huntington's disease transgenic model. *Hum. Mol. Genet.*, 11:1977–1985, 2002.

[XPGD+99] J. Xie, M. Pierce, V. Gailus-Durner, M. Wagner, E. Winter, and A. K. Vershon. Sum1 and hst1 repress middle sporulation-specific gene expression during mitosis in saccharomyces cerevisiae. *EMBO J*, 18:6448–6452, 1999.

[YDFQ05] X. Yan, M. Deng, W. K. Fung, and M. Qian. Detecting differentially expressed genes by relative entropy. *J Theor Biol*, 234:395–402, 2005.

[YDL+02] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30:e15, 2002.

[YSG+02] H. Yoshimoto, K. Saltsman, A. P. Gasch, H. X. Li, N. Ogawa, D. Botstein, P. O. Brown, and M. S. Cyert. Genome-wide analysis of gene expression regulated by the calcineurin/crz1p signaling pathway in saccharomyces cerevisiae. 277:31079–31088, 2002.

[ZAZL01] A. Zien, T. Aigner, R. Zimmer, and T. Lengauer. Centralization: a new method for the normalization of gene expression data. *Bioinformatics*, 17:323–331, 2001.

[ZKH+05] X. J. Zhou, M. C. Kao, H. Huang, A. Wong, J. Nunez-Iglesias, M. Primig, O. M. Aparicio, C. E. Finch, T. E. Morgan, and W. H. Wong. Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol*, 23:238–243, 2005.

[ZKM+05] S. O. Zakharkin, K. Kim, T. Mehta, L. Chen L, S. Barnes, K. E. Scheirer, R.S.Parrish, D. B. Allison, and G. P. Page. Sources of variation in affymetrix microarray experiments. *BMC Bioinformatics*, 6:214, 2005.

[ZMC05] S. A. Zurita-Martinez and M. E. Cardenas. Tor and cyclic amp-protein kinase a: two parallel pathways regulating expression of genes required for cell growth. *eukaryot cell*, 4:63–71, 2005.

[ZSV+00] G. Zhu, P. T. Spellman, T. Volpe, P. O. Brown, D. Botstein, T. N. Davis, and B. Futcher. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, 406:90–94, 2000.