Inferring Transcriptional and Post-Transcriptional Network Structure by Exploiting Natural Sequence Variation

Mina Fazlollahi

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Inferring Transcriptional and Post-Transcriptional Network Structure by Exploiting Natural Sequence Variation

Mina Fazlollahi

Understanding how cellular processes of an organism translate its genome into its phenotype is one of the grand challenges in biology. Linkage studies seek to identify allelic variants that manifest themselves as phenotypic variation between individuals in a population. The advent of high-throughput genotyping and gene expression profiling technologies has made it possible to use messenger RNA levels as quantitative traits in linkage studies. This has created new opportunities to study genetic variation at the level of gene regulatory networks rather than individual genes.

This thesis consists of four parts, each of which outlines a different strategy for integrating genome-wide expression data and genotype data in order to identify transcriptional and post-transcriptional regulatory mechanisms. The data for these analyses comes from segregating populations of *Saccharomyces cerevisiae* (baker's yeast) as well as *Caenorhabditis elegans* (roundworm).

The first study focused on inferring the *in vitro* binding specificity of RNA-binding proteins (RBPs). We first analyzed a recent compendium of *in vivo* mRNA binding data to model the sequence specificity of 45 yeast RBPs in the form of a position-specific affinity matrix (PSAM). We were able to recover known consensus nucleotide sequences for 12 RBPs and discovered novel binding preferences for 3 of the RBPs namely, Scp160p, Sik1p and Tdh3p.

The second study aimed to identify transacting chromosomal loci that regulate expression of a large number of genes. Traditionally, such loci are discovered by first mapping expression quantitative loci (eQTLs) for individual genes, and then looking for so-called "eQTLs hotspots". Our method avoids the first step by integrating

information across all genes, leading to a more elegant method that has increased statistical power. For yeast, we recovered 70% of the reported eQTL hotspots from two independent studies, and discovered a new transacting locus on chromosome V. For worm, we detected six transacting loci, only two of which were previously reported as eQTL hotspots.

The third study focused on post-transcriptional regulatory networks in yeast, by mapping the regulatory activity level of RNA binding proteins (RBPs) as a quantitative trait in so-called "aQTL" analysis. We used the collection of 15 sequence motifs with the associated mRNA region combinations that we obtained in our first study together with mRNA expression data to estimate RBP activities across yeast segregants. Consistent with a previous study, we recovered the MKT1 locus on chromosome XIV as a genetic modulator of Puf3p activity. We also discovered that Puf3p activity is modulated through distinct loci depending on whether it is binding to $5'$ or $3'$ untranslated region (UTR) of its target mRNAs. Furthermore, we identified a locus on chromosome XV that includes the IRA2 gene as a putative aQTL for Puf4p; this prediction was validated using expression data for an IRA2 allele replacement strain.

Our fourth study focused on the detection of loci whose allelic variation modulates the *in vivo* regulatory connectivity between a transcription factor and its target genes. We call these loci connectivity QTLs or "cQTLs". We mapped the DIG2 locus on chromosome IV as a cQTL for the transcription factor Ste12p. Dig2p is indeed a known inhibitor of yeast mating response activator Ste12p. The coding region of the DIG2 gene contains a single non-synonymous mutation (T83I). We are experimentally testing the functional impact of this mutation in allele replacement strains. We also identified the TAF13 locus as a putative modulator of GCN4p connectivity.

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

First I would like to thank my advisor, Dr. Harmen Bussemaker, who has been an exceptional mentor. Harmen's enthusiasm and dedication to research have both been truly inspiring. I specially thank him for giving me the opportunity to work in his lab, even though I lacked a strong background in molecular biology and genetics. His patience and encouragements were a constant source of motivation for me. In many instances, he taught me how to be objective in scientific reasoning and how to clearly express scientific ideas. Through his involved and hands-on approach to programming, I have acquired many useful coding habits. Harmen has been remarkably generous with his time and always considerate to every lab member. I am sincerely grateful for all of his help.

I thank Szabolcs Marka for accepting to sponsor my thesis in the physics department. I also thank him and the other members of my thesis committee: Robert Mawhinney, Lam Hui, Chris Wiggins and Harmen Bussemaker for their time and generosity to serve in my defense. I want to also thank past and present staff members of the physics department: Lalla Grimes, Joey Cambareri, Randy Torres, Rasma Medis, Lydia Argote, Michael Adan, Yasmin Yabyabin, all of whom took care of the administrative tasks that makes the path to the Ph.D. less frantic.

I thank my lab colleagues who have greatly helped me, Barrett Foat, who introduced me to the biological concepts underlying my first project, Gabor Halasz and Luke Ward, both of whom patiently answered my many R programming and biology related questions, Ron Tepper, for so often asking questions that forced me to think deeper about my work, Eunjee Lee, for the many useful discussions we had, Pilar Gomez, for our programming discussions and her moral support, Xiang-Jun Lu, for helping me countless times with the MatrixREDUCE software, Todd Riley and Allan Lazarovici, for the helpful discussions, and Ben Snyder, our system administer, who was always quick in resolving any computer-related problems. I have learnt a lot from all of you and it has been a pleasure to work with you.

I thank my colleagues in the physics graduate program, some of whom have become dear friends. Becky Grossman and Tatia Englemore, I thank you for all the nice time we spend together while studying and for the outside department activities. Thanks for being supportive . I also want to thank my friends Eric Vazques and Hui Zhou, with whom I spent many late nights going through our first problem sets. I am so happy to have Niloofar Faghihi as a dear friend in my life. I look forward to many more scientific discussions, and about anything and everything. I want to thanks my Iranian friends at Columbia University, who have made graduate school more enjoyable and reduced my homesickness. I specially want to mention Bahar Moezzi, Fereshte Ghahari and Behnaz Bozorgui. Bahar was my biggest support when I was preparing for the physics qualification exams during the first semester.

I thank my dear parents for their unconditional love, support and guidance. My father's curiosity and enthusiasm have always been inspirational for me. My mother's hardworking character and her encouragements have given me the strength to go forward. My lovely sisters, Ladan and Niloofar, I am happy to have you in my life. Your support and kindness have been invaluable and thinking about you lifts my spirit. My dear Guillaume, your patience, kindness, and love in the ups and downs of graduate school and life have been my anchors. I would not have been able to fulfill this without your support and I am truly thankful. I also want to express my gratitude to Joanne and my aunt, Rafat.

به نام خدا

To my dear parents,

ملیحه مروّجی اصل

محمود فضل اللهی

Introduction

Traditionally, the field of molecular biology has relied on techniques tackling one or few genes at a time. The development of DNA microarray technology in the early 1990's provided the means for measuring expression data of several thousand genes in a single experiment. Since then, many high-throughput screening techniques have been developed. The vast amount of genetic data produced brought forth the need for computationally demanding statistical approaches that extract biological information using data-driven modeling. Many of these statistical models are directly or indirectly based on biophysical models describing the interactions between DNA, RNA and proteins.

The central dogma of molecular biology is illustrated in **Figure 1**. Protein synthesis from DNA involves many regulatory stages governed by different classes of protein complexes. Initiation or inhibition of the transcription of DNA to RNA is conducted by a class of regulatory proteins known as transcription factors (TFs). By binding to a specific DNA sequence proximal to the genes, TFs can recruit or block the binding of transcriptional machinery and thus activate or repress RNA production. This stage is known as transcriptional regulation. As RNA is being synthesized, another class of proteins known as RNA binding proteins (RBPs) controls a further set of processes. These proteins bind to specific sites on RNA located mostly in the untranslated regions (UTRs) to influence many different processes that contribute to the conversion of pre-mRNA to mature messenger RNA (mRNA). These include RNA splicing (i.e. removing segments of the RNA sequence known as introns), localization, stability, and degradation. These processes are called post-transcriptional regulation

since they are conducted on RNA after DNA transcription initiation. Messenger RNA is then transported out of the nucleus for translation to protein by ribosomes.

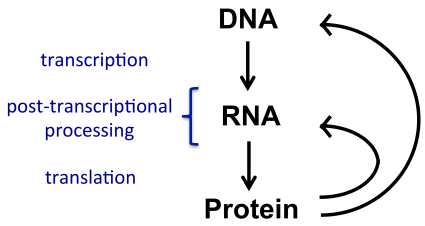

Figure 1: Central Dogma in Molecular Biology. A large class of proteins known as transcription factors (TFs), initiate and regulate gene expression resulting in DNA transcription to RNA. During and after the RNA production another class of proteins, known as RNA binding proteins (RBPs), interact with RNA and modify it into messenger RNA (mRNA). This is known as post-transcriptional regulation. RBPs are also responsible for stabilization or destabilization of mRNAs. The final phase is translation of mRNA to protein via ribosomes. As explained, there are many regulatory controls along the path of protein production from genes.

In a steady-state cellular condition one can use statistical mechanical and thermodynamical approaches to model the binding of TFs and RBPs to their target DNA and RNA, respectively. There has been a strong focus on TF-DNA interaction and much less attention has been channeled to RBP-RNA interaction. This is due to two reasons: Firstly, the significant role of RBPs as an important regulator of the cellular gene expression program was not known. Secondly, unlike the helical structure of DNA, single stranded RNA can fold into complex secondary structure requiring more complex modeling and technologies. Without further regulation after transcription initiation, one would expect a strong correlation between mRNA levels and protein abundances. However thorough measurements have shown more than 20-fold variation between specific mRNAs and their encoded proteins, suggesting post-transcriptional regulation plays a critical role (Gygi *et al.*, 1999). Also, an increasing number of studies are confirming the involvement of post-transcriptional regulation by RBPs in human genetic disorders (Cooper *et al.*, 2009; Joshi *et al.*, 2012; Lukong *et al.*, 2008; Polymenidou *et al.*, 2012; Sterne-Weller *et al.*, 2011; Yamazaki *et al.*, 2012). All these studies point to the fact that dissecting the post-transcriptional network is crucial to understanding how the cell orchestrates gene expression regulation.

Microarray technology has also been used to compare the genomic sequence of a specific strain to a reference sequence (i.e. genotyping) (Sapolsky *et al.*; Winzeler *et al.*, 1998). This has e.g. allowed detection of several thousands of single nucleotide polymorphisms (SNPs) between the DNA sequences of the two yeast strains used to generated the data analyzed in this thesis. Recent deep-sequencing approaches are capable of detecting SNPs location at much higher resolution for many individuals in parallel (Mortazavi *et al.*, 2008; Otero *et al.*, 2010; Shendure and Ji, 2008; Swan *et al.*, 2002). High-resolution genotype maps make it possible to link an observed quantifiable variable (i.e. quantitative trait) to genotype variation in a population. This is the main goal in genome-wide association studies (GWAS) and linkage analysis approaches. GWAS and linkage analysis are applicable to any organism but they are of high value for human disease studies because they incorporate the only available strategy to perturb the human genome, viz. natural genetic variation in human population.

This thesis is organized into six chapters as follows:

Chapter 1 provides background on our current understanding of how gene expression is regulated in eukaryotic cells, focusing mainly on transcriptional and post-transcriptional regulation. It also briefly introduces the experimental techniques employed to obtain the various data sets used in the analyses presented in this thesis. The techniques include mRNA expression level profiling with microarrays, mRNA-protein binding measurements, and protein-protein interaction assays.

Chapter 2 gives a summary of existing motif discovery approaches to determine RNA-binding proteins (RBPs) binding preferences. It then explains our method for modeling RBP binding preferences in the form of position-specific affinity matrix (PSAM) from genome-wide *in vivo* RBP binding data. We applied our motif finding algorithm to a collection of binding data for 45 RBPs and compared our findings to the existing RBPs binding preferences.

Chapter 3 describes the linkage analysis approach and summarizes standard expression QTL (eQTL) methods. It then introduces a novel method for the detection of

transacting loci applied to two model organisms: *Saccharomyces cerevisiae* (baker's yeast) and *Caenorhabditis elegans* (roundworm). The loci we detect modulate expression levels of a large number of genes. Our method uses χ^2-statistics in a novel way to detect transacting loci that have a broad impact on genome-wide mRNA expression levels.

Chapter 4 explains our approach for detection of genetic loci that modulate the activity of RBPs (aQTLs). We used the inferred binding preferences from Chapter 2 to map genetic loci (aQTLs) that modulate RBP activity.

Chapter 5 describes our connectivity QTL (cQTL) analysis method, which identifies genetic loci that modulate the global pattern of regulatory influence of a transcription factor on its target genes.

Chapter 6 summarizes our findings and discusses possible further directions for the studies discussed in this thesis.

Chapter 1

Background

Ever since the discovery of the molecular structure of DNA by James D. Watson and Francis H. Crick in 1953, the fascination to decode the genome has never ceased. Besides opening a horizon for curing genetic disorders and diseases such as cancer and developing personalized medicine, the desire to shed light on the harmonious and auto-regulatory control in a living cell has fueled this drive. The regulatory control of the cell in response to internal and external stimuli is not static and unidirectional; Rather, the products of the genetic code (proteins) dynamically interact with the genome and significantly affect the outcome. It is important to understand the interactions among different elements and organelles in the cell to comprehend how it can orchestrate various functions.

The majority of the work in this thesis applies to yeast, still the methodology is applicable to other organism. The baker's yeast, *Saccharomyces cerevisiae* has served as an important model for eukaryotic organisms. Yeast was the first eukaryote to have its genome completely sequenced (Goffeau *et al.*, 1996). Genetic manipulation such as genetic mutations or deletions of yeast is easy and cheap. Even though there are few aspects of gene regulation that are exclusive to higher eukaryote organisms, most of the fundamental genetic regulatory machinery has been conserved from yeast to human. Perhaps the most valuable insight from performing genetic research on yeast

is achieved by homologous comparison between other eukaryotes. Homologous proteins have some degrees of sequence similarity between different species and they are thought to possess a common evolutionary origin (Reeck *et al.*, 1987). The homology existing between the proteins among different organisms implies the conserved functional roles (Tatusov *et al.*, 1997). More than 30% of all the protein encoding genes of yeast are found to have homology to mammalian proteins (Botstein *et al.*, 1997). We can gain insight about the function of novel proteins by identifying their homologous proteins in yeast. All these applications highlight the importance of genetic research using yeast.

In the first half of this chapter we will discuss cellular structure and interactions from a genetic point of view for eukaryote organisms and in the second half we will explain various experimental techniques that has been developed to understand these interactions both qualitatively and quantitatively.

1.1 Genetic Code

Genes are discrete units through which the genetic information is passed from the parents to offspring. These genes are positioned linearly along structures called chromosomes in the cell nucleus in eukaryote organisms. Each organism has a different number of chromosomes. For example the common fruit fly has 8 and humans have 46 chromosomes. The chromosomes of a particular organism make up the genome of that organism. The genome is made of deoxyribonucleic acid (DNA), which consists of long chain of units called nucleotides. Each nucleotide has three components, a deoxyribose sugar, a phosphate group (PO_4^{-3}) and a nucleobase. There are four type of nucleobases, adenine (A), cytosine (C), guanine (G) and thymine (T). DNA is formed by nucleobases attached together covalently by a backbone made of alternating deoxyribose sugars and phosphate groups. The phosphate group has a net negative charge by exchanging a proton (H^+) to a water molecule present in the nucleus. Each sugar molecule contains five carbon atoms labeled from $1'$ to $5'$ as follows, $C^{1'}$ is

linked to the nucleobase, $C^{3'}$ is attached to the phosphate group of next nucleotide and $C^{5'}$ is bound to the phosphate group of the nucleotide itself. DNA strands have directionality, meaning that for protein synthesis purposes the strands are read from $5'$ to $3'$ direction. This also implies that all genes on a single DNA strand lie from $5'$ to $3'$ direction.

DNA is double stranded in its natural state and the two strands are wrapped around each other in an antiparallel direction to form a double helix structure. The two strands are attached by hydrogen bounds that exist between the complementary bases opposite each other on the two strands, A-T and C-G, known as Watson-Crick base pairings (Watson and Crick, 1953). The DNA double helix structure is displayed in **Figure 1.1**. Because the distance between the two strand and each turn of the helical structure are not equal, the two grooves developed between the strands as shown in the figure are not equal. The major and minor grooves are recognized by DNA binding proteins (Dervan and Burli, 1999; Gao *et al.*, 1992; Mamoon *et al.*, 2002; Singh and Lambowitz, 2001).

Each DNA molecule together with proteins form a highly organized and compact structure within the nucleus called chromatin. **Figure 1.2** shows this multi level and dense structure. Double helix DNA is wrapped around a bead that consists of 2 copies of each of the core histones H2A, H2B, H3, and H4 proteins to make a histone octamer complex (Kornberg and Thomas, 1974). The basic organizational unit of chromatin is called nucleosome. It is 12 nm in diameter. Nucleosomes consist of 147 base pair of double helix DNA coiled around the histone core about 1.67 turns. Neighboring nucleosomes are about 60 base pair apart. In 1928, Emil Heitz published his findings on chromatin taking different forms by observing that part of the chromosomal material of moss stayed compact throughout the cell cycle (Heitz, 1928). He named the condensed part heterochromatin and named the part that decondensed at times during the cell cycle euchromatin.

It is interesting to know that the intragenic and intergenic non-protein-coding sequences make up almost about 98% of human genome (Taft *et al.*, 2007). These

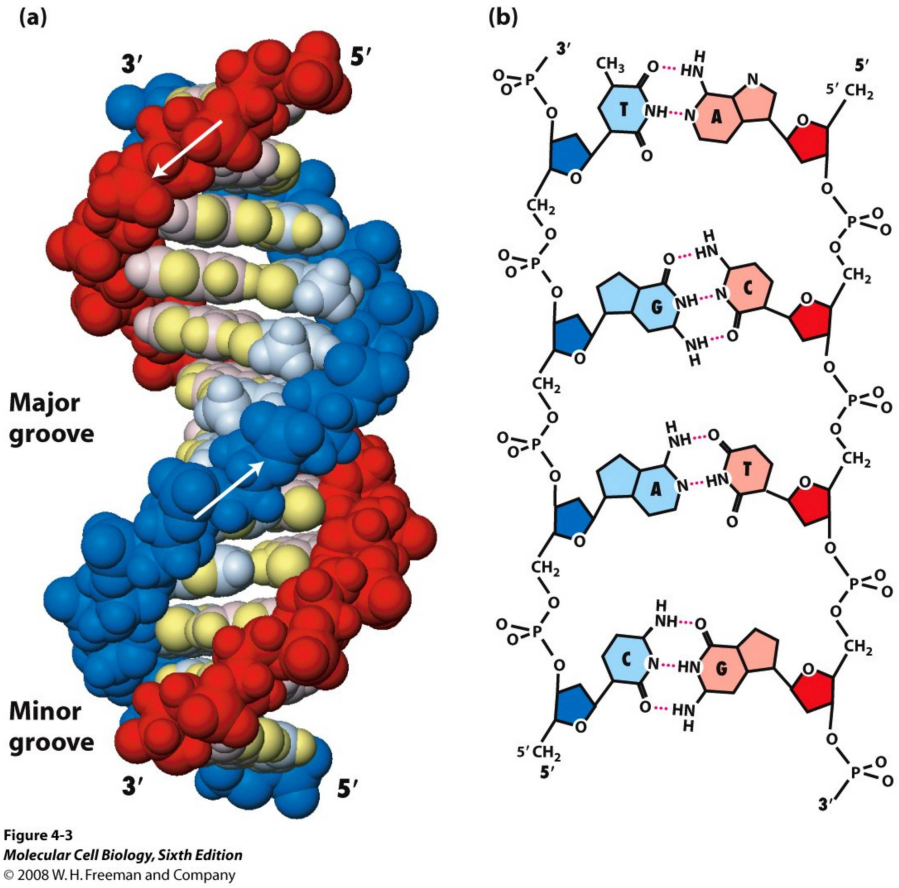

Figure 1.1: DNA Double Helix Structure. The two strands are attached together by the hydrogen bonds between the A-T and C-G nucleobases. The base pairing is known as Watson-Crick base pairing. The nucleobases on each DNA strand are attached together by a backbone made of altering deoxyribose sugar and phosphate. The five carbon atoms on the sugar base are labeled from $1'$ to $5'$. The phosphate group in a nucleotide is bound to the $C^{5'}$ and $C^{3'}$ of each sugar is connected to the phosphate group of the neighboring nucleotide. Two antiparallel DNA strands wind around each other and create a double helix structure. The major and minor grooves, which are formed because of the helical structure, serve as binding platform for DNA-Protein interactions. The double helix structure and the major and minor grooves are shown in (a). A-T and C-G interactions through Watson-Crick edge of the nucleobases by hydrogen bond contacts are presented in (b). Figure from Lodish *et al.* (2007).

regions are referred to "genetic dark matter" or "junk DNA". Increasing evidence points to the important regulatory roles of these regions, which are often linked to human complex disease (Melhem and Devlin, 2010). In addition, it is now known that genes that encode proteins are located in the euchromatin, which is more accessible

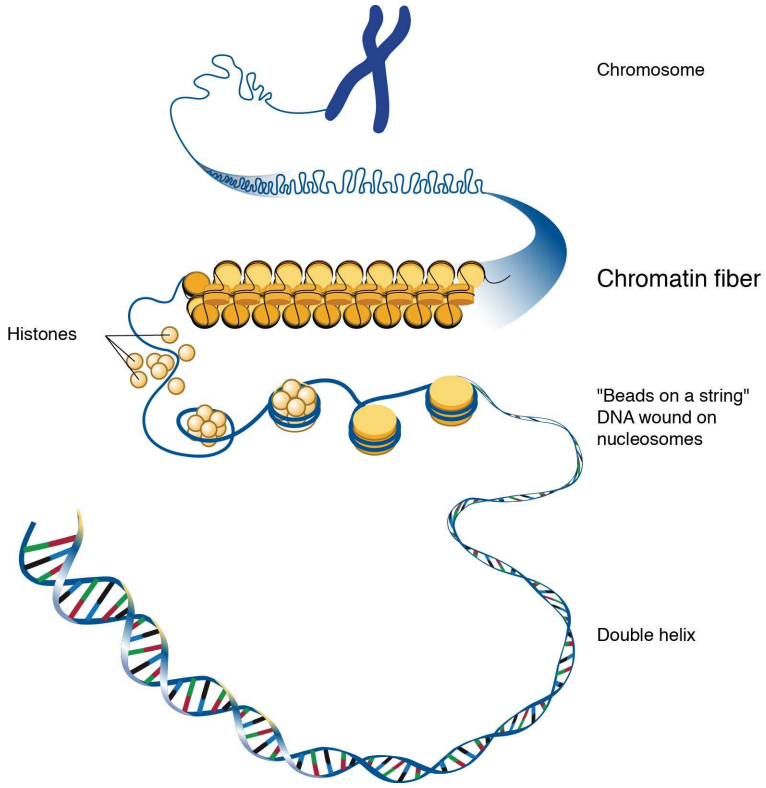

Figure 1.2: From DNA to Chromosome Structure. Double helix of DNA is wrapped around beads made of eight histone proteins, one pair of the each of the four type histones. About 147 base pair of DNA coiled around the histone core create nucleosome, the building block of chromatin. This form of chromatin is known as euchromatin. Chromatin can also acquire an even more condensed form named heterochromatin. During the cell devision heterochromatin is highly condensed with the help of scaffolding proteins to create the chromosomal structure. Figure by Darryl Leja/National Human Genome Research Institute.

for proteins regulating gene expression. In other words, genes are not distributed evenly throughout the chromatin but instead there are regions of high gene density interspersed with depleted region. Prior to cell division, the chromatin forms an even more condensed structure known as chromosome.

1.2 Eukaryotic Gene Expression Model

Gene expression is the process in which a ribonucleic acid (RNA) molecule is synthesized from a specific region of the DNA (i.e. gene) to be used for protein production

in cytoplasm. Upregulation and downregulation of expression of a gene refer to the induction or repression of RNA production. RNA, like DNA, is a polymeric molecule in which the deoxyribose sugar and nucleobase thymine (T) of DNA are replaced by the ribose sugar and uracil (U), respectively. Ribose sugar has a hydroxyl group (OH^{-1}) attached to carbon $C^{2'}$ compared to deoxyribose sugar. This makes the RNA molecules chemically more active than DNA and prone to breakdown by water. The fact that RNA is less stable than DNA is also biologically justifiable. Since RNA molecules are constantly produced and recycled after protein synthesis in the cell, a relatively unstable structure is advantageous; whereas, a stable and chemically passive structure is preferred for the DNA molecules. Also instability of RNA molecules allows the cell to adjust itself in short time after abrupt environmental changes when the synthesis of a specific protein must be shut down. We will discuss later how RNA molecules are protected from degradation between their synthesis and protein production.

Now let's study the series of processes that are conducted in the cell to synthesize proteins in response to external stimuli. **Figure 1.4** illustrates the contemporary theory of gene expression. External signals such as variation in the concentration of chemicals and hormones or environmental changes are detected by specific protein structures on the cell's membrane. G-protein-coupled receptors (GPCRs) make up the largest family of transmembrane receptors that activate signal transduction pathway within the cell (Rosenbaum *et al.*, 2009). Upon detection of external signal, GPCRs release signal-dependant subunits into the cytoplasm. The subunits are then identified by specific enzymes such as protein kinase, enzymes that add phosphate groups to some amino acid residues of a substrate protein resulting in its activation. From here on depending on the stimuli and regulatory circuits there are several signaling cascade that result in activation and translocation of an enzyme from cytoplasm to nucleus.

Transcription factors (TFs) are proteins that bind to a specific nucleotide sequence on DNA proximal to a gene and upregulate or downregulate the gene expression.

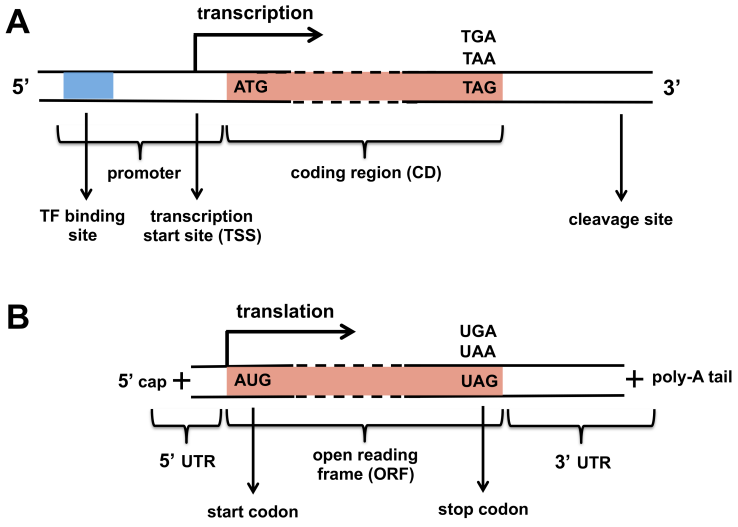

Figure 1.3: Anatomy of a Typical Gene and Synthesized Mature mRNA Molecule. (A) Different segments of a gene are shown. The promoter region contains regulatory elements such as TF binding site and RNA polymerase II transcription initiation site. The segment shown in pink is the coding region (CD) of the gene that will be translated to protein amino acid chain by ribosome. As the transcription machinery continues elongation, the nascent RNA is cleaved off. (B) Different segments of mature messenger RNA (mRNA). The segment upstream of the first AUG motif is the 5′ untranslated region (5′ UTR), which is protected from degradation by a cap attached to it, and the segment downstream of stop codon is the 3′ untranslated region (3′ UTR), which is protected by a poly-adenosine tail. The segment in between the translation start codon (AUG) and the stop codon (either UAG, UAA or UGA) is the open reading frame. This region can be different from the CD of the related gene because of the splicing process (i.e. exons removal), which are not shown here. The ORF is translated to amino acids during protein synthesis.

This region, which contain regulatory sites both for induction or repression of gene expression, is called promoter region (see **Figure 1.3A**). Here the TF and its cofactors initiate the transcription process by recruiting RNA polymerase II complex to the transcription start site (TSS) to the promoter region of the gene. TSS location is usually at least several hundred bases upstream of the beginning of the coding region (CD) of the gene.

The promoter region of many genes is partially obstructed by nucleosomes that inhibit the binding between the TF and the promoter. Several subunits of transcription machinery are responsible for chromatin structure modulation, either by displacing

the histones complex which are ATP-dependant or by post-translational modification of histones (Kouzarides, 2007). For example, the SWI/SNF (SWitch/Sucrose NonFermentable) complex is one of the major ATP-dependent chromatin remodeling composed of 12 subunits (Smith *et al.*, 2003; Szerlong *et al.*, 2003). Prior to transcription initiation of a gene, this complex alters the position of nucleoseomes occupying the *cis*-regulatory site of that gene by forming a DNA loop on the nucleosome surface (Zofall *et al.*, 2006) and it functions with DNA bending proteins to enhance proper chromatin architecture. The second class of chromatin modifier chemically alters the histones amino acid tail that are wrapped around the DNA tightly. This modification such as acetylation of histone tails is a reversible event by deacetylation required for gene repression (Struhl, 1998; Wolffe, 1996). Both of the chromatin remodeling processes decompact the chromatin allowing access to the promoter region of the gene.

As the nascent RNA polymer curls out of the RNA pol II, it gets covered with proteins named RNA binding proteins (RBP) and a series of further processes are initiated. These events are known as post-transcriptional regulation and extend from birth to death of the RNA transcript by various classes of RBPs (Moore, 2005). First a 7-methylguanosine cap is attach to the beginning of the RNA (see **Figure 1.3** and **Figure 1.4**). Besides protecting the RNA molecule from degradation, the cap is also important for efficient protein synthesis from the RNA template (Fechter and Brownlee, 2005). mRNA splicing, by which some specific segments of the pre-messenger RNA (pre-mRNA) molecule known as introns are cleaved out (i.e. spliced), was first observed by Berget *et al.* (1977). The remaining two segment (i.e. exons) after each splicing event are then ligated back together. At first it was thought that each gene codes only for one specific protein amino acid chain. It is now evident that in higher organisms condition-specific combinations of intron splicing can occur. This is known as alternative splicing and as a result a single gene can encode for two or more proteins (Black, 2003). The intron splicing continues as the RNA pol II continues RNA elongation phase, even after reaching the end of the coding region of the gene. Once RNA pol II passes and transcribes the polyadenylation signal sequence (AAUAAA)

and cleavage signal, the pre-mRNA is cleaved and the poly-adenosine tail (poly(A) tail) with about 150-300 adenosine residue is added to the end (Mangus *et al.*, 2003). The poly-A tail plays an important role in stability and transportation of the mature mRNA to the cytoplasm where the proteins are synthesized.

Figure 1.3B displays different parts of the mRNA molecule. The main body is comprised of an open reading frame (ORF) in between the 5′ and 3′ untranslated regions (UTRs). As mentioned earlier, the 5′ UTR is attached to a cap and the 3′ UTR is linked to the poly(A) tail. Protein synthesis is carried out by ribosomes with translation starting at the first occurrence of AUG nucleotides triplet. As ribosomes scan the mRNA chain, the nucleotide triplets are translated to amino acids. Amino acids are the building block of proteins. Codons are the sequential non-overlapping nucleotide triplets that specify the type of amino acid for protein synthesis. There are only 20 different types of amino acids found in eukaryote proteins. From the 4 different types of nucleotides existing in mRNA sequence (A, C, G and U), 64 unique codon can be generated. This means the codon-to-amino acid mapping is degenerate, with only 20 different amino acids found in natural proteins.

Ribosomes begin translation of mRNA from the start codon (AUG) and continue reading the ORF until reaching one of the stop codons (UAG, UAA or UGA). As their name implies, the 5′ and 3′ UTRs are not translated into amino acids. However the UTRs play a major role in post-transcriptional regulation of mRNA. More specifically, the 3′ UTR is involved in mRNA stability (Conne *et al.*, 2000; Mignone *et al.*, 2002). mRNAs that are actively being translated by ribosome are protected from degradation by protein complex bound to the 5′ UTR cap and 3′ UTR poly(A) tail; Whereas, translationally inactive mRNAs are targeted for decay enzymes.

In the next section we will expand more on the regulatory circuit in the cell by focusing on some well studied examples.

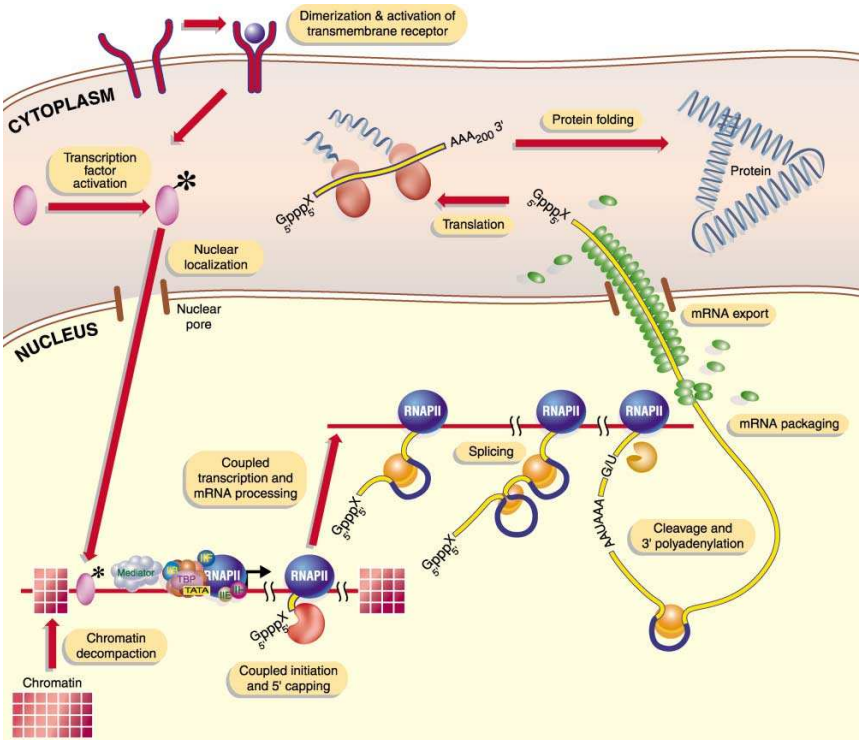

Figure 1.4: A Unified Theory of Gene Expression (see text for details). Figure from Orphanides and Reinberg (2002).

1.3 Regulatory Control Strategies in the Cell

Regulatory control in a cell is a multi-layer process. Regulation in a cell is often triggered by an extra-cellular signal. This triggers a chain of overlapping events within the cytoplasm and nucleus. It starts with post-translational modification of signal transduction protein in the cytoplasm (e.g. protein phosphorylation) carried on into nucleus by activating or repressing transcription through chromatin modification or activating the transcription factors. The next phase in the chain of events is the post-transcriptional regulation through pre-mRNA processing, mRNA transport into cytoplasm, localization, mRNA stability and degradation. The next layer of regulation is conducted during mRNA translation, protein synthesis and post-translational protein modification. The majority of the regulatory mechanisms are conducted through modulation of the activity levels of the protein kinases, transcrip-

tion activators, repressors and RBPs involved in a response pathway of the detected signal. Furthermore, it is now known that in all eukaryote model organisms except *Saccharomyces cerevisiae*, post-transcriptional regulation and more specifically gene silencing is triggered by RNA interference (RNAi) (Drinnenberg *et al.*, 2009; Filipowicz and Sonenberg, 2008; Tomari and Zamore, 2005).

1.3.1 Transcription

The majority of biological regulation occurs at the level of transcription initiation by the transcription activating factors and repressors. They regulate mRNA transcription by assisting or blocking the binding of the transcription machinery to transcription start sites (TSS). In the cellular "ground-state" with respect to transcription, the promoter region of most genes are partially occupied by histones core and transcription is repressed. This results in a very low basal transcription level *in vivo*. Transcription is initiated by sequence-specific recognition of unique upstream activation sequences (UASs) within the promoter regions by TFs (Ptashne, 1988; Struhl, 1998). Presence of different types of UASs in the promoter region of a gene indicates the combinatorial and more complex regulatory network even in the simplest eukaryote organism (Tuch *et al.*, 2008). Transcription repression can happen by recruiting complexes known as transcription repressors to the upstream repression sequences (URSs) on the promoter regions that compete with RNA polymerase II complex (Smith and Johnson, 2000). Repression can also happen through chromatin modification. For example, by recruiting histone deacetylation complexes, which cause localized condensation of the chromatin around the promoter region and thus blocking the binding of transcriptional machinery (Kadosh and Struhl, 1998).

A well-studied example of transcriptional regulation is yeast galactose[1] metabolism. Gal4 protein is a transcription factor that is required for expression of the genes encoding enzymes such as GAL1 that are involved in this pathway (Giniger *et al.*, 1985). Gal4 protein structure consists of two separable domains: a DNA binding

[1]Galactose is a monosaccharide sugar that is very similar to glucose.

domain (BD) and an activating domain (AD). The BD recognizes and binds to the so-called galactose upstream activating sequence (UAS_G) on the DNA sequence and the AD interacts with the RNA polymerase II subunits (Keegan *et al.*, 1986). Both domains are required for transcription activation by Gal4p. In yeast cells that grow in the absence of galactose, GAL genes are expressed at very low levels. Upon addition of galactose to the medium, the expression of GAL genes gets highly induced by Gal4p through active recruitment of RNA polymerase II to the transcription initiation site on the DNA (Gill and Ptashne, 1987). The second regulator of GAL genes is the repressor Mig1p, which is active in the presence of glucose (Griggs and Johnston, 1991; Nehlin *et al.*, 1991). In the absence of both glucose and galactose, Gal4p is bound to the UAS_G through its BD region. However, the AD region of Gal4p is bound by inhibitor Gal80p and Gal4p cannot recruit the RNA pol II complex (Johnston *et al.*, 1987; Ma and Ptashne, 1987). In the presence of galactose but not glucose, cytoplasmic Gal3p can localize the nuclear Gal80p to cytoplasm resulting in rapid dissociation of Gal80p and Gal4p (Jiang *et al.*, 2009). So Gal3p in a way acts as a galactose sensor. Meanwhile, repressor Mig1p is kept in cytoplasm in a phosphorylated state and is unable to interfere with Gal4p activity. When glucose is detected by the cell, the protein kinase that phosphorylates Mig1p is deactivated and dephosphorylated. Mig1p can then localize in the nucleus and can bind to URS sites and repress the expression of GAL genes by recruiting a repressing complex containing Tup1p. It is interesting to know that the URS site for Mig1p binding is also found in the promoter region of the GAL4 gene itself. So Mig1p reduces the levels of the GAL4 mRNA in a glucose-dependent manner as well (Griggs and Johnston, 1991).

Some other mechanism of transcriptional regulation includes TF conformational change during stress response (Eastmond and Nelson, 2006), TF-dependaet chromatin remodeling (Young *et al.*, 2002) and TF activation by intermediate products during amino acid starvation (Wang *et al.*, 1997).

1.3.2 Post-Transcription

Post-transcriptional regulation refers to the set of processes performed by RNA-binding proteins (RBPs) on the precursor mRNA molecule once its transcription is initiated and are continued until the mRNA degradation phase. As shown in **Figure 1.3B**, each mRNA molecule has three regions: the 5′ and 3′ untranslated regions (UTRs) and the open reading frame (ORF). The ORF segment is scanned and read by ribosome and proteins are synthesised by translating the codon information. The two UTRs play an important role in the regulation of mRNA levels. Here we will discuss the post-transcriptional regulatory mechanisms in more depth.

In the traditional view of gene expression, pre-mRNA processing is performed once the transcription is completely finished. The RNA molecule is then cleaved off and splicing, 5′ capping and 3′ poly(A) trail addition is performed. So, transcriptional and post-transcriptional regulations were long viewed as two discrete and independent events. It is now known that some of the pre-mRNA processing take place co-transcriptionally (Orphanides and Reinberg, 2002). That is, while the RNA polymer is being synthesised, the 5′ capping and some splicing are performed (Bentley, 2005). In contrast, the 3′ end poly(A) tail formation is tightly linked to the transcription termination (Buratowski, 2005).

Splicing is catalyzed by the splicosome, a complex consisting of small nuclear ribonucleoprotein particles (snRNPs) (Wahl *et al.*, 2009). First, specific snRNPs of the splicosome complex bind to both ends of the splice site (i.e. intron) and are joined by other snRNPs to loop out the intron segment. The generated stem-loop is then cleaved and the neighboring exons of the spliced site are ligated. In a study by Beyer and Ocheim (1991), electron microscopy revealed looped RNAs attached to chromatin. This observation was among the first experimental evidence of co-transcriptional splicing. However, the introns mostly close to the 3′-end of the mRNA are spliced post-transcriptionally. In fact, some pre-mRNAs in the neuronal cells are spliced after transportation into the cytoplasm by Calcium ion signaling (Glanzer *et al.*, 2005).

The functional consequences of co-transcriptional versus post-transcriptional splicing is still an open question. It could be that the former subjects splicing to transcription-dependent mechanism and the later might link splicing with some downstream regulatory mechanism (Han *et al.*, 2011). Only about 5% of yeast genes are found to contain introns (~287) and all of them are removed before translation (Juneau *et al.*, 2006). The presence of introns in higher eukaryotes, like mice or human, compared to yeast is extremely common. For example, there are about 140,000 introns present in the human genome. The introns cover about 25.9% of the human genome compared to exons, which make up only about 1.5% (Gregory, 2005; Juneau *et al.*, 2006). Due to sparsity and low occurrence of introns in yeast genome, the genetic studies on this organism are in general less concerned about introns and splicing events.

UTR regions of mRNAs play vital roles in the post-transcriptional regulation of gene expression. They are involved in mRNA transport between the nucleus and cytoplasm, subcellular localization, stability and translation efficiency (Mignone *et al.*, 2002). These processes are mainly controlled through interaction of RBPs with specific nucleotide motifs on the mRNA UTR regions and the RNA secondary structure. The average length of 5′ UTRs varies between 100 and 200 nucleotides among various organisms. In contrast, the average 3′ UTRs length is organism-dependent. In fungi this length is about 200 nucleotides and it reaches about 800 nucleotide in humans and other vertebrates (Mignone *et al.*, 2002). As mentioned earlier, the 5′ UTR is important for translational efficiency. Specific cap-binding proteins gather at the 5′-end and any secondary structure that has formed in this region is unfolded. This creates a platform for binding of the ribosome subunit to the mRNA (Maitra *et al.*, 1982). Also, translation can be repressed by some RBPs. For example, the iron-repressive element (IRE) is located in the 5′ UTR of mRNAs that encode proteins required for iron metabolism pathway and effect translation through RBP-IRE binding regulated by intracellular iron levels (Hentze *et al.*, 1987; Leipuviene and Theil, 2007).

Another post-transcriptional regulatory mechanism is the subcellular localization of mRNAs, which results in asymmetric concentration of the synthesized proteins in the

cytoplasm. This regulation is highly important for development. There are several strategies for mRNA localization such as active directed mRNA transport and local stability regulation of the mRNAs in the cytoplasm. Both of these mechanisms are carried out by RBPs interacting with the signal elements located mainly within the 3′ UTRs (Ainger *et al.*, 1997; Bashirullah *et al.*, 2001).

Finally, RBPs also regulate the stability or instability of mRNA transcripts through interaction with specific sites on the mRNA untranslated regions. In eukaryotes, mRNA decay is initiated by shortening of the poly(A) tail at the 3′ end and 5′ cap removal by deadenylase complex and decapping enzymes respectively (Parker and Song, 2004). The mRNA stability regulation is mediated through AU-rich elements mainly in the 3′ UTR region and mRNA decay is initiated by degradation of the poly(A) tail. 5′ UTRs and ORFs may also play a role in mRNA decay by a process known as nonsense-mediated mRNA decay (NMD). When a nonsense codon (i.e. premature stop codon) is followed by junction due to splicing, this type of decay can occur (Hentze and Kulozik, 1999). The spliced regions of the mRNA are detectable because a marker protein binds to the junction at the to end of neighboring exons. This marker protein remains at the junction even after the mRNA transport to the cytoplasm and during protein synthesis (Kataoka *et al.*, 2000). The ribosome complex displaces these marker proteins during mRNA translation. However, if the ribosome disassembles from the mRNA molecule due to a nonsense stop codon, the linker protein initiates the NMD pathway.

mRNA half-lives are measured by chemically arresting the transcription followed by DNA microarray assays over a time course (see **Section 1.5.1**). Such studies performed in yeast revealed that the mRNA encoding metabolic proteins have relatively long half-lives and the mRNA encoding ribosomal proteins are relatively short-lived, which vary from couple of minutes to more than 100 minutes (Wang *et al.*, 2002). mRNA degradation is a highly efficient process in the cell. Since a single mRNA is used as template to synthesize multiple protein by ribosome, fast mRNA degradation is crucial for the cell when that specific protein is not needed any more. This is even

more important in the case of defective mRNAs, which could lead to catastrophic consequences for the cell if that mRNA is not degraded efficiently and quickly (Houseley and Tollervey, 2009).

As an example, let us focus on the yeast protein Puf3. Puf3p is a RNA-binding protein that regulates the stability of its target mRNAs post-transcriptionally (Foat *et al.*, 2005; Gerber *et al.*, 2004; Jackson *et al.*, 2004). It is a member of the Pumilio-FBF (PUF) protein family domain, whose protein members are highly conserved both functionally and structurally (Wang *et al.*, 2002; Wickens *et al.*, 2002). Puf3p is important both for regulation of mRNA degradation and mitochondrial biogenesis especially in respiratory conditions (Jiang *et al.*, 2010). Foat *et al.* computationally and experimentally demonstrated that depending on the type of the carbon source present in yeast growth medium, Puf3p becomes active and destabilizes its target mRNAs by binding to consensus motif located in their $3'$ UTRs. For example, Puf3p represses expression of the COX17 gene through promoting the degradation of its mRNA by deadenylation (Olivas and Parker, 2000). The COX17 gene encodes a protein that is copper metallochaperone and transfers copper ions to a subunit in mitochondrion[2] (Heaton *et al.*, 2000). In the study by Olivas and Parker, the half-lives of the COX17 mRNA between yeast wild-type strain and a strain lacking PUF3 gene (*puf3*Δ) were compared after shutting off the transcription. Measurements of that experiment showed that the half-lives are about 3 minutes and 17 minutes in the wild-type and mutant strains respectively. So the lack of Puf3p resulted in stabilization of the COX17 mRNA by more than 5-folds. The same study also measured deadenylation rate of 17.5 residues/min and 2-3 residues/min for the two strains, again serving as an evidence that Puf3p enhances the deadenylation and degradation of the COX17 mRNA. One possible explanation for this is that Puf3p might recruit deadenylase enzyme to its target mRNA or modify the mRNP structure to a better substrate for the deadenylase (Wickens *et al.*, 2002). It is also possible that Puf3p binding to the $3'$ UTR could accelerate the decapping of the $5'$ UTR and enhance degradation (Houshmandi and Olivas, 2005). Additionally, by some feedback

[2]Mitochondria are the organelles in the cytoplasm that are responsible for ATP production.

signaling mechanism PUF3 gene expression is highly induced during the respiration state (Jiang *et al.*, 2010).

RNA Turnover Rate Formulation

In a living cell, biochemical molecules such as RNA and proteins are constantly produced, utilized and recycled. In response to abrupt changes in the environmental conditions, cell can rapidly response by regulating the state of protein synthesis by adjusting the mRNA turnover rate. The difference between the transcription rate of a gene g and the decay rate of its mRNA sets the turnover rate (Foat *et al.*, 2005).

$$\frac{d}{dt}[\text{mRNA}]_g = \frac{\alpha_g}{V_{\text{cell}}} - \tau_g[\text{mRNA}]_g \tag{1.1}$$

Here $[\text{mRNA}]_g$, α_g and τ_g represent the mRNA concentration, transcription and decay rate of gene g, respectively. The parameter V_{cell} refers to the cell's volume. Note that the concentration is measured per unit volume. We have discussed some transcriptional regulatory mechanisms in the previous section. Here we will focus on the decay rate, which depends inversely on the mRNA stability. During the time that the mRNA is bound to the ribosome and translation is actively carried out, two RNA binding protein complexes attach to either end of the mRNA molecule and stabilize it by protecting the $5'$-end cap and the $3'$-end poly(A) trail from degradation. mRNA decay is induced when particular RBPs bind to a specific binding site located mainly on the $3'$ UTR of the mRNA and promote deadenylation or repress translation and thus enhance the mRNA turnover rate.

In an steady state condition, there is no variation in mRNA_g concentration macroscopically and the right side of the above equation is equal to zero.

$$V_{\text{cell}} \times [\text{mRNA}]_g = \frac{\alpha_g}{\tau_g} \tag{1.2}$$

From the equation above, it is obvious that the mRNA abundances contain both

transcriptional and post-transcriptional regulatory information. So measuring the mRNA abundances in a steady state condition can be used for dissecting both TFs and RBPs regulatory networks.

1.3.3 Translation and Post-Translation

Translation refers to the protein synthesis process during which the mRNA template is scanned by ribosome and the amino acid chain is synthesized based on the sequential codon information of the mRNA open reading frame (ORF). The translational regulation includes processes that control the recruitment of ribosome complex on the start codon, the elongation, and termination of protein synthesis. Like the regulatory mechanisms discussed above, translation is a highly regulated process and is tightly coupled with post-transcriptional regulation. In eukaryotes, translation initiation is in most cases dependent on the $5'$ UTR structure and the presence of the $5'$ cap. The cap-binding protein complexes interact with ribosome through its 40S ribosomal subunit. Also, it has been observed in yeast that decreased translation rate of mRNAs can trigger degradation (Muhlrad and Parker, 1994; Schwartz and Parker, 1999). Also as mentioned previously, mRNAs with iron responsive elements (IREs) can recruit regulatory proteins to inhibit ribosome scanning and repress translation (Leipuviene and Theil, 2007). In addition to the mRNA-specific translation regulation, translation can be regulated through modification of the proteins in the basic translation machinery. For example phosphorylation of the translation initiation factor, eIF4E, is linked to increased activity of the translational machinery (Duncan *et al.*, 2005). Regulation of translation processes is crucial for the cell, especially during development, since it controls the accumulation of the required proteins at the correct location and time within the cell (Vasudevan *et al.*, 2006).

Post-translational regulation, as its name implies, refers to set of processes that modify synthesized proteins and regulate their activities or mark the protein for degradation (Benayoun and Veitia, 2009). It includes protein folding, amino acid editing and adding chemical groups such as methylation (Lee *et al.*, 2005), phosphory-

lation (Fiedler *et al.*), acetylation/deacetylation (Kurdistani and Grunstein, 2003). Many of these modification occur in combination together during any pathway.

Protein methylation is a post-translational modification by which a methyl group (CH_3^{-1}) is added to specific lysine or arginine amino acid residues. This phenomenon was first time observed in bacteria by Ambler and Rees (1959). Lysine methylation has exclusively been seen on histones, but arginine methylation has been detected in various other proteins (Lee *et al.*, 2005). Methylation of several lysine residues in the H3 nucleosome subunit is associated with euchromatin and transcriptional activation, whereas methylation of other residues in H3 and H4 nucleosome subunits is associated with heterochromatin and transcriptional repression (Lee *et al.*, 2005). Transcriptional repression or activation by histone methylation could be due to the inhibition of binding of other proteins to histone tails or generating a binding site for the recruitment of other enzymes and proteins involved in chromatin remodeling processes.

Histone deacetylation is a reversible mechanism underlaying chromatin remodeling. Removal of an acetyl group ($COCH_3^{-1}$) from specific lysine amino acid residues on the histone tails destabilizes the contact between the core histone and DNA and allows for decompactation of the DNA, providing a platform for transcription factors interaction with the DNA. Histone acetylation, which is the addition of an acetyl group to the lysines, reverses this process and leads to compact nucleosome (Wolffe, 1996).

Phosphorylation is the process that modifies the chemical structure of a protein by adding a phosphate group (PO_4^{-3}) to specific serine and threonine amino acids residues. This can alter the protein's activity level. Protein phosphorylation was first reported by Burnett and Kennedy (1954). Protein phosphorylation is another reversible modification, which can set the activity of wide range of kinases on and off and it is the basis of signal transduction. Many important cellular processes are initiated or repressed through a cascade of protein phosphorylation (Fiedler *et al.*). Phosphorylation of TFs or RBPs is a way of sequestering them to outside of the nucleus when not needed during gene expression.

All of these classes of regulatory mechanisms of gene expression are highly interwoven and this regulatory network is more complex in the higher eukaryotes. In the next section we will focus on the protein structure and consider few examples of protein-DNA and protein-RNA interactions.

1.4 How Do Proteins Recognize Their Targets?

This section concentrates on the topic of target recognition by proteins. As was discussed in the previous section, the target set of a protein could be promoter regions on the DNA, RNA transcripts, or other proteins. Transcription factors (TFs) bind to specific sites on the promoter region of the genes and initiate or suppress transcription. Most transcription activators and repressors contain a DNA-binding domain that directly interacts with the DNA. Some factors get recruited to the proper site on the promoters with the assistance of cofactors. DNA-binding domains (DBD) are sorted into different protein families based on the domain structures. RNA-binding proteins (RBPs) bind to RNA transcripts and carry out post-transcriptional processing and mRNA stability regulation. RBPs also contain segments known as RNA-binding domains (RBD), which are categorized into structural domain families. These domains recognize and bind specific sites on the mRNAs. The third group of protein domains, in the case of protein kinases, recognize amino acid site on other proteins. We will mostly elaborate on the TFs and RBPs direct recognition of their targets from structural and chemical point of view with few examples.

To understand the mechanisms of protein-DNA or protein-RNA recognition, it is necessary to first study some common binding domain structures. Proteins re made up of amino acids. Each amino acid consist of a generic part H3NCHRCOOH or peptide, which acts as the back bone, and a unique side chain or residue (R) attached to each peptide. There are 20 unique amino acids found that occur naturally in living organisms. However, due to post-translational modification it is common to find variations of these 20 basic amino acids in the proteins. During the amino acid chain

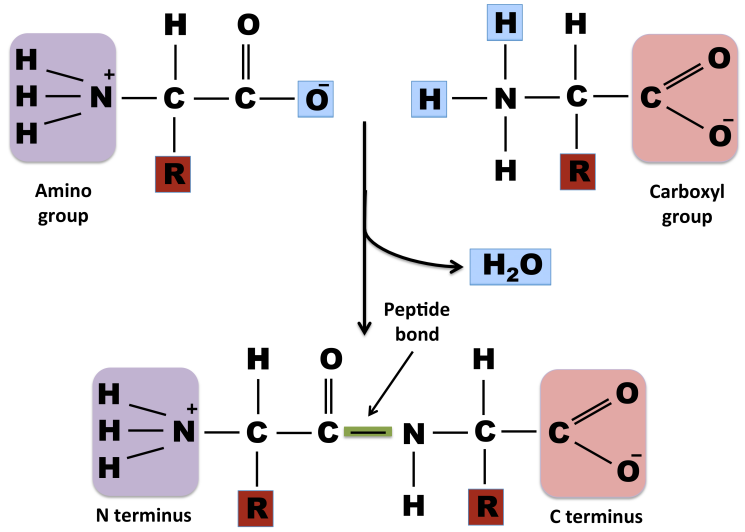

Figure 1.5: Formation of Peptide Bond Between Two Amino Acids. The generic sequence of a free amino acid can be written as H3NCHRCOOH. The carboxyl group with negative charge forms a peptide bond with the positive amino group of the successive amino acid and a water molecule (H2O) is released.

synthesis, the carboxyl group (COOH) of an amino acid reacts with the amino group (H2N) of the following amino acid and they generate a chemical bond by releasing a water molecule (H2O) as shown in **Figure 1.5**. Because the polypeptide chain begins with the amino group of the first residue and it ends with the carboxyl of the last residue, the two ends of the protein are referred to as the N-terminal and C-terminal respectively.

The amino acids' size, hydrogen-bonding potential and net electric charge are determined by the structural composition of the amino acid side chains as shown in **Figure 1.6**. Five amino acids, namely argenine, histidine, lysine, aspartic acid and glutamic acid, have relatively long and flexible charged side chains. The rest of the side chains are electrically neutral. However, six amino acids in this group can participate in hydrogen bonding due to the polarity of the chemical groups at the end of their side chains containing nitrogen, oxygen and phosphorus atoms. The remaining nine amino acids have hydrophobic resides. All of these physical and chemical characteristics of amino acids play a crucial role in protein folding and also recognition of specific nucleotide sequences for protein-target interactions. In the physiological

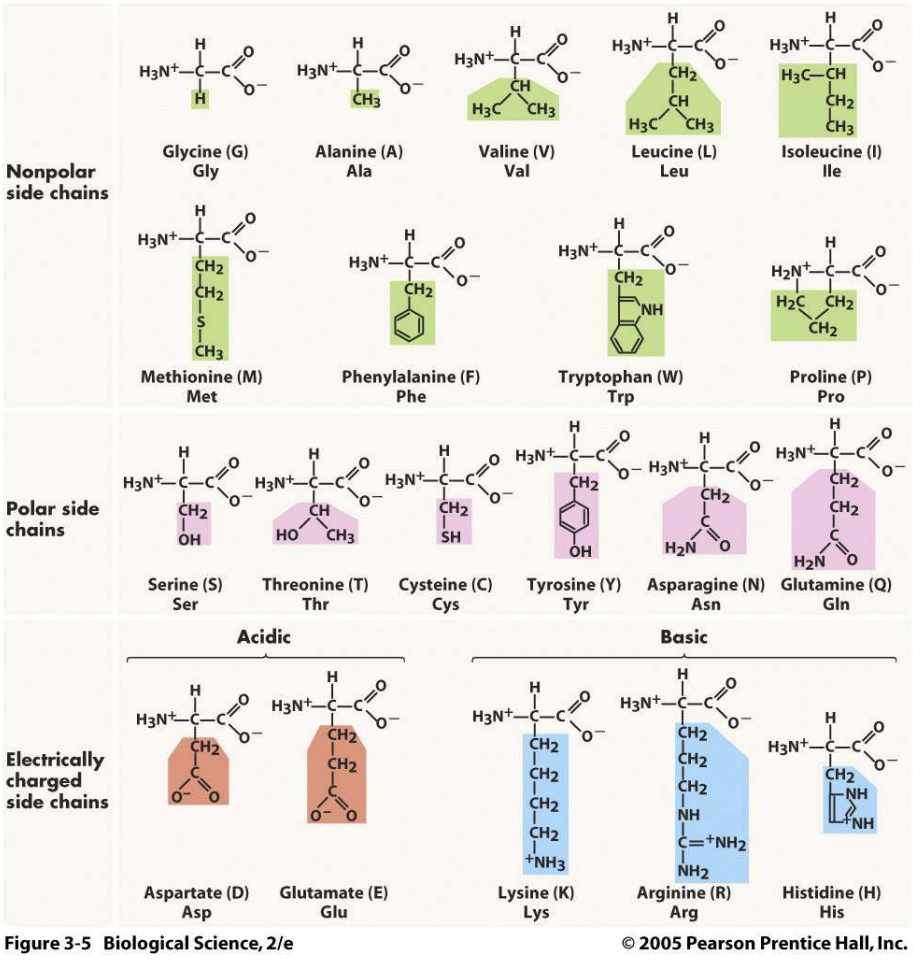

Figure 1.6: Chart Representing the 20 Amino Acids Found in the Proteins of Living Organisms.

state of the cell, e.g. concentration levels of various chemicals and ions, PH level, temperature, and cofactors concentration, the protein structure is dictated by its amino acid sequence (Anfinsen, 1973). The optimal three dimensional protein configuration must be energetically favored such that the newly synthesized polypeptide chain can reach it in a short time. In some cases, the protein folding is guided with molecular chaperones that assist the trapped protein conformations in local free energy minima by an ATP-dependant mechanism (Hartl *et al.*, 2011). The formation of chemical interactions such as covalent bonds and hydrogen bonds between different amino acids, hydrogen bonds between polar amino acids and water molecules and also hydrophobic effect between non polar residues and water molecules, determine

the protein structure and maintain its stability. It is also important to note that these structures are highly sensitive to temperature and chemical composition of the cytoplasm and nucleus. For example, even a slight variation in the temperature of the cell's environment can destabilize and deform protein structures and consequently give rise to the loss of their functionality. If we consider yeast cells in conditions such as heat shock, the stress response pathway gets induced and expression of about 14% of yeast genes gets either activated or repressed (Gasch *et al.*, 2000). Since the heat shock reduces the structural integrity of proteins and promotes protein unfolding, stress response induces expression of genes encoding protein folding chaperones and further localizes them in the cytoplasm and mitochondria (Gasch *et al.*, 2000). So maintaining protein structure is of high concern for the survival of living cells.

In 1913, Nishikawa and Ono observed some ordered molecular structure when studying the x-ray diffraction pattern of silk (Nishikawa and Ono, 1913). This was the first indication for the existence of the protein structures using crystallography method. X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are now the standard approaches for determining protein structures. Two common structural pattern that are observed in proteins are the α-helix and the β-sheet. To generate an α-helix, the amino group of residue i forms a hydrogen bond with the O = C group of residue $i+4$. In almost all protein structures, these helices are right-handed, with the hydrophilic side of the residues facing the exterior and the hydrophobic sides face toward the interior of the helix. A β-sheet is formed from separate polypeptide strands through hydrogen-bonding interactions, which can have parallel or anti-parallel orientation relative to each other.

To form a compact and stable structure, regions of helices and sheets within the protein structure are connected with loops or turns. Protein folding causes distant regions of the protein to interact and produce a stable and functional structure. So the binding-domain of a protein can be made up of distant regions of the protein. **Figure 1.8** displays examples from 5 different DBD families. For example, two α-helices connected by a turn generate the helix-turn-helix (HTH) structure. Proteins

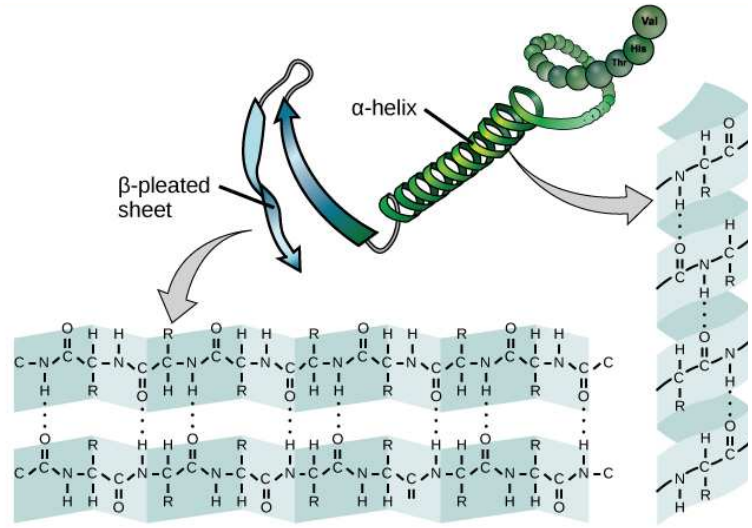

Figure 1.7: Two Local Secondary Structures of Proteins: α-helix and β-sheet. The dotted lines represent hydrogen-bonds. Figure from `http://cnx.org/content/m44402/latest`.

with HTH domain interact with the DNA by the insertion of one of their helices into the major groove of the promoter region of their target genes. Another structural pattern is known as leucine zipper (bZIP) where two parallel α-helices are connected together through leucine residues repeating every seven residues. If each of the helices of the bZIP structure is replaced by two α-helices connected through a single loop, the structure is named as basic helix-loop-helix (bHLH). The two helix-loop-helix halves can be same or different proteins, which are dimerized together. Protein structures can become further stabilized by introduction of other ions such as zinc that can hold the folds together.

In general, the DNA-binding domain (DBD) of the transcription factors can be classified based on the structure of its domain. The main structural families are the zinc stabilized, HTH and zipper type. **Figure 1.8** displays a protein member from five different DNA-binding domain families. The zinc stabilized DBD family contains the largest number of protein with zipper type being the second largest group. The residues in the DBD of the TF interact with the nucleotides and also through the minor and major grooves of the DNA. The DNA grooves have a net negative charge so only the amino acid residues with net positive charge interact with the grooves.

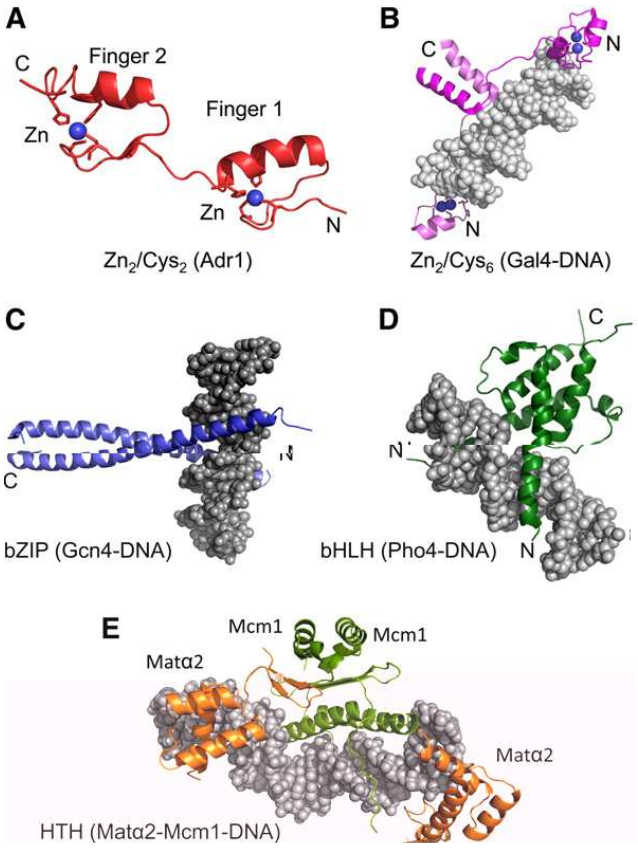

Figure 1.8: DNA-Binding Domains of Five Different Transcription Factors of Yeast. (A) and (B) are two members from the zinc stabilized family with (B) being observed in fungus only, (C) from leucine zipper, and (D) and (E) are from HTH families. Figure from Hahn and Young (2011).

One major difference between the target site recognition on DNA and RNA by proteins is due to the difference of groove width. The double stranded regions of RNA generated by the binding of complementary segments of the RNA molecule, create deep and narrow grooves such that insertion of protein residues is quiet impossible. So most RBPs recognize single stranded regions of the RNA. **Figure 1.9** presents human Pum1 protein structure, which is a typical protein from the Pumilio/FBF (PUF) homology domain family. The RNA-binding domain of Pum1p consists of eight repeats labeled as R1 to R8. Each repeat is made up of three α helices. The middle helix of each repeat directly interacts with a single mRNA base and recognizes a specific nucleotide within the binding site on the mRNA UTR regions. The

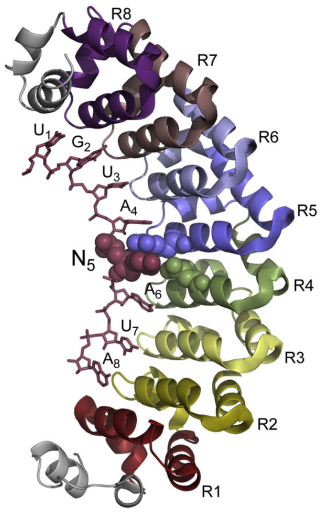

Figure 1.9: Protein Structure of Pum1p of Human a Member of PUF Homology Domain Family. The RNA binding domain of the protein is depicted here. It is comprised of eight repeats (labeled by R1 to R8), each about 40 amino acids long. The repeats stack together and form a crescent structure whose concave surface interacts with the RNA sequence and its convex surface serve as a interaction platform for cofactors. The target RNA and the nucleotide composition of the binding site is also shown. Each repeat is responsible for recognition of one nucleotide within the binding site on RNA sequence. The repeats are made of 3 α-helical segments where the middle helix interacts with the nucleotides. The interactions between the amino acids and the nucleotides is based on hydrogen bonds and van der Waals contacts. Figure from Ryder (2011).

amino acid residues located on the concave side of the Pum1p bind to the mRNA site by creating hydrogen bonds and van der Waals bonds. The convex side of Pum1p provides a platform for the interactions with other proteins.

Figure 1.10 depicts the detailed interactions between specific amino acids on the Pum1p repeats and each of the nucleotides within the consensus binding site on the mRNA. Each nucleotide is identified by three amino acids at specific positions on the middle helix of each repeat. Two of the residues make hydrogen bonds or van der Waals bonds with the nucleobases and the third amino acid establishes stacking interactions with the aromatic rings of the nucleobases. For example, the combination of serine and glutamate residues at positions 1079 and 1083 detects guanine nucleobase, that of glutamine and cysteine/serine recognizes adenine, and that of glutamine and asparagine recognizes uracil (**Figure 1.10b**).

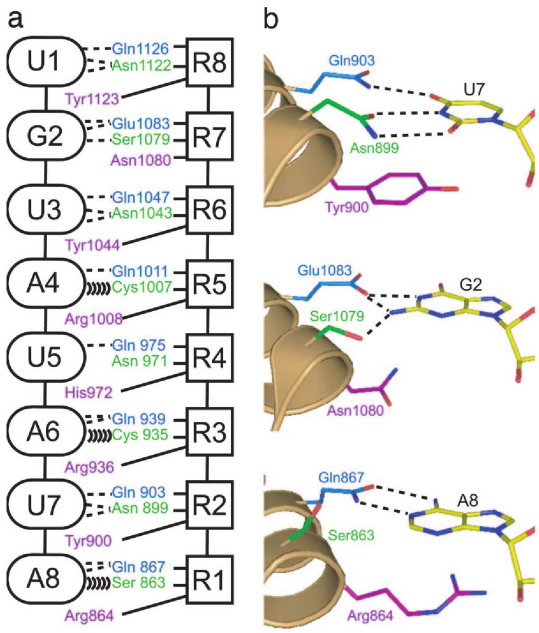

Figure 1.10: Detailed Interactions Between the Amino Acid Residues of Pum1p Repeats and the Nucleotide Sequence of the target mRNA. Protein repeats are shown by squares and RNA bases by ovals. Dashed lines indicate hydrogen bonds and parentheses represent van der Waals contacts. Three amino acid residues from each repeat are involved with RNA base recognition. On the right side a close up look at the Pum1p interaction with uracil (top), guanine (middle) and adenine (bottom) are shown. The nucleotide atoms coloring is as follows: dark blue for nitrogen, red for oxygen, yellow, light blue, green or purple for carbon atoms. Figure from Cheong and Hall (2006).

Studies focused on understanding protein-target recognition usually approach the problem by engineering the binding domain of the protein or the binding site on the target and study the effect of the desired site-directed mutations on the protein domain and/or target site by measuring the equilibrium binding constants. Cheong and Hall (2006) developed cultures with mutant Pum1p and measured the binding constant between the mutant protein and correspondingly mutated binding site on the target mRNA or wild-type target mRNA. As the first test, they mutated Glu-1083 and Ser-1079 in repeat 7 to glutamine and asparagine, respectively. This mutant Pum1p was expected to recognize an uracil at position 2 of the mRNA site. Indeed, they found that the mutant protein bound the target sequence with UUUAUAUA site, with 25 times higher affinity than the wild-type UGUAUAUA, confirming that the identity of this amino acid is important for nucleotide recognition.

In summary, DNA-binding proteins and RNA-binding proteins are able to recognize their targets through their binding domains. The DNA-binding domain or RNA-binding domain consists of proximal or distal regions of the protein that are brought close spatially by protein folding. The combination of amino acid residues within the binding domain interact and recognize the identity of the nucleotides on the DNA or RNA sequence.

1.5 Experimental Techniques for Deciphering the Genetic Code

In this section, we will briefly describe the experimental methods employed to obtain various datasets used in this thesis. In general, the experimental methods can be categorized based on the environment in which they are performed. Experiments carried out within living cells are labeled as *in vivo* and those performed outside a living cell but in a controlled artificial environment are known as *in vitro*. A main weakness of the *in vitro* experiments is that they fail to simulate the exact cellular conditions of an organism, so the interpretation of the results must be done carefully. However due to the absence of the complexities and many regulatory interactions within a living cell, *in vitro* experiments can focus on one or few types of interactions and greatly simplify the system under study. Another classification of the methods is based on the scope of the measurements. A method can be low-throughput screening or high-throughput screening, where the former focus on few genes or gene products whereas the later allows genome-wide measurement.

1.5.1 DNA Microarray Technology

The idea of biochips and their application is probably one of the key factors in the advancement of biological research in the past two decades. It allowed researches to perform measurements for thousands of biochemical reactions in parallel. The

chip, which is about 1 cm^2 in size, consists of thousands of spots, each with a unique single stranded nucleotides or amino acids polymers (i.e. probes) attached to a glass or nylon substrate. Each spot contains many copies of same probe sequence. The detection of presence or interactions between a polymers sample and the chip is based on the hybridization between the probes and the polymers in the sample under study based on reverse complement or protein recognition. Perhaps the most widely used type of this technology is the DNA microarray, which is used for genome-wide mRNA expression level profiling. Living cells respond to external stimuli by reprogramming expression of specific genes, resulting in modified mRNA transcription or turn over rates. DNA microarrays allow for measuring the absolute mRNAs abundances in a sample or relative mRNA abundances between two samples/conditions (Schena *et al.*, 1995; Spellman *et al.*, 1998; van't Veer *et al.*, 2002). They are also used for detection of single nucleotide polymorphisms between a DNA sample and a DNA reference pool (Sapolsky *et al.*; Winzeler *et al.*, 1998) and for high-throughput identification of interactions between proteins and target DNAs/RNAs (Gerber *et al.*, 2004; Harbison *et al.*, 2004; Hogan *et al.*, 2008; Iyer *et al.*, 2001).

First let us discuss the DNA microarray fabrication processes briefly. The microarray can be fabricated by using spotting or *in-situ* synthesis. In spotting method, the nucleotide probes are spotted with fine-pointed pins onto the substrate. A single probe is first amplified using a polymerase chain reaction (PCR) method. PCR is a repetitive step of reverse transcription of the DNA by a thermostable DNA polymerase to yield a double stranded DNA followed by heat treatment to denature the produced double stranded DNA molecules. The produced strands are then used as the probes on the microarray. The *in-situ* methods uses various photolithography techniques, where the single nucleotides adenine (A), cytosine (C), guanine (G) and thymine (T) are chemically attached in different order based on a set of location based patterns. This technique was first developed by Fodor *et al.* (1991). The probes in this method are not full-length DNA, but rather oligonucletides[3]. Currently with the

[3]Oligonucleotide is a short single-stranded chain of nucleotides with a length usually about couple of 10 bases.

in-situ fabrication techniques more than 1 million probes can be synthesised onto every square cm of the chip. Arrays synthesised by *in-situ* techniques have higher probe density and better reproducibility with a much higher cost than the spotting technique.

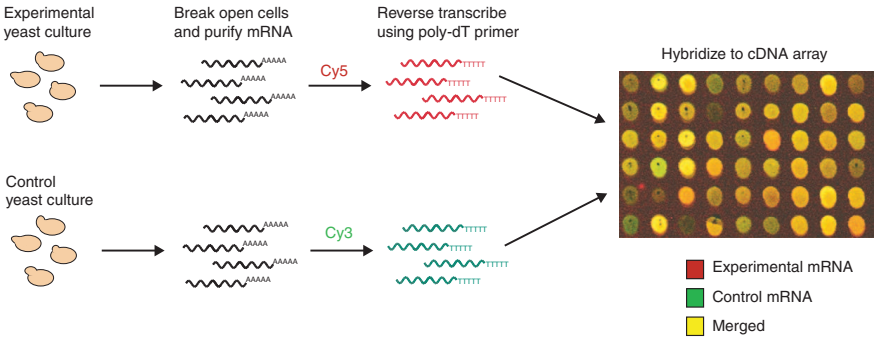

Figure 1.11: The Principle of Genome-Wide Expression Profiling Using cDNA Microarray Technology. mRNA samples from both cultures are reverse transcribed to cDNA and fluorescently labeled nucleotides are incorporated (Cy5 and Cy3). The usage of these two dyes allows the comparison of the two RNA samples on one single array. The resulting labeled cDNA mixture is hybridized to the array based on the reverse complement match to the single stranded DNA probes. The array is then scanned with a laser. The fluorescent scans colors correspond to hybridization intensities for each probe spot. The Cy5/Cy3 fluorescence ratio for each probe reflects relative abundance of that specific cDNA. Figure from Wyrick and Young (2002).

The microarray experiments are based on the principle illustrated in **Figure 1.11**, where relative mRNA abundances are measured between an experimental and the control yeast cultures. The experimental culture can be cells grown in a different condition than the control culture, mutated cells relative to the control, or even related strains or tissues. The cells in both cultures are lysed either by an enzyme or virus and mRNA content of each is collected and purified separately. Because RNA by nature is less stable than DNA, the mRNA strands of each sample is first reverse transcribed in to its complementary DNA strand (cDNA). Note that cDNA molecules are structurally exactly same as DNA molecule, but since it is made from mRNA, the cDNA lacks the intron segments. Before applying the samples to the microarray chip, the cDNA molecules in each of them are labelled using different fluorescent dyes. As shown in the figure, the single-stranded cDNAs are labelled with

red fluorescent dye (Cy5) and green fluorescent dye (Cy3) in the experimental and control samples respectively. The two samples are then mixed in equal proportions and applied to the microarray. The labelled cDNAs compete to hybridize to the complementary probes on the substrate slide. Usually this step is carried out at a specific temperature to minimize the non-specific binding of the cDNAs to the probes on the microarray. Then the slide is washed and scanned with a laser to quantify the relative abundance of a specific cDNA strand between the two samples. This relative abundance is the ratio of the intensity of the red and green light emitted by the dyes. The mean of the pixels intensities that make up a probe is used in the ratio. On the microarray the spot complementary to a particular cDNA sequence is presented with red when the cDNA relative abundance is higher in the experimental sample, green when the relative abundance is lower in the experimental sample and yellow when there is no differential expression for that gene is detected between the experimental and control samples. cDNA microarrays are used for mRNA levels profiling of both *in vivo* and in vitro experiments. They can also be used for measuring absolute mRNA abundances of a single sample where the intensities of a single dye is measured.

One disadvantage of using the DNA microarray is that the ratios do not accurately represent true expression ratios for very low or very high abundances. For low abundances the signals are very noisy and high abundances can lead to signal saturation and therefore bias the results. Signal saturation occurs when the actual signal at a pixel on the chip exceeds the scanner's detection upper threshold. In these cases, all pixels with intensities larger than the threshold will be truncated, regardless of the actual intensities. One way to account for the high abundances is to synthesize custom designed microaarays that take into account the over-representation of the specific mRNAs in the sample, that is, to fabricate chips with enriched probes for the desired spots. Thus the knowledge of the mRNA sequences being interrogated is crucial for array design. Another method/approach is to remove all of the saturated pixels prior of averaging the intensities for each spot. Also cross-hybridization between the probes and cDNAs can lead to false positives when analysing the gene expression profiles (Okoniewski and Miller, 2006).

1.5.2 TAP-tagged Affinity Purification Binding Method

As mentioned in the previous section, microarrays can be used for genome-wide high-throughput detection of the interactions between the RNA-binding proteins (RBPs) ans their target mRNAs. One method for identifying such interactions is known as RBP immunoprecipitation on chip (i.e. RIP-chip) originally established by Tenenbaum *et al.* (2000) to study RNAs associated with RBPs in human cancer cells. The more recent version of this method performed by Gerber *et al.* (2004) and Hogan *et al.* (2008), was based on the precipitation of the endogenously formed RBP-mRNA complexes by using a tandem affinity purification (TAP) tagging method originally introduced by Rigaut *et al.* (1999). It allows for rapid purification of protein complexes even at low concentration. The tag was constructed by attaching two immunoglobulin G (IgG) binding domains of protein A of *Staphylococcus aureus* bacterium, calmodulin binding peptide and a tobacco etch virus (TEV) cleavage site. The tag was fused in-frame at the C-terminus of the respective open reading frame (ORF) in its original chromosomal location of the gene encoding the protein of interest (Ghaemmaghami *et al.*, 2003). It was shown that the function, regulation and stability of most proteins of yeast were not affected by the fused tag (Gavin *et al.*, 2002) and (Ghaemmaghami *et al.*, 2003).

Figure 1.12 depicts the RIP-chip procedure. Yeast cells with TAP-tagged RBP of interest are grown and subsequently are broken down. The tag is presented as protein A in the figure. As mentioned earlier the TAP tag has IgG binding down and the tagged RBP-mRNA complex with bind to the IgG column. The complex is then released from the column by cleavage with TEV protease[4] and the RNA is isolated. This RNA sample is then reverse transcribed in to cDNA and labelled with red fluorescent dye (Cy5). To control for non-specific RNA-RBP binding, whole cell mRNA from wild-type cells lacking the TAP tag are also isolated and the generated cDNAs are labelled with green fluorescent dye (Cy3). The two labelled samples are then mixed in equal proportion and applied to cDNA microarray. The slides are

[4]Protease is an enzyme that catalyzes the hydrolytic breakdown of proteins.

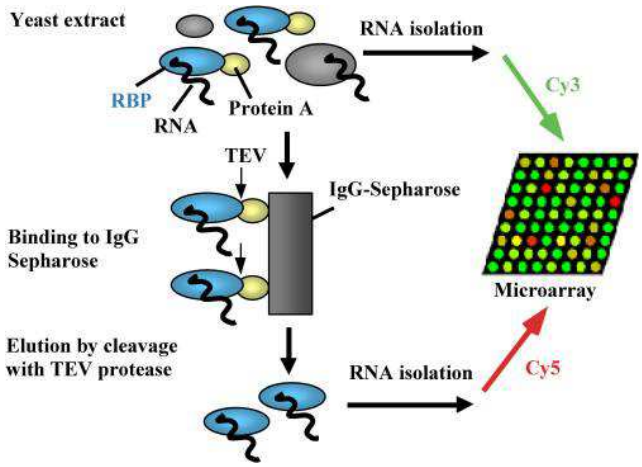

Figure 1.12: Genome-Wide Identification of RNA Associated to RBPs Using Affinity Purification Experiment. Yeast cells with TAP-tagged RBP of interest are grown and then lysed. The tag is shown as protein A. The tagged RBPs-RNA complexes are isolated with IgG-Sepharose and then are released from it by cleavage with TEV protease. RNAs associated with the released tagged RBP are isolated, and cDNA copies are labeled with red fluorescent dye (Cy5). RNAs are also isolated from wild-type cells lacking the TAP tag and their cDNAs are labelled with green fluorescent dye (Cy3). Both samples are then hybridized to yeast cDNA microarray. The Cy5/Cy3 fluorescence ratio for each spot on the microarray represents its enrichment for binding to the RBP. Figure from Gerber *et al.* (2004).

then scanned with a laser and the ratio of the Cy5/Cy3 light intensities reflect the enrichment of the spots for binding to the tagged RBP under the study.

The RIP-chip method is a rather simple procedure that allows for *in vivo* detection of RBP targets expressed at their natural level with negligible distortion on the RBP activity. However the drawback is the concern of the re-association of the RBP to RNAs with higher binding affinity after the cell lysis (Mili and Steitz, 2004).

Similarly, such measurements can be done for high-throughput *in vivo* detection of genomic DNA associated with a specific DNA-binding protein (DBP). In this case, the method is called chromatin immunoprecipitation on chip (i.e. ChIP-chip), where the DNA is fragmented and the DBP-DNA complexes are precipitated with an antibody relevant for the protein under study followed by the microarray measurements (Iyer *et al.*, 2001).

1.5.3 RNA-Seq

RNA-Seq is a high-throughput sequencing-based approach categorized as "next generation sequencing" or "deep" sequencing technology. RNA-seq technology is based on the sequencing method by Sanger *et al.* (1977). The Sanger method utilizes *in vitro* DNA replication. The principle of Sanger sequencing is as follows. First the DNA or cDNA, in the case of RNA, is denatured[5]. Since the replication can not start without a starting sequence to which the polymerase can add new nucleotide to construct the reverse complement of the single stranded DNA, a primer is attached to the beginning of the single stranded DNA molecule. The primer is a short nucleotide acid chain with known composition, with its $3'$-end annealed to the beginning of the DNA sequence. The primer is radioactively or fluorescently labeled so that the final product is detectable. The primer-DNA sample is amplified with PCR process and then the solution is divided in to four tubes with labels "A", "C", "G", and "T". For the DNA replication, a DNA template, normal nucleotides and the DNA polymerase enzyme are required. However, Sanger sequencing method uses two different types of deoxynucleotide, normal nucleotides (dATP, dCTP, dGTP and dTTP) and dideoxynucleotides (ddATP, ddCTP, ddGTP and ddTTP). The difference between normal nucleotides and dideoxynucleotides is that when the later are integrated into the sequence by DNA polymerase, the addition of further nucleotides and the replication is terminated. The Sanger method takes advantage of this feature for sequencing. A solution of all four types of the normal nucleotides in equal amounts are added to each tube. The next step is the addition of ddATP to tube "A", ddCTP to tube "C", ddGTP to tube "G" and ddTTP to tube "T". The concentration of the dideoxynucleotides in each tube is much smaller than that of the normal nucleotides. Finally the DNA polymerase is added to each tube and replication is initiated. Note that the DNA polymerase enzyme starts the replication process from the $3'$-end of the DNA template and moves toward the $5'$-end. As the complementary DNA strand is being

[5]Denaturing of a polymer means to unfold it by a chemical or heating, resulting in a linear structure. In the case of double stranded DNA this process causes the separation of the strands, which results in two single stranded DNA molecules.

synthesized by the polymerase, nucleotides are added one by one on to the growing chain. However, a dideoxynucleotide is integrated sporadically into the chain instead of a normal nucleotide, which terminates the chain. So one is expected to observe the incomplete DNA chains end with ddA in tube "A", ddC in tube "C", ddG in tube "G" and ddT in tube "T". At this point the strands are denatured again and the newly produced chains, which were labeled at their primers, are collected from each tube separately. By sorting the produced chains based on their length for each tube separately, the sequence information of DNA template can be revealed. DNA molecules naturally have a negative net charge because the phosphate backbone releases a proton (H^+), which is absorbed by water molecules. So placing the DNA molecules in an electric field causes the molecules to move toward the higher voltage. If the electric field is applied in a gel medium, then the drag force from the gel on the travelling DNA molecules is proportional to their length and the molecules are separated based on their length with the shorted travel furthest. This method is known as gel shift assay or electrophoretic mobility shift assay (EMSA), which we will discuss more in the next section. The Sanger sequencing approach is depicted in **Figure 1.13**.

Sanger sequencing is low-throughput and expensive and it is not accurate for very short or very long sequences. In contrast, RNA-seq uses massive and parallel sequencing of millions of short DNA fragments simultaneously at much lower cost.

Now we will explain the steps involved in the RNA-Seq approach for genome-wide mRNA sequencing. First the poly(A) mRNA molecules are purified from the cell and fragmented with an enzyme before or after the cDNA synthesis. Sequencing adaptors are ligated to both ends of each cDNA fragment. The fragments are the amplified and sequenced with Illumina Genome Analyser platform (`http://www.illumina.com`). cDNA sequencing with Illumina is carried out in two steps: clustering and sequencing (Nagalakshmi *et al.*, 2010). The clustering station is depicted in **Figure 1.14**. In this step, the denatured single stranded cDNA fragment is loaded into the flow cell. The flow cell surface is packed densely with primers with a solution containing normal

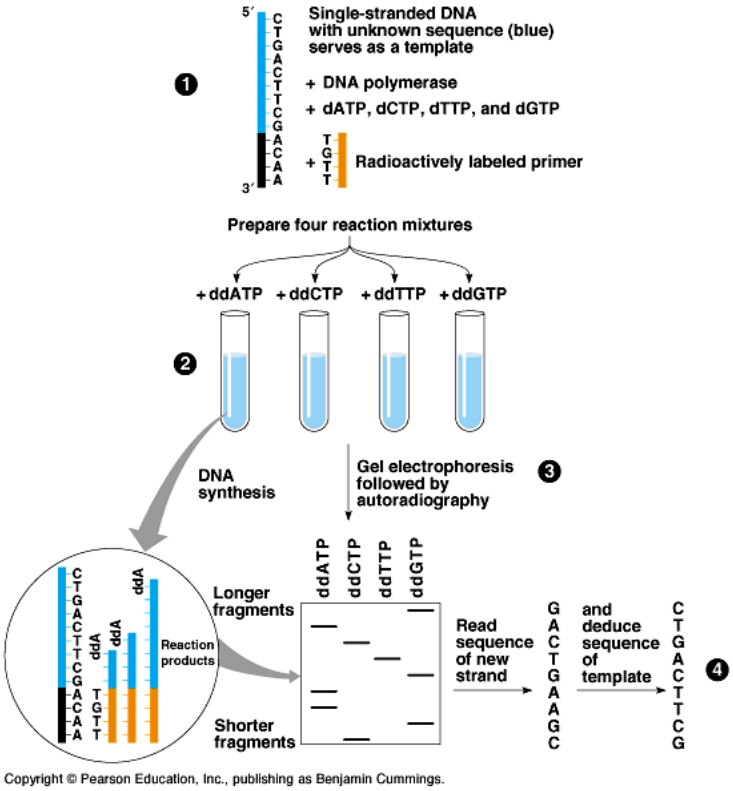

Figure 1.13: The Sanger Sequencing Approach. (1) A radioactively labeled primer is annealed to the 3′-end of the single stranded DNA of interest shown in blue. Once the primer is attached to the DNA, the sample is amplified using a PCR process. The solution is then divided into 4 tubes. Mixture of normal deoxynucleotides (dATP, dCTP, dTTP and dGTP) in equal amount is added to all 4 tubes. (2) In this step, each tube receives one of the four dideoxynucleotide (i.e. ddATP, ddCTP, ddTTP and ddGTP). The concentration of the ddNTP is usually much less than the normal nucleotides. Adding the DNA polymerase enzyme to the tubes initiates the DNA synthesis. DNA polymerase carries out the replication by adding the complementary nucleotides of the bases on the unknown DNA template and synthesis the reverse complement chain. Occasionally, a dideoxynucleotide will be incorporated into the chain, which prevents the addition of further nucleotides. For example in tube containing ddATP, the incomplete chains all end with ddA. (3) The strands in each tube are denatured and the radioactively labeled chains are extracted separately from each tube and applied to gel electrophoresis separately. Since the chains have net negative charge, applying a voltage difference across the gel with higher voltage at the bottom of the gel plate causes the chains to travel downward in the vertical gel columns. The drag force from the gel on the nucleotide chains is proportional to their length. So shorter fragments migrate faster across the gel. The sequence of the new strand can be reconstructed based on the bands appearing on the gel for each tube as shown. (4) The reverse complement of the inferred strand, is the sequence of the unknown DNA template. Figure from `http://www.bio.utexas.edu/faculty/sjasper/bio212/biotech2.html`.

nucleotide and DNA polymerase enzyme. As the sample is introduced into the flow cell, the adaptor-cDNA fragments covalently bind to the surface of the flow cell from end of the fragment. The adaptor on the free end of the cDNA fragment is recognized by one of the primers attached to the surface. At this point DNA polymerase initiates DNA replication and a second strand is produced from the first cDNA fragment. These two strands are denatured and since only one end of each of them is attached to the flow cell surface, the free end adaptor is recognized by a complementary primer on the surface and strand synthesis is initiated again. This process is repeated to generate a cluster of identical DNA fragment all covalently attached to a spot on the surface. Each unique DNA fragment generates one such cluster on the flow cell.

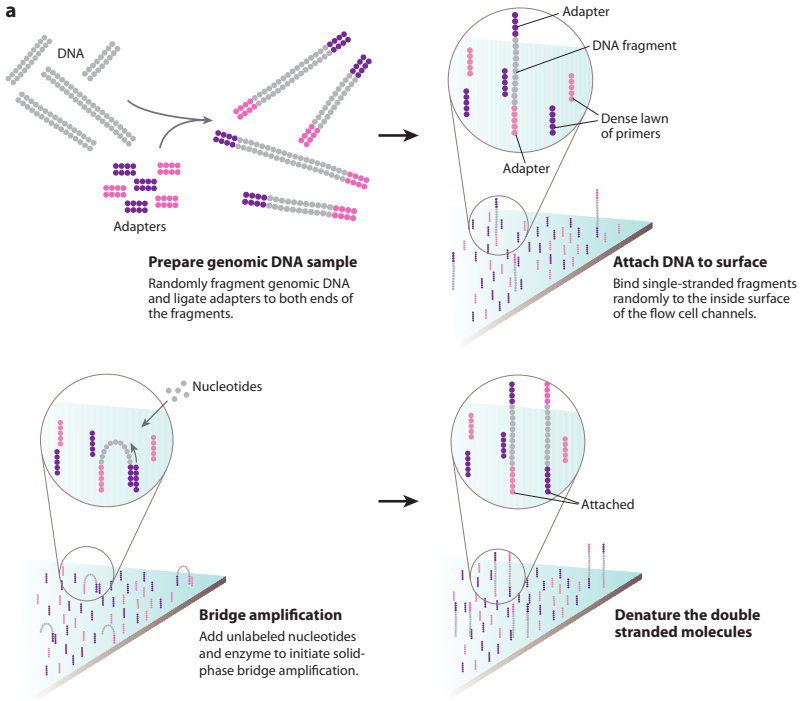

Figure 1.14: Illumina Gene Analyzer Clustering Step. Figure from Mardis (2008).

The flow cell is used by Illumina Genome Analyser to sequence the cDNA fragments. This step is depicted in **Figure 1.15**. Here the concept is similar to Sanger sequencing. First the cDNA fragments are denatured and a sequencing primer that is complementary to the adaptor is attached to each cDNA fragments (i.e. templates) on the flow cell. Sequencing is then performed by DNA synthesis, adding one base

pair at a time to the DNA strands in the cluster. Each type of the nucleotides used are color coded with fluorescent dyes attached to the 3′-OH group, which prevents addition of further nucleotide by DNA polymerase. This way, the strand synthesis freezes after each nucleotide addition, allowing the Genome Analyzer's camera to record the color of each cluster for identifying the type of the incorporated nucleotide at each step. The color labels and the OH group are then cleaved off before the next nucleotide addition cycle. Usually at most 50 base pairs of one or both side of the fragments are sequenced. The reads are then mapped to a DNA library allowing for gapped alignment in the case of RNA sequencing to consider for introns. By calculating the number of reads for each gene in the genome, we are able to calculated the expression level for the genes.

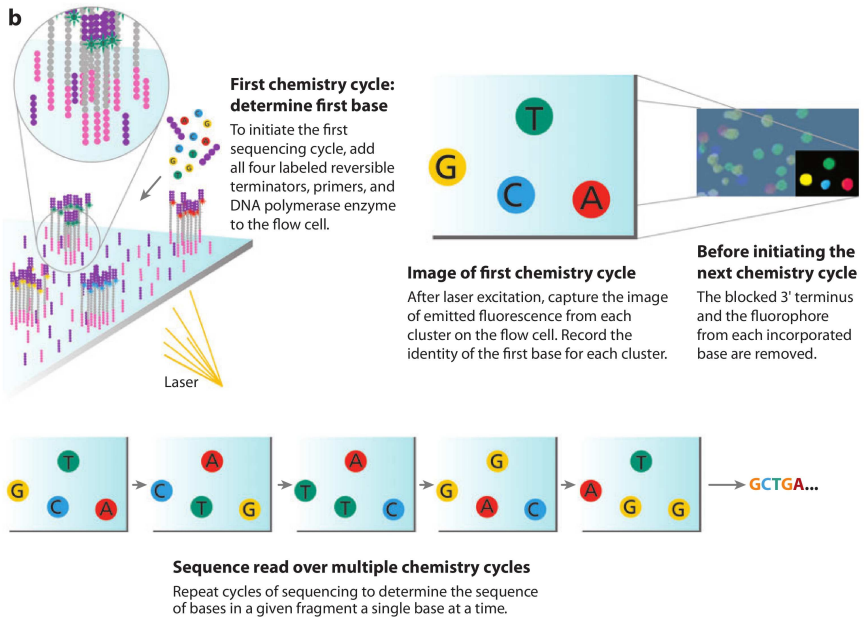

Figure 1.15: Illumina Gene Analyzer Sequencing Step. Figure from Mardis (2008).

The RNA-Seq approach is not bound by the limitations inherent to microarray measurements. In contrast to the microarray technology, the DNA or cDNA sequences are directly determined at single base resolution. Another big advantage of RNA-Seq over hybridization-based methods is the very low background signal for RNA-Seq. Probe saturation is irrelevant unlike for the microarray which can result in biased measurements. RNA-Seq has already been applied to sequence the transcriptome

of various organisms (Nagalakshmi *et al.*, 2008), (Mortazavi *et al.*, 2008), (Cloonan *et al.*, 2008) and (Wilhelm *et al.*, 2008).

As for any developing technology, the RNA-Seq method has its own challenges and weaknesses. Once the sequences of the cDNA fragments are obtained, the reads are mapped to a reference DNA library. So RNA-Seq depends on the quality of the sequenced genome of an organism or cell line. Also, there is the issue of segments that are mapped to multiple locations on the genome, partly because the DNA fragments are not fully sequenced and only about 50 base pairs from one or both ends are sequenced. Despite these shortcomings, RNA-Seq and in general the deep sequencing technologies have revolutionized the transcriptomics field.

1.5.4 Protein-Protein Interaction Identification

Proteins are the main elements in the cell that carry out the various processes. So it is highly important to understand protein interactions. As explained earlier usually several protein cooperate to conduct a specific task. There are several medium-throughput methods that study the protein interactions, which we will explain two common approaches in this section.

Yeast Two-Hybrid Assay

Yeast two-hybrid system is a low-throughput method for *in vivo* detection of physical interaction between two different proteins originally generated by Fields and Song (1989). This method takes advantage of yeast's Gal4p structural properties. This protein is a transcription factor that is required for expression of the genes encoding enzymes for galactose metabolism (Giniger *et al.*, 1985). As explained earlier in **Section 1.3**, Gal4p protein structure consists of two separable domains: the DNA binding domain (BD) and the activating domain (AD). The BD recognizes and binds to the so-called galactose upstream activating sequence (UAS_G) on the DNA sequence and the AD interacts with the RNA polymerase II subunits (Keegan *et al.*, 1986).

Both domain are necessary for the transcription activation by Gal4p. In the presence of galactose, Gal4p induces the expression of its targets by recruiting RNA polymerase II to the transcription initiation site on the DNA (Gill and Ptashne, 1987).

The principle of the yeast two-hybrid assay is demonstrated in **Figure 1.16**. Let's assume that we are interested to test whether two proteins X and Y physically interaction or not. Two hybrid protein complexes are generated as follows. The DNA binding domain of Gal4p (GAL4-BD) is fused in-frame to the ORF for protein X and the activating domain of GAL4 (GAL4-AD) is similarly fused in-frame to the ORF for protein Y. If X and Y do not interact with each other, the two hybrids are not brought into close proximity and there is no activation of transcription of the reporter gene[6] containing the UAS_G. However, if X and Y interact with each other in the nucleus of the yeast cell, the fused hybrids are assembled at the UAS_G site of the reporter gene, which leads to activation of transcription. So detection of the reporter gene transcript in the cell is an indication for the interaction between proteins X and Y.

Affinity Capture

A different approach for identifying protein interactions is to purify the protein complexes and use a method such as mass spectrometry to identify the protein partners. The purification of the protein complex is similar to the TAP-tagged affinity purification that we described earlier. The tag construct is fused in-frame to the ORF of the protein of interest and the protein complexes containing the tag are isolated with an antibody or IgG column similar to the purification procedure described in **Section 1.5.2**. The purified protein complexes are then the proteins in these complexes are separated using a gel shift assay. Gel shift assay, also known as electrophoretic mobility shift assay (EMSA), is a standard method for separating polymers such as proteins or DNA molecules based on their net charge, size and shape (Chelm and Gei-

[6]Reporter gene is a gene whose expression level can be monitored and measured easily (e.g. green fluorescent protein). The gene is attached to the promoter region of interest and the construct is transfected into a cell or tissue.

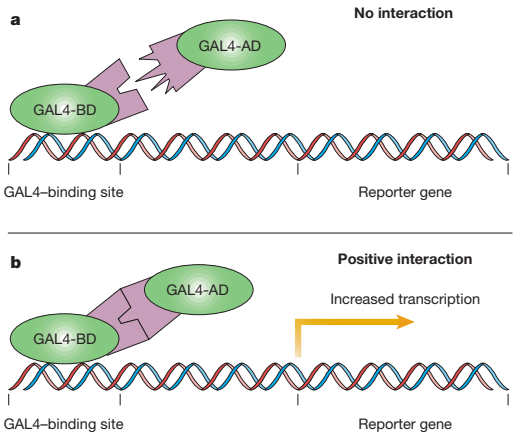

Figure 1.16: Yeast Two-Hybrid Assay for Detection of a Physical Interaction between two proteins X and Y. The two proteins, X and Y, are shown in purple color. The Gal4p DNA-binding domain (GAL4-BD) is infused in-frame to the ORF of the gene encoding protein X, shown as the hybrid protein on the left and the activating domain (GAL4-AD) is infused in-frame to the ORF of the gene encoding protein Y, shown as the hybrid protein on the right. (a) The X and Y region of the two hybrids are not interacting and RNA pol II is not recruited to the reporter gene. So the reporter gene is not expressed. (b) In this case, X and Y interaction brings the two GAL domains close together and consequently the RNA pol II is recruited and the transcription of the reporter gene is induced. Figure from Pandey and Mann (2000)

duschek, 1979; Garner and Revzin, 1981). An electric field is applied in a gel plate (**Figure 1.17A**). The polymers with larger net charge, shorter length and linear structure travel through the gel faster since they are less impeded by drag force from the gel medium. So after applying the solution containing the protein complexes, different bands appear on the gel each representing a different protein (**Figure 1.17B**). Some gel shift assays can even separate molecules different in length by one nucleotide or amino acid. Proteins naturally fold to generate complex 3-dimensional structures and unlike the DNA sequences that have net negative charge, proteins can be neutral or have a positive or negative net charge based on their amino acid composition. So the proteins are usually denatured and coated with negative charge that is proportional to the protein's length, so that the result of gel assay separation is only based on the length of different proteins present in the complexes. Once different proteins in the purified complexes are separated in the gel, each band is physically cut out of the gel and identified by mass spectrometry based on the mass-to-charge-ratio calculated

from the flight time in a known electric field (Domon and Aebersold, 2006).

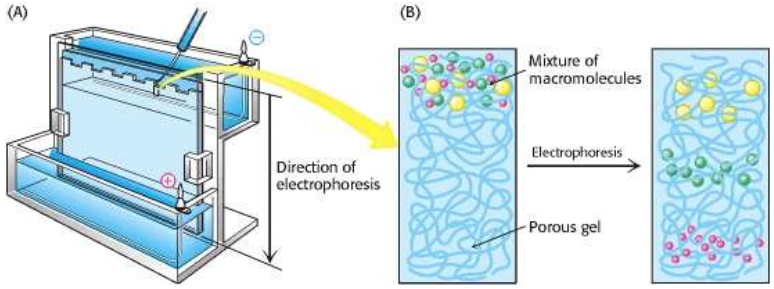

Figure 1.17: Electrophoretic Mobility Shift Assay (EMSA). In biology, it is used for separating different macromolecules such as protein, DNA or RNA strands based on the net charge, size and shape. An electric field is applied across the gel plate (A). The solution containing the macromolecule complexes is injected to the top of one of the gel channels. DNA and RNA molecules inherently have negative net charge due to the phosphate back bone. However proteins can have zero, positive or negative net charge depending on their amino acid composition. Therefore, the proteins are first unfolded and coated with negative charge proportional to their length. The macromolecules will move toward the cathode; however, the drag force from the gel pores on the moving molecules slows them down. The drag force depends on the size and the shape of the molecules. As a result the macromolecules migrate with different velocities and each band in the gel is generated by a specific type of macromolecule (B). EMSA is also used to study the interaction between a known DNA (or RNA) and the protein of interest. First the DNA-protein or RNA-protein solution and a control solution without the protein are applied to two separate gel channels. Observing a band on the mixed solution channel, which has traveled less distance compare to the control sample, indicate direct interaction between the protein and the DNA (or RNA) molecules. Figure from Berg *et al.* (2002).

Another method for protein detection is based on antibody labeling known as Western blot. It was originally developed by Towbin *et al.* (1979). In this method, the protein complexes must first go through a purification and separation as explained earlier. After the separation step, all the bands are transferred and secured to a membrane. Detection of a specific protein within the blot is done with an antibody. The antibody solution is added to the membrane, followed by a rinsing step to remove the unbound antibody. A secondary antibody, linked to a chemically luminescent agent, is bound to the first antibody. The image of the membrane with luminescent plots is captured on a photographic film.

This method does not differentiate between direct and indirect protein-protein interaction. Indirect interaction can happen when the tagged protein and another protein both bind to different segments of the RNA and does not physically interact or when when both proteins interact together via a third component.

1.5.5 *Delitto Perfetto* Approach for Allele Replacement Experiments

In biology, one can learn about the interactions between the genome, transcriptome and proteome of an organism by applying a modification such as a single nucleotide polymorphism (SNP) in a specific gene that may be carried into the encoded protein and study the effect of such modification. For example it is possible to measure the changes of equilibrium dissociation constant (K_d) for DNA-DBP, RNA-RBP or protein-protein bindings, when one or more amino acid residues in the binding domain of the proteins or nucleotide residues of the binding site on DNA or RNA is altered. These type of reverse genetic[7] approaches help to reveal which amino acid residues or nucleotides are critical for recognition of the binding sites. To perform such measurements, it is necessary to develop the experimental methods that introduce such mutations in the genes of interest. The *Delitto Perfetto* method is a site-directed *in vivo* mutagenesis approach developed by Storici and Resnick (2006). It is a clean and rapid method based on homologous recombination originally applied to yeast cells.

Mutagenesis in this method consists of two steps depicted in **Figure 1.18**. The first step involves the integration of a COunterselectable REporter (CORE) into the location of the gene of interest. Because of the homology between the CORE cassette flanking sides and the upstream and downstream regions of the gene, a recombination can sporadically occur. This means that the CORE cassette replaces the coding region

[7]Reverse genetic approaches investigate the impact of induced genetic modification of a specific gene (e.g. by inserting mutation, deletion or gene silencing) and infer the gene function through the detection of physical or biochemical changes of the mutant cell.

of the gene under study **Figure 1.18A**.

Storici and Resnick designed the CORE cassette for yeast cells such that it contains two controls for testing the successful integration of the cassette. Yeast cells are are naturally sensitive to hygromycin antibiotic which can kill the cell. One design for the CORE cassette includes a reporter that provides resistance to this antibiotic (Goldstein and McCusker, 1999). This means that only those cells with the cassette incorporated properly will survive a hygromycin medium. Another very common control to use an auxotrophic strain. A yeast cell is able to synthesize all of the amino acids required for its protein synthesis. If the media lacks a certain type of amino acid, for example uracil, then the uracil synthesis pathway is induced and uracil is synthesis from the inorganic compounds in the media. This means that a yeast strain lacking the gene that encodes a protein crucial for uracil synthesis, will not survive in a medium with uracil deficiency. This can serve as another control to ensure that the CORE cassette is integrated into the location. We can start with an auxotrophic strain engineered specific to lack URA3 gene (i.e. $ura3^{-}$). Then the CORE cassette is designed to include URA3 gene. Only cells with correct integration of the cassette will survive media lacking uracil amino acid.

The second step involves complete removal of this cassette with oligonucleotides that contain a desired allele of the gene of interest between the side flanking regions. Again, based on the homologous recombination the oligonucleotide will replace the CORE cassette in some of the cells (**Figure 1.18B**). These cells are expected to be for example sensitive to hygromycin again. We used this approach to design an experiment to test our finding, which is discussed in Chapter 4.

1.6 Summary

In this chapter, we gave an overview on the eukaryotes genome architecture and different regulatory mechanism performed by various types of proteins. We introduced the traditional and contemporary views on gene expression. The traditional model views

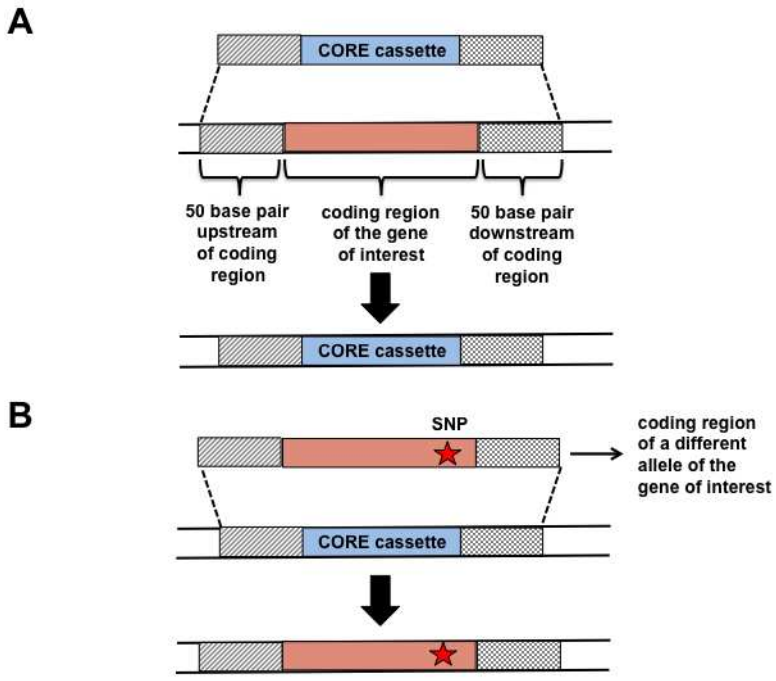

Figure 1.18: *Delitto Perfetto* Approach for Allele Replacement.

each step of the gene expression program as a separate process; whereas, the contemporary model states that different regulatory processes have temporal overlapping and in a way, each step can regulate its downstream events. That is, transcription, post-transcriptional processing and translation are tightly coupled and affect one another. We also provided some experimental approaches for analyzing the genome-wide gene expression of a cell, protein-DNA and protein-RNA interactions. The methods include: high-throughput mRNA level profiling by DNA microarrays, protein-target interaction assay, protein-protein interaction and DNA/RNA sequencing. These topics are necessary for understanding the projects discussed in the next four chapters.

Chapter 2

Inferring Quantitative Sequence-to-Affinity Models for RNA-Binding Proteins

This chapter is adopted from a manuscript co-authored by Mina Fazlollahi, Eunjee Lee, Xiang-Jun Lu, Pilar Gomez-Alcala and Harmen J. Bussemaker.

2.1 Introduction

Post-transcriptional processing carried out by RNA-binding proteins (RBPs) is critical for the regulation of RNA abundances. These processes include assembly, splicing, edition, localization, and stability of RNA transcripts. There have been more studies that focus on DNA biding factors and their interaction network. Unlike the double helix structure of the DNA, the RNA molecules fold into complex secondary structures. Therefore, detection of binding motifs[1] associated with RBPs are challenging due to more complicated structure of RNAs.

[1]Sequence motif is a short nucleotide or amino acid sequence that is usually referred to as the binding site for the protein-DNA or protein-RNA interactions.

In recent years, some studies have focused on identifying stability-associated nucleotide motifs that are critical in regulating RNA steady state rates (e.g. (Foat *et al.*, 2005; Hogan *et al.*, 2008; Riordan *et al.*, 2011; Shalgi *et al.*, 2005)). By interacting through these sequence motifs, which are mostly located within the untranslated regions (UTRs) of messenger RNAs (mRNAs), RBPs stabilize or destabilize their target mRNA transcripts. Some studies look for the enrichment of specific k-meres (i.e. nucleotide sequence of length k) that correlate to mRNA expression profile and mRNA half-live data (Shalgi *et al.*, 2005). Some other studies attempt to identify k-mers enriched within the mRNA sequences of the target set of RBPs based on mRNA binding data (Gerber *et al.*, 2004; Hogan *et al.*, 2008).

Our approach combines biophysical modeling of the interactions among RNA and RBPs with the use of RBPs genome-wide mRNA binding data. Unlike some other studies, our motif discovery method does not require a priori RBP target sets. We searched for potential regulatory elements in mRNA sequences that are recognized by diverse RBPs. We used a similar approach by (Foat *et al.*, 2006) searching for binding sites in the form of sequence specific affinity matrices (PSAMs). We were able to obtain known binding motifs for 15 various RBPs including novel motifs for Scp160p, Sik1p and Tdh3p.

2.2 Methods

2.2.1 Experimental Data Used

For our motif search, we analyzed genome-wide imunoaffinity purification data for 45 different RNA binding proteins by (Hogan *et al.*, 2008). In this assay, mRNA molecules bound to C-terminal tandem affinity purification (TAP)-tagged proteins were isolated at mid-log phase from cells growing in YPD media[2]. The mRNA mixture

[2]Yeast Extract Peptone Dextrose abbreviated as YPD, is a complete medium for yeast growth containing yeast extract, glucose, amino acid monomers (peptides) and water.

was then hybridized to a DNA and/or oligonucleotide microarray. For each RBP, 2 to 6 experimental replicates were performed, for a total of 132 IP experiments.

2.2.2 Pre-Processing of RBP Binding Data

For our motif analysis, we started from $\log_2$-ratios between the microarray intensities for the immunoprecipated sample and input sample, respectively, for each RBP. To reduce the effect of outliers, we applied a rank-quantile transformation based on the standard normal distribution. For each RBP, let $x = (x_1, x_2, \cdots, x_n)$ denote the vector of binding log-ratios across all genes, sorted in ascending order. We first ranked the data points in each column. Let $\Pr(X < \chi)$ denote the cumulative distribution function (CDF) for a standard normal random variable X with mean $\mu = 0$ and standard deviation $\sigma = 1$. As illustrated in **Figure 2.1**, we then defined χ_i as the i^{th} quantile,

$$\Pr[X < \chi_i] \equiv \frac{\text{rank}(x_i) - \frac{1}{2}}{n} \qquad \text{for} \quad i = 1, 2, \cdots, n \tag{2.1}$$

and used it to replace the i^{th} element in the vector x. With this transformation, we ensure that the effect of outliers is diminished.

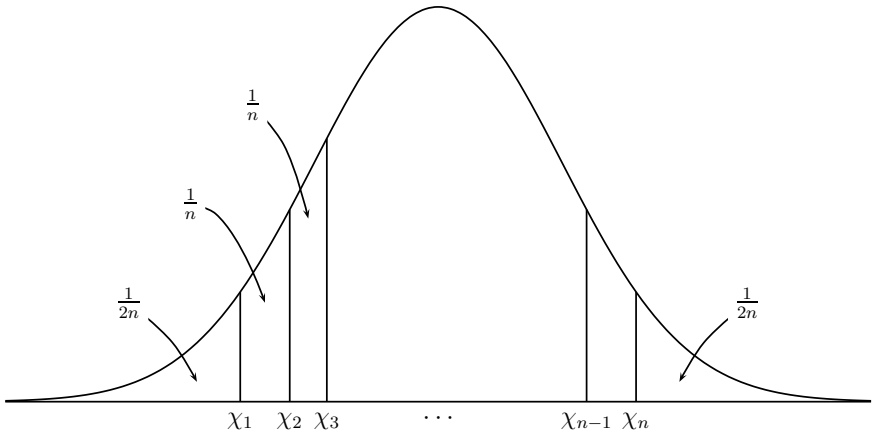

Figure 2.1: Schematic Representation of Rank-Quantile Transformation Step Applied to Each Column (Size n) of the Binding Data. We assigned i^{th}-quantile value (χ_i) to the i^{th} element based on the rank of data point x_i.

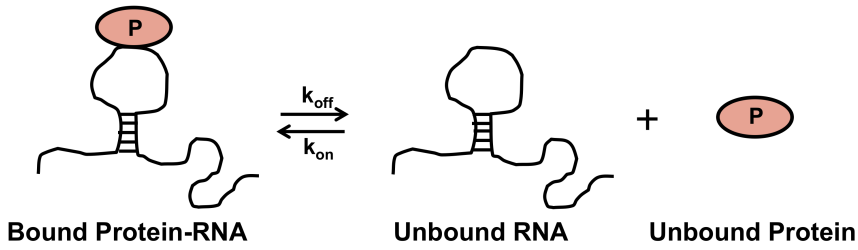

Figure 2.2: Model Used for Quantification of RNA-Protein Binding. Here k_{on} and k_{off} are equilibrium binding and unbinding rates, respectively.

2.2.3 Quantitative Model of RNA-Protein Binding

Let's consider a simple situation where a single protein P binds to a RNA strand S as shown in **Figure 2.2**. At equilibrium the RNA and the protein from an RNA-protein complex at rate k_{on} and detach at a rate k_{off}.

The association constant K_a is defined as the ratio of the concentraion of the RNA-protein complex, $[SP]$, and the product of the free RNA and protein densities, $[S]$ and $[P]$, respectively:

$$K_a(S) \equiv \frac{[SP]}{[S][P]} = \frac{k_{\text{on}}}{k_{\text{off}}} = e^{-\Delta G/RT} \tag{2.2}$$

Here, ΔG is the Gibbs free energy of binding per mole, R is the gas constant, and T the temperature.

The occupancy $N(S)$ of RNA sequence S by protein P can be written as the fraction of bound RNA to the total RNA concentration (bound and unbound) (Bussemaker *et al.*, 2007; Foat *et al.*, 2006).

$$N(S) = \frac{[SP]}{[S] + [SP]} = \frac{1}{\frac{[S]}{[SP]} + 1} = \frac{[P]}{K_a^{-1}(S) + [P]} \tag{2.3}$$

In the low protein concentration regime, where $[SP] \ll [S]$, or equivalently, $K_a \ll [P]$, we have

$$N(S) \approx [P]K_a(S) \tag{2.4}$$

So far we showed that in low protein limit, the occupancy of S is proportional to association constant K_a. The association constant depends on both the protein P binding domain and RNA strand nucleotide compositions. Next, we want to quantify K_a in terms of the nucleotide information of binding site on the RNA strand. We are going to make a second assumption: additivity of binding energies for each nucleotide base in the binding region. As far as the RNA sequence is concerned, we further assume the binding energies only depend on the nucleotide type (A, C, G or U) and their position in the binding side (Benos *et al.*, 2002). With this assumption, we neglect any dinucleotide or higher order dependencies. It might seem a crude assumption. However, as a first order approximation in modeling the RNA-protein interactions, we were able to recover many experimentally validated binding motifs as explained in **Section 2.3.1**. Let's assume the reference binding site S_{ref} be the nucleotide sequence that has the highest K_a to protein P. Also assume S_{mut} is single nucleotide mutation of base b at position j relative to S_{ref}.

$$w_{bj} \equiv \frac{K_a(S_{\text{mut}})}{K_a(S_{\text{ref}})} = e^{-\Delta\Delta G_{jb}/RT}, \qquad \Delta\Delta G_{jb} = \Delta G(S_{\text{mut}}) - \Delta G(S_{\text{ref}}) \tag{2.5}$$

By quantifying the effect of all three possible point mutations of the reference base at every position in the binding site, we can calculate the relative occupancy $N(S)$ of a sequence S with more than one mutation form the reference sequence S_{ref}. Thus, the occupancy of a particular binding site S of length L_ϕ is:

$$N(S) = [P]k_a(S_{\text{ref}}) \prod_{j=1}^{L_\phi} w_{jb_j(S)} \tag{2.6}$$

The occupancy $N(S)$ for the entire sequence S equals the sum of occupancies for each binding site of size L_ϕ sliding over the whole sequence of length L shifted one position at a time.

$$N(S) = [P]K_a(S_{\text{ref}}) \sum_{i=1}^{L-L_w+1} \prod_{j=1}^{L_\phi} w_{jb_{i+j-1}(S)} \tag{2.7}$$

We can label the sum term as *in vitro* specificity (or affinity) K of the mRNA to protein p. So the occupancy equation can be rewritten:

$$N(S) = [P]K_a(S_{\text{ref}})K(S) \tag{2.8}$$

In the next section we will explain RNA-binding proteins motif search using MatrixREDUCE software. MatrixREDUCE performs PSAM training based on the model discussed in this section.

2.2.4 Motif Search for RNA-Binding Proteins

Our RNA-binding proteins (RBPs) motif discovery approach is shown in **Figure 2.3**. To detect the motifs, we used the MatrixREDUCE program from the REDUCE Suite package (`http://bussemakerlab.org/software/REDUCE`) to perform a genome-wide fit of a position-specific affinity matrix (PSAM) to the rank-quantile $\log_2$-ratios of RBP binding data. MatrixREDUCE builds a multivariate linear model originally developed by Foat *et al.* (2006, 2005). We used the enhanced version of MatrixREDUCE implemented by Dr.Xiang-Jun Lu.

The MatrixREDUCE algorithm consists of two steps: seed motif finding and PSAM optimization. The seed motif finding seeks to identify the sequence motif of desired length whose occurrence best correlates with the binding signal within all of the motifs with possible nucleotide combinations. The motif size is allowed to vary from to 1 to 8 nucleotides. Once the optimal motif has been identified, it is used as a seed for the optimization procedure. Let's assume the optimal motif has length L. First a matrix of size $4 \times L$, representing each nucleotide A, C, G, and T/U at positions 1 to L is constructed. At every column (i.e. position in the seed motif) the best nucleotide

element is given K_a equal to one and unacceptable nucleotides (i.e. the other three element) are given a very small number close to zero. Optimization step aims to find the optimal weight matrix by minimizing an error function:

$$(C, \{F_e\}, \{W_{jb}\}) = \text{argmin} \sum_e \sum_g (Z_{ge} - F_e \sum_{i=1}^{L_g - L_\phi + 1} \prod_{j=1}^{L_\phi} w_{jb_{i+j-1}(S_g)} - C)^2 \qquad (2.9)$$

where Z represents the rank-quantile binding data in our case. The parameters e and g stand for experiment indices and genes, respectively. The optimization is based on the Levenberg-Marquardt (LM) algorithm to find the optimal W_{jb}'s. LM algorithm is a blend of gradient descent and Gauss-Newton iteration (Madsen *et al.*, 2004). Once the Optimization step converges for this PSAM, the residues of Z is the used for the next seed finding and optimization iteration. By default the affinity of best binding site is equal to 1. That is the weights for every position in the optimized PSAM is normalized relative to the best nucleotide. The C parameter represents the genome-wide basal expression level when no preferred motif is present on the sequence.

We split every column of the binding data randomly into two sets of equal sized training and test sets and ran MatrixREDUCE on the training set of all experimental replicates of a RBP (ϕ) simultaneously (using command line option -mf). For every RBP, we searched for binding motifs on the whole mRNAs, $5'$ UTRs, ORFs and $3'$ UTRs sequences separately. For some of the RBPs, we also ran the software without -mf argument because one of the experimental replicates out of 2 was missing more than 40% of the data points, Idh1p, Nrd1p, Tdh3p and Vts1p.

We obtained the *Saccharomyces cerevisiae* UTR sequences from a study using RNA-seq method to obtain the transcriptional landscape of the yeast genome by (Nagalakshmi *et al.*, 2008). The mRNA open reading frame (ORF) sequences were downloaded from Yeast Genome Database (SGD; `http://www.yeastgenome.org`). For all RBPs we searched for PSAMs from length 1 to 10 iteratively with p-value cut-off of

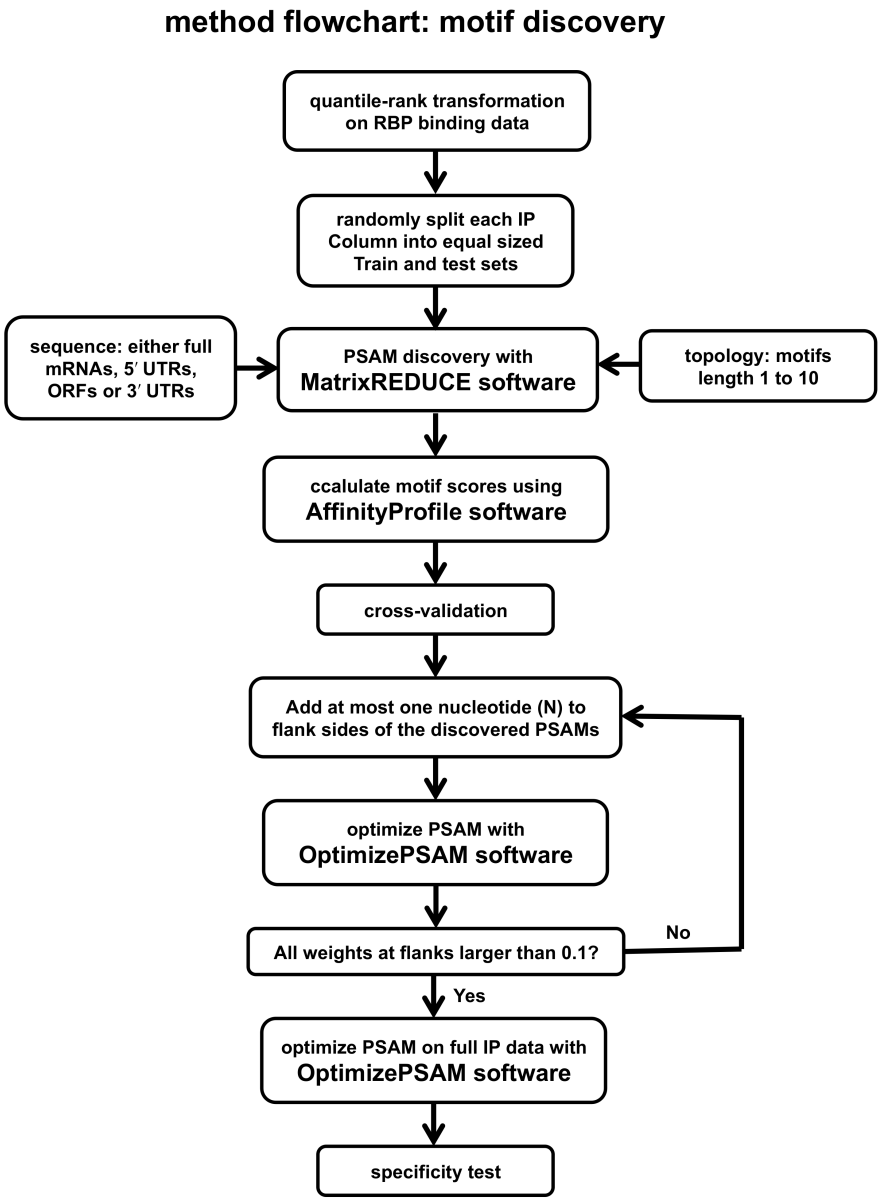

Figure 2.3: The Flowchart Representation of Our Motif Search Approach (see text for details).

0.001.

2.2.5 Computational Validation of Obtained PSAMs

We calculated the affinity scores of the discovered PSAMs using the AffinityProfile software from the REDUCE Suite package. We then calculated the Pearson t-value and Spearman p-value for the correlation of the affinity scores to the test data set. We further tried to capture any low specificity flanking sides of the motif for the PSAM that passed the cross-validation step. We extended both of the flanking sides at most one nucleotide position (i.e. added column (1,1,1,1) to the flanks of the PSAM) and ran the OptimizePSAM software from the same package using the PSAM with added columns on the sides as seed. We continued adding columns to the sides of the PSAM until no nucleotide's weight at the added side flanks is less than 0.1. In the case of Nrd1p PSAM optimization we neglected this criteria where one of the matrix element for G nucleotide at position 8 was equal to 1.0×10^{-7}. Further flank addition and optimization of this PSAM resulted in optimization divergence even after several rounds. At the end of every optimization round, we cross-validated the newly extended PSAM by calculating Pearson t-value and Spearman p-value. After this step, we ran OptimizePSAM using the full data on the PSAMs passed cross-validation steps.

The final set of PSAMs was obtained from the PSAMs that pass the test for specificity to their own IP experiment among all the experiment and If the affinity of a specific PSAM on UTRs and/or ORFs had the highest correlation to at least one of the relevant IP experiments among the 132 experiments, that RBP-mRNA region combination would pass our test. In the case of YLL032C we had to use the whole mRNA sequences to calculate the affinity for the factor to pass the specificity test.

2.2.6 Functional Assessment of the Novel motifs

Gene Ontology Enrichment Analysis

Gene Ontology (GO) (Ashburner *et al.*, 2000) is a unifying tool among many eukaryotic organisms. Many proteins in any given eukaryote are involved in the same biological pathway. Because the experimental research on model organisms like budding yeast *Saccharomayces cerevisiae*, the nematode worm *Caenorhabditis elegans* and fruitfly *Drosophila melanogaster* developed independently, the need for a common language for describing the roles of genes or gene product (annotation) was necessary. Each GO category represents a particular organism-independent biological process, molecular function or cellular component. **Figure 2.4** displays a GO category for DNA metabolism process shared among all eukaryotes. The genes involved in each category from three organisms yeast, fruitfly and mouse are listed.

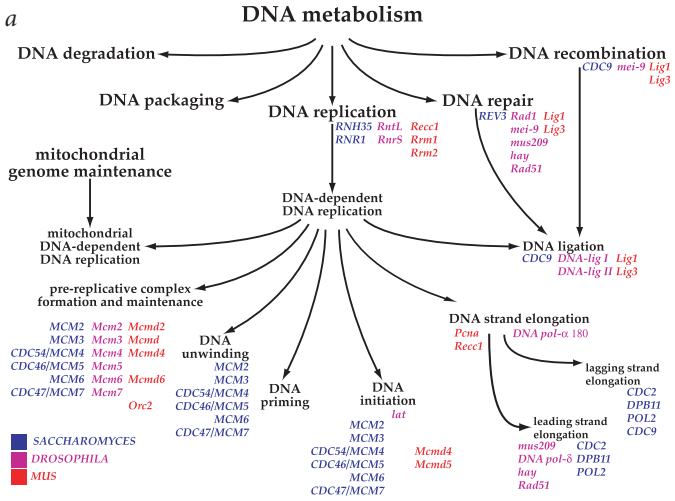

Figure 2.4: Example of a Gene Ontology (GO) Category. It depicts the genes of 3 different organisms categorized under "DNA metabolism process". Figure from Ashburner *et al.* (2000).

We used GO enrichment scoring analysis to detect the underlying regulatory program, cellular state or cellular component for the novel motifs. For each GO category, we tested whether the Affinity scores for a PSAM on ORF or UTR sequences are associated with a specific biological pathway or not. We applied the non-parametric

Mann-Whitney-Wilcoxon test to determine whether the affinity scores of mRNA regions within a particular GO category have a different distribution than the affinity scores for all other mRNAs. We used an iterative procedure for removing the effect of redundant nested GO categories which was implemented originally in the T-profiler algorithm (Boorsma *et al.*, 2005). We only considered GO categories that have at least 10 genes. To correct for multiple testing, we performed a Bonferroni correction on the resulting p-values accepting only categories with p-values smaller than 0.01/N where N is the number of unique GO categories. We also used Student's t-test to verify whether genes enriched in GO categories were significantly upregulated or downregulated for a specific category based on the t-value sign.

To perform the GO enrichment analysis, we downloaded packages *GO.db*, and *org.Sc.sgd.db* for yeast from the Bioconductor website within the R statistical programming environment (`http://www.Bioconductor.org`).

Correlation to Condition-Specific Expression Data

To further validate the novel motifs, we correlated the affinity scores of the 25 factors to expression data from 173 different stress conditions (Gasch *et al.*, 2000). The stress conditions include: heat shock, oxidative or reductive chemical agents, nutrients or amino acids starvation, and osmosis. To calculate the correlations, we performed a multiple linear regression of the genome-wide mRNA expression levels of each condition to the affinity scores of all of the selected RBP-region combination and compared the Pearson t-values among different conditions.

2.3 Results

2.3.1 RBP Binding Motif Search

Our motif discovery method is depicted in **Figure 2.5**. We used the MatrixREDUCE software from the REDUCE Suite package, which takes as inputs the nucleotide sequences and RBP binding logratios for all mRNAs. To reduce the effect of outliers, we transformed the binding data to quantiles of a standard normal distribibution based on their ranks. We randomly picked 50% of the data as a training set.

To define the sequences, we used the annotations from (Nagalakshmi *et al.*, 2008) and extracted 5′ and 3′ untranslated regions (UTRs), open reading frames (ORFs) and whole mRNA nucleotide sequences. For every RBP, we performed our genome-wide motif search on whole mRNAs, ORFs, 5′ and 3′ UTRs separately. One motivation for this was to allow for functional differences between these various parts of the mRNA transcript. Also we observed that some cases, using the whole mRNA sequence seemed to reduce the signal to noise ratio. For instance, for Nrd1p and Puf2p, we were not able to find any significant binding site using the full mRNA, whereas by using only the 3′UTRs for our search, we were able to find a significant motif.

After the training step, we calculated the affinity scores using the derived PSAMs (see methods section) and only accepted those that passed our cross-validation step using the other 50% of binding data (test set). We optimized these PSAMs by adding flanking positions (maximum of one nucleotide added to either sides at every optimization iteration) to capture any low specificity sites that were not captured during our training step. After each optimization round, we cross-validated to check for over fitting. Finally we optimized these PSAMs on full data. Out of 45 proteins, we were able to obtain PSAMs for 20 different RBPs **Table 2.1**.

Most mRNA regulations are carried out through protein interactions with the 3′ UTR of mRNAs. However it can be possible that some proteins bind to 5′ UTR or ORFs and thus be involved in activation or suppression of difference mRNAs. For

Table 2.1: PSAM Training Statistics From the MatrixREDUCE Software

RBP	**mRNA Region**	**Pearson t-values on IP experiments**	**Spearman p-values on IP experiments**	**Accept**[a]
Gbp2	ORFs	25.3, 19.5, 15.8, 14.6	1.2×10^{-133}, 3.1×10^{-105}, 1.9×10^{-31}, 4.0×10^{-27}	Yes
Idh1	ORFs	9.1, 15.8	2.7×10^{-27}, 1.3×10^{-86}	No
Khd1	ORFs	27.2, 25.7, 28.5, 23.1, 22.2	3.0×10^{-89}, 2.0×10^{-74}, 2.0×10^{-109}, 3.6×10^{-68}, 1.0×10^{-50}	Yes
Mrn1	mRNAs	21.8, 21.0, 18.5, 12.9	1.3×10^{-83}, 3.0×10^{-68}, 1.8×10^{-71}, 8.4×10^{-19}	No
Msl5	ORFs	9.3, 13.8	3.2×10^{-1}, 1.8×10^{-11}	Yes
Nab2	ORFs	23.8, 22.4, 21.3, 18.8	1.7×10^{-142}, 1.5×10^{-114}, 3.3×10^{-122}, 3.4×10^{-114}	Yes
Nrd1	3′ UTRs	4.2, 7.7, 8.4	3.0×10^{-8}, 1.6×10^{-13}, 3.1×10^{-16}	Yes
Pin4	3′ UTRs	6.3, -1.0, 11.5	1.5×10^{-7}, 1.9×10^{-4}, 4.0×10^{-15}	Yes
Pub1	3′ UTRs	28.3, 27.2, 30.0	8.7×10^{-143}, 1.6×10^{-145}, 2.8×10^{-152}	Yes
Puf1	ORFs	17.9, 3.3, 6.9, 20.3	2.3×10^{-65}, 7.0×10^{-25}, 2.1×10^{-19}, 1.8×10^{-89}	No
Puf2	3′ UTRs	15.6, 18.2, 17.6, 16.5	9.0×10^{-12}, 2.4×10^{-27}, 6.8×10^{-15}, 1.3×10^{-21}	Yes
Puf3	3′ UTRs	22.1, 22.8, 21.7, 25.0, 23.2	1.1×10^{-24}, 1.6×10^{-17}, 2.2×10^{-19}, 8.0×10^{-26}, 1.6×10^{-39}	Yes
Puf4	mRNAs	21.9, 30.9, 27.8, 24.3	2.3×10^{-73}, 1.7×10^{-123}, 2.1×10^{-73}, 1.7×10^{-43}	Yes
Puf5	mRNAs	19.6, 16.8, 18.2, 18.9	4.6×10^{-67}, 1.0×10^{-37}, 3.5×10^{-30}, 3.5×10^{-38}	Yes
Rna15	ORFs	16.6, 22.1, -2.4	3.1×10^{-89}, 1.5×10^{-136}, 1.4×10^{-5}	No
Scp160	ORFs	15.2, 24.7, 35.9, 38.7, 35.1	1.9×10^{-72}, 4.4×10^{-166}, 4.4×10^{-290}, 0, 0	Yes
Sik1	5′ UTRs	13.7, 13.9	1.4×10^{-49}, 9.1×10^{-62}	Yes
Tdh3	ORFs	34.0, 5.6	9.0×10^{-272}, 7.9×10^{-16}	Yes
YLL032C	mRNAs	22.7, 13.0	3.7×10^{-112}, 6.1×10^{-19}	Yes
Yra2	mRNAs	11.4, 10.3	2.4×10^{-14}, 1.9×10^{-11}	No

[a] Acceptance Based on Specificity Test.

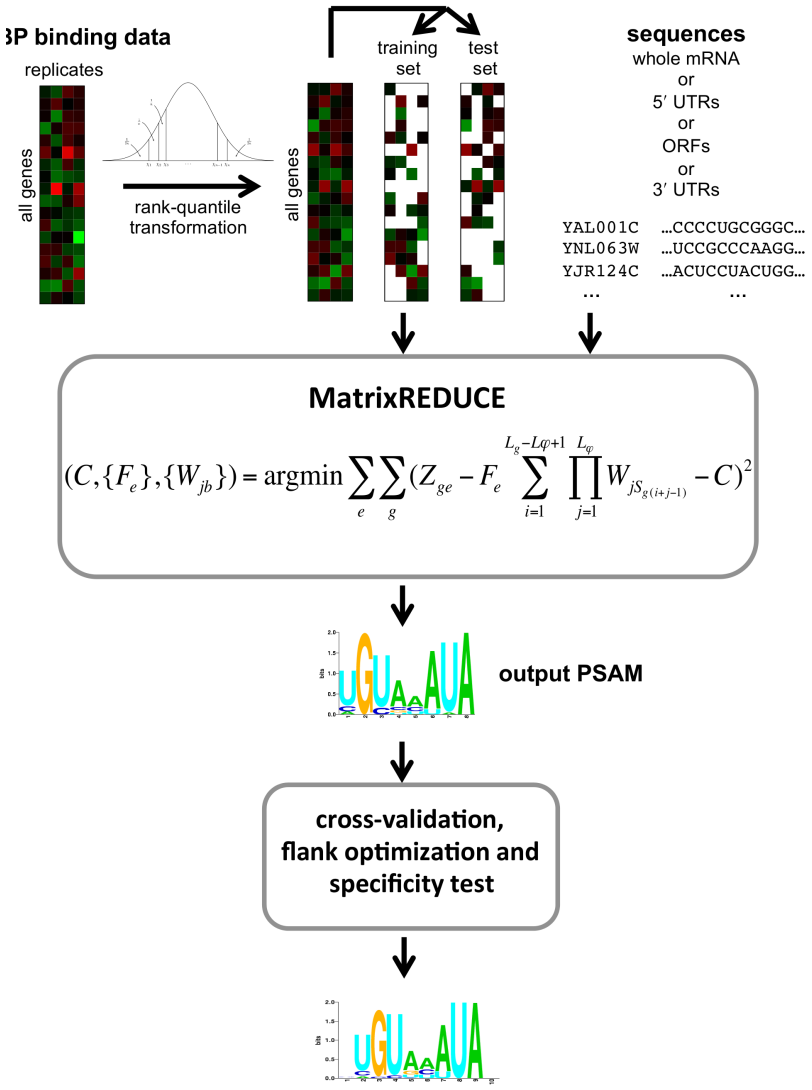

Figure 2.5: Overview of Our Motif Discovery Approach. We applied MatrixREDUCE software to rank-quantile transformed binding data (training set) and mRNA sequences. We repeated this analysis by replacing the full mRNA sequences by 5′ UTRs, ORFs and 3′ UTRs separately. We accepted the PSAMs only if they passed cross-validation and specificity test (exclusive correlation of affinity scores for a RNA-binding protein affinity to its own Binding data).

example yeast Khd1p represses FLO11 post-transcriptionally by binding to its coding region (Wolf *et al.*, 2010). To consider this, we checked the correlation between 132 IP experiment and affinity of each mRNA region separately and only accepted those RBP-region combinations that are significantly correlated to their own RBP binding

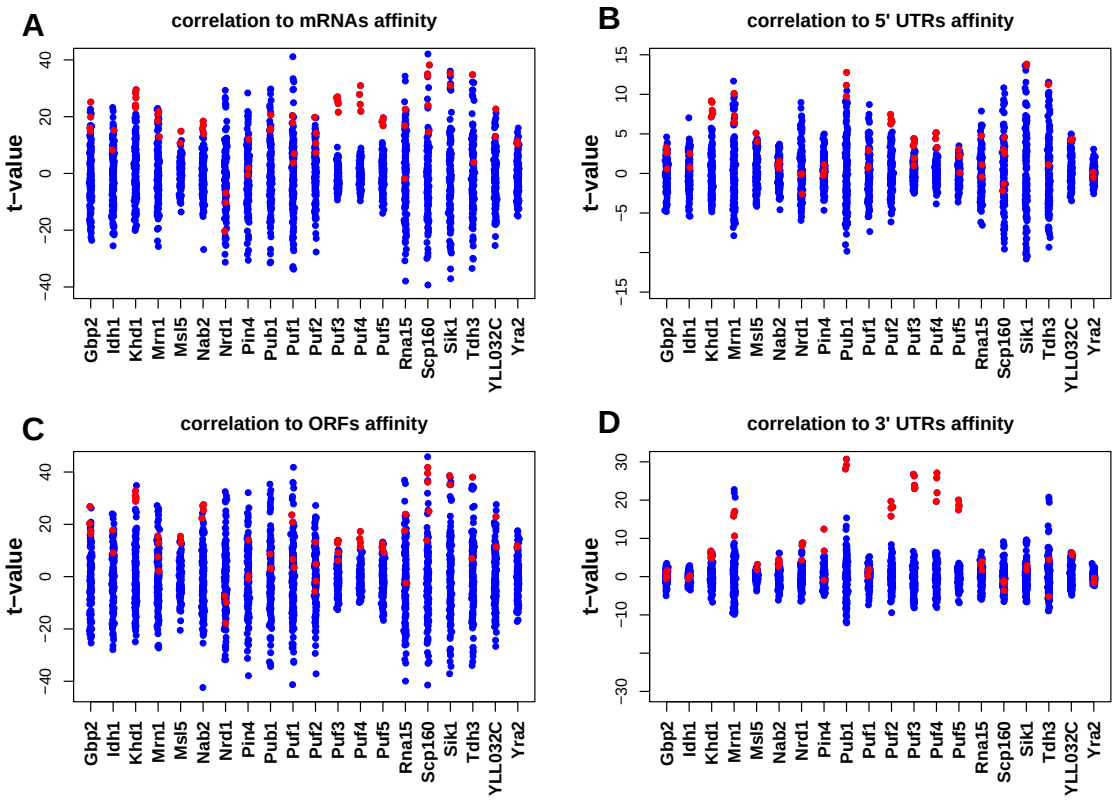

Figure 2.6: Specificty Test for Obtained PSAM affinities. Scatter plot for correlation of RBP binding data to the affinity scores of the 20 PSAMs calculated on (A) full mRNA sequences, (B) 5′ UTRs, (C) ORFs, and (D) 3′ UTRs. The x-axis represents PSAM affinity and each blue dot corresponds to the Pearson t-value of the correlation of PSAM affinity to a particular RBP binding experiment. The red dots indicate the binding measurements that correspond to the labeled RBP on the x-axis. Only the RBPs in combination of mRNA regions that passed specificity test were accepted (i.e. the red dot appeared at the top of the blue dots). There was a total of 25 combinations that were accepted.

experiments. We further narrowed done the accepted set of combinations by selecting only those combinations that showed exclusive correlation of the affinity scores to their binding data (specificity test). **Figure 2.6** displays the results of specificity test for the 20 PSAMs for mRNAs, 5′ UTRs, ORFs and 3′ UTRs separately.

Figure 2.7A shows the PSAM logos for 15 out of 20 PSAMs that passed the specificity test. Among the 20 PSAMs from the training step, 5 did not pass the specificity test. They include Idh1p, Mrn1p, Puf1p, Rna15p and Yra2p (results not shown). The Affinity calculated using these 5 PSAMs were highly correlated to other RBP binding

data. **Figure 2.7B** shows all the 25 RBP-region combinations that passed the specificity test. There was an exception with Scp160p-ORF where the affinity is slightly more correlated to Bfr1p IP experiment (green dots). However Bfr1p is reported to associate with cytoplasmic mRNP complexes containing Scp160p (Lang *et al.*, 2001). We observe that there is a large gap between the relevant IP experiments (red dots) and the rest of the IP experiments (blue dots) for the affinity of 3′ UTR of Pub1p, Puf2p, Puf3p, Puf4p and Puf5p. This indicates that these PSAM are highly specific to the binding data for their RBPs.

2.3.2 Recovered Motifs for RBPs

Out of the 15 PSAMs, 12 PSAMs are consistent with the motifs reported for these RBPs using the same binding data, which is a validation for our PSAM search method (Gerber *et al.*, 2004; Hogan *et al.*, 2008).

The first protein in our PSAM list is Gbp2p. Our discovered motif GRNGNNGR (R is A/G), which is enriched in the ORFs. Gbp2 is involved in mRNAs export from the nucleus to the cytoplasm. The motif reported for this protein by Riordan *et al.* (2011) HGGUGW (H is A/C/U, W is A/U) is compatible with our finding, however our PSAM is more specific.

Khd1p is involved in the asymmetric localization of ASH1 in daughter cells, which is a transcription inhibitor of mating type switch protein encoded by HO gene. Khd1p binds to CNN repeats in coding regions of mRNA /textitin vitro (Hasegawa *et al.*, 2008). In a more recent study by (Wolf *et al.*, 2010) they report enrichment of YCAY (Y is C/U) element in the mRNAs bound to Khd1p.

MSL5p is part of the complex that is responsible for splicing pathway initiation of pre-mRNAs (Abovich and Rosbash, 1997). Another study found that Msl5p binds to branch-point sequence UACUAAC, which confirms the motif we identified for this protein (Garrey *et al.*, 2006).

Another protein in our study is Nab2p which is involved in mRNA poly(A) tail

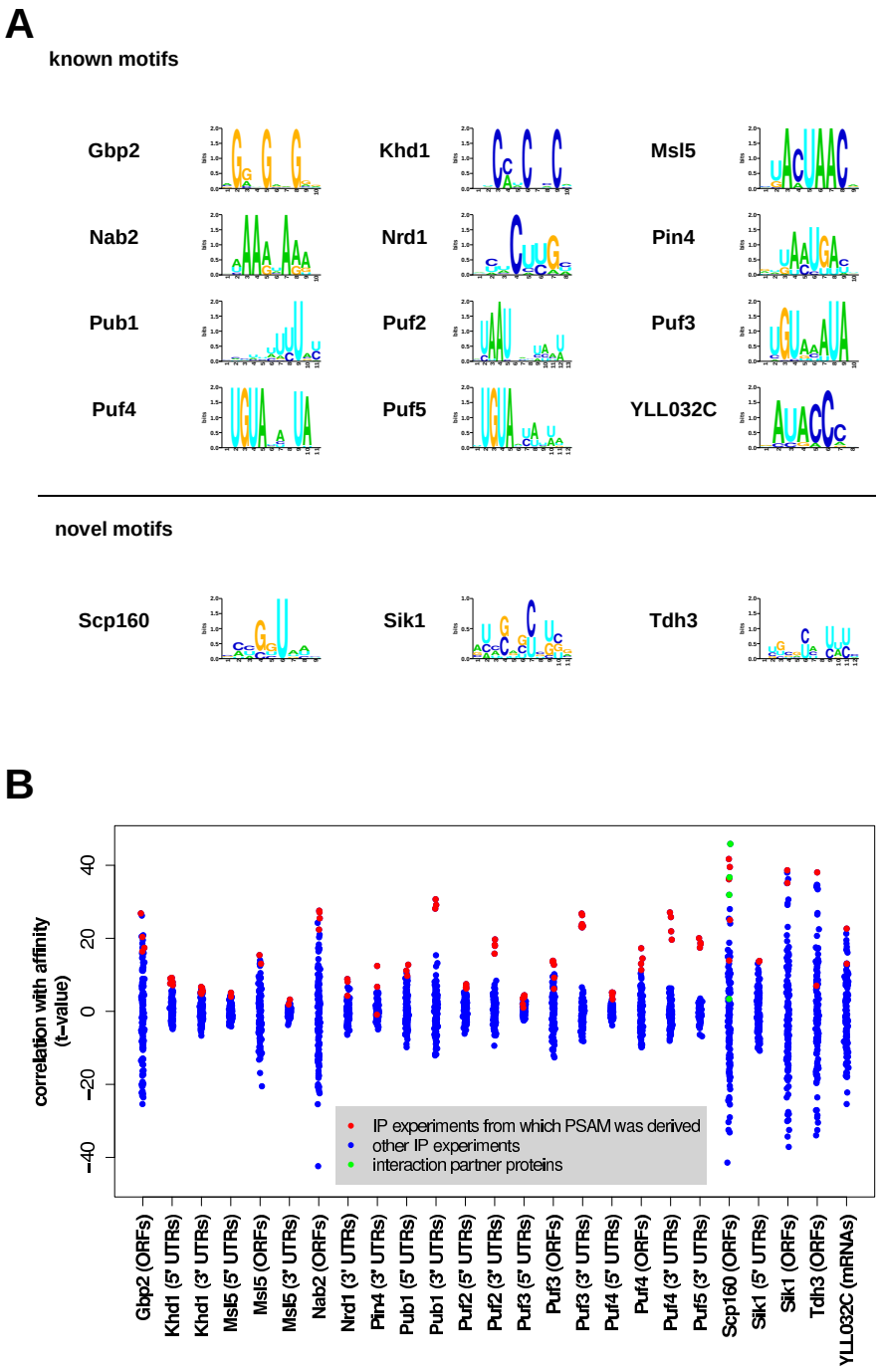

Figure 2.7: List of Known and Novel RBP Motifs Obtained by Our Motif Search Method. (A) List of Known and Novel RBP Motifs Obtained by Our PSAM Search Method. These 15 PSAMs passed cross-validation and specificity tests. (B) Specificity test. Scatter plot for the factors specificity test where the Pearson t-values of univariate linear fit coefficients between 132 RBP binding experiments and 25 selected PSAM-region combinations are presented. Only the factors with at least one self RBP IP experiment t-value (red dots) appearing at the top are shown. The only exception is for Scp160 (ORF) where we have a higher correlation to Bfr1p binding data (green dots). We accepted this PSAM, since Scp160p and Bfr1p are known to interact and are co-imunopercipitated in IP measurements.

formation control and export of mRNA from nucleus to cytoplasm (Hector *et al.*, 2002; Kelly *et al.*, 2010). Our Nab2 element is highly enriched with Adenine, which confirms its tendency to bind to mRNA poly(A) tail.

The PSAM we found for Nrd1p has a core motif CUUG. This protein is subunit of Nrd1p-Nab3p-Sen1p complex, which mediates the termination of small nucleolar RNAs (snoRNAs) (Vasiljeva *et al.*, 2008). Previously it has been reported that Nrd1p binds to GUA[AG] and Nab3p recognizes UCUU (Carroll *et al.*, 2004; Lunde *et al.*, 2011). Since Nab3p and Nrd1p form a complex, it is not surprising that we found a motif that is a partial match to both of the consensus motifs.

The motif we obtained for Pin4p looks similar to the motif reported by (Hogan *et al.*, 2008).

In the case of Pub1p, our motif is a U-rich element. Pub1p is a poly(U) binding protein that is essential for stability of many mRNAs (Anderson *et al.*, 1993; Li *et al.*, 2010; Matunis *et al.*, 1993).

Among the proteins that we were able to capture the binding site nucleotide sequence, there are 4 members of Pumilio/FBF (PUF) homology domain family. Puf2p is one of the members of PUF protein family, where not much is known about its specific physiological role. It interacts preferentially with mRNAs that encode membrane-associated proteins (Gerber *et al.*, 2004). In a recent study, it was shown computationally and experimentally that Puf2p binds to a dual UAAU motif connected through a linker in between (Yosefzon *et al.*, 2011). Our PSAM search algorithm was able to capture the same motif. It is interesting to note that Puf2p binding site is completely different from the consensus UGUA motif that Puf3p, Puf4p and Puf5p bind to (Foat *et al.*, 2005; Gerber *et al.*, 2004; Miller *et al.*, 2008). This is because unlike these 3 proteins, Puf2p has 6 PUM repeats and its amino acid sequence is the most identical to another member of this family, Puf1p. Puf3p binds nearly exclusively to mRNAs that encode mitochondrial proteins (Gerber *et al.*, 2004) and is involved with mitochondrial localization of nuclear-encoded mRNAs (Saint-Georges *et al.*, 2008). Puf3p is also enhances COX17 mRNA degradation by binding to UGUR-AUA on

its 3′ UTR (Olivas and Parker, 2000). Puf4p is known to bind to UGUAUAUUA site on 3′ UTR of HO endonuclease mRNA and together with Puf5p, they negatively regulate this mRNA (Hook *et al.*, 2007; Miller *et al.*, 2008). Puf4p is also known to bind preferentially to mRNA encode ribosomal proteins (Gerber *et al.*, 2004). The binding sites we found for Puf3p, Puf4p and Puf5p all in agreement with the reported motifs in the works mentioned earlier.

YLL032C is an un-annotated protein that may interact with ribosomal complexes (Fleischer *et al.*, 2006). Our algorithm found a motif AUACC as reported by (Hogan *et al.*, 2008).

Recovering these motifs for 12 RBPs acts as positive controls for our approach. It indicates that our method successfully can detect the composition of RBPs binding sites using genome-wide binding data without depending on training on the target set.

2.3.3 Novel Binding Motifs for Scp160p, Sik1p and Tdh3p

Our method detected 3 novel binding sites for the RBPs including Scp160p, Sik1p and Tdh3p. To further validate these new motifs, we calculated the Pearson test t-values for correlation of the affinity scores of the 25 RBP-region combinations to the expression data for 173 different stress conditions (Gasch *et al.*, 2000). We also performed Gene Ontology (GO) scoring analysis on the *in vitro* affinity scores using Wilcoxon-Mann-Whitney (WMW) test.

Scp160p is a RNA binding protein involved in mating response pathway (Guo *et al.*, 2003). It contains multiple heterogeneous nuclear ribonucleoprotein K-homology (KH) domains[3]. Using the stress condition expression data, Scp160 affinity for ORF sequences is highly anti-correlated to YPD stationary phase, YPD, nitrogen depletion and heat shock conditions with t-values about -14, -14, -10 and -7 respectively. In con-

[3]KH domain is a protein domain that binds RNA and single stranded DNA. This domain is about 70 amino acids long.

trast the positively correlated conditions were: cold shock (t-value ~ +10) and hypo osmotic shock (t-value ~ +9). Pseudohyphal differentiation, a filamentous growth form of the budding yeast overlapping partly with mating pathway, is induced by nitrogen starvation (Gimeno *et al.*, 1992). GO analysis based on the mRNA affinity scores for Scp160p showed enrichment for nitrogen compound metabolic process (p-value = 4.16×10^{-9}).

Sik1p (Nop56) is component of the box C/D snoRNP complexes that direct 2′-O-methylation of pre-rRNA during its maturation. Sik1p affinity score on both 5′ UTRs and ORFs. We observed positive correlation of ORF affinity scores to the YPD stationary phase growth (t-value ~ +6) and also significant correlation to heat shock stress conditions (t-value ~ +7). Spb1p is a nucleolar AdoMet-dependent methyltranferase also involved in rRNA processing (Kressler *et al.*, 1999). Spb1p is homologous to E. coli Ftsj/Rrmj heat shock protein, which also acts as 2′ O-methyltransferase (Bugl *et al.*, 2000). It could be that Sik1p also has a direct or indirect role in rRNA methylation regulation under heat shock and that is why we observed up-regulation of Sik1p activity in the heat shock conditions. GO analysis shows significant negative correlation to ribosome and rRNA related categories.

Tdh3p encodes Glyceraldehyde-3-phosphate dehydrogenase, which is required during gluconeogenesis and is essential for yeast cells to grow on non-carbohydrate sources such as ethanol and glycerol (McAlister and Holland, 1985). This enzyme is also found in cytoplasm and cell wall (Delgado *et al.*, 2001). Using stress condition data, we found that Tdh3p is up-regulated when cells are exposed to menadione, a synthetic nutritional compound, (t-value ~ +5) and down-regulated in the presence of sorbitol, a type of sugar that is naturally found in some fruits, (t-value ~ -6) and in nitrogen depleted conditions (t-value ~ -5). The GO scoring analysis for the affinity scores for this factor showed these categories as significantly enriched: intrinsic to membrane (p-value = 1.79×10^{-79}), thiolester hydrolase activity (p-value = 4.44×10^{-8}), glucosyltransferase activity (p-value = 8.92×10^{-7}) and glycerophospholipid metabolic process (p-value = 8.24×10^{-7}).

The GO enrichment analysis results for all the three novel motifs are consistent with the functional validation by stress condition expression data.

2.4 Conclusion

Our motif discovery approach is based on biophysical modeling of the binding of RBPs to target RNAs. It detects potential regulatory elements within RNA sequences that are recognized by diverse RBPs. Our algorithm searches for binding sites in the form of sequence specific affinity matrices (PSAMs). Most approaches either impose a threshold to filter RBPs binding data or use gene expression data in combination with mRNA half-lives to identify stability motifs associated with RBPs. Measuring mRNA half-lives requires transcription arrest, which can interfere with the post-transcriptional control of mRNAs under study (Grigull *et al.*, 2004). Hence, the interpretation and usage of mRNA half-lives should be done cautiously. By contrast, our model is not based on defining a target set or mRNA half-lives.

The biophysical model our method that is based on is only strictly valid in the low protein concentration regime. In other words, for the cases where RBP concentration is much smaller than dissociation constant (K_d). This may not be completely correct for some RBPs. Still, the PSAMs discovered for 12 RBPs agree with previously reported consensus motifs in other studies. In addition, we discovered three novel motifs for Scp160p, Sik1p and Tdh3p. The functional validation results from GO enrichment analysis and condition-specific genome-wide mRNA expression data suggest that these novel motifs could indeed be the binding site for Scp160p, Sik1p and Tdh3p. Experimental follow up are required for further validation of these new findings.

Chapter 3

Novel Method for Mapping *Trans*-Acting Loci

3.1 Introduction

The DNA sequence of a typical gene usually varies from one individual to another even between the members of a genetically related population. Knowledge of genetic variation among these individuals helps us understand why some of the members are more susceptible to a disease or why some are more responsive to a particular treatment. It can also be used to determine for example which genes influence product yield in crops or why some are more tolerant to environmental stresses. The central goal of linkage studies is to statistically associate an observed phenotype to the genotype (i.e. genetic sequence) among genetically related individuals. The pheynotype can be any quantitative trait measurable for every individual in the population such as: morphological characteristics, chemical level of compounds in the blood/tissues, tumor count, or even the mRNA abundance of each gene in the cells/tissues of the members of the population under study. The last one is possible courtesy of the advances in high-throughput mRNA expression profiling techniques. In recent years, a growing number of studies have carried out linkage analysis to identify causal loci

for human genetic diseases such as: Alzheimer disease (Schellenberg *et al.*, 1991), Schizophrenia (Schwab *et al.*, 1995), breast cancer (Smith *et al.*, 2006), pancreatic cancer (Klein *et al.*, 2007), asthma (Celedon *et al.*, 2007), hypertension (Guo *et al.*, 2012) and bipolar disorder (Badner *et al.*, 2012). This approach has also been applied to identify genes influencing the crops yield or animal products quality (Aslam *et al.*, 2011; Heidari *et al.*, 2011).

Combining quantitative trait profile data with the added information from natural genetic sequence variation in the segregating population, one can link the observed phenotype to influential genes or gene products, encoded at specific as genetic loci. This approach is called quantitative trait loci (QTL) analysis. From gene sequence to mRNA level, there are many layers of regulation involved such as chromatin state, transcriptional rate, mRNA processing, localization and stability as discussed in **Section 1.3**. One can still simply apply QTL analysis to discover the causal genetic variations manifesting in mRNA expression differences among the members. This approach is called expression QTL or eQTL. If a locus regulates the expression of a large number of genes, then it is considered to be an eQTL "hotspot".

In this chapter we will first explain the common eQTL approach and discuss its shortcomings. We will then present our eQTL hotspot detection approach. Finally we will discuss the result of applying this method to *Saccharomyces cerevisiae* (baker's yeast) and *Caenorhabditis elegans* (roundworm). To validate our method, we compared our results for yeast to the eQTL hotspots reported by two independent studies (Brem *et al.*, 2002; Zhu *et al.*, 2008). Our findings overlapped with 70% of the loci from the mentioned studies. Our analysis also discovered a new locus on chromosome V. For worm data, we recovered 2 loci and discovered 4 new loci by comparing our results to (Rockman *et al.*, 2010).

3.1.1 eQTL Approach

Studying the cellular state of a single organism (RNA and protein levels) in different biological conditions tells us about genes critical for survival in those specific condition. Studying cellular states of many related organisms simultaneously, can unravel small genetic perturbations manifesting a continuous spectrum of traits (Jansen and Nap, 2001). As we mentioned earlier, linkage studies identify causal loci within a set of genetically related individuals (samples) with respect to one or both parents (reference). Mating results in the segregation of the parental genetic material passed to the offsprings or segregants. When two haploid cells from two different yeast strain mate, four segregants or spores are generated. Each of the spores will have a unique genotype, which is the result of the recombination between the parental chromosomal material. The different allele combination along the chromosomes causes a perturbation in the gene expression compared to the parental strains. Since the genotype of each segregant is different, the gene expression pattern is unique for each of the segregants. **Figure 3.1** explains the traditional approach for eQTL detection.

The traditional approach for eQTL detection is to split the segregants at each marker based on their inherited parental allele type at that marker and then compare the differential mRNA expression levels distribution of a gene between the two subsets. This approach basically relies on the F-statistic of analysis of variance (ANOVA). Here the differential mRNA level of each gene is considered as an individual trait and separately tested for linkage to the chromosomal marker. A common way to calculate a significance threshold to distinguish significant linkages is to permute the expression levels of each genes to obtain the linkage to all the markers. This is done usually at least N=100 times to get the empirical null distribution for linkage of each gene to the markers. Then the FDR corresponding to a F-statistic threshold is obtained as the ratio of the number of linkages above the threshold averaged over the N shuffled data to the number of linkages for the actual data. Detection of eQTL hotspots is then based on a clustering procedure that identifies loci with many significant eQTL linkages.

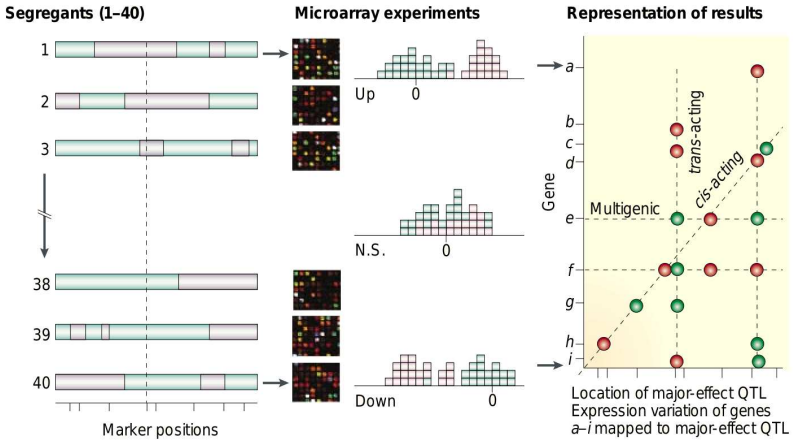

Figure 3.1: The Common Approach for the Detection of Expression Quantitative Trait Loci (eQTL). Here the population of 40 segregants (samples) is obtained by crossing two parents (reference). The DNA material of each segregant is sequenced and a genotype map indicating the type of the parental allele inherited at each marker location on chromosomes is obtained (cyan and pink). These different alleles are detectable based on the natural polymorphism existing in the genetic sequences of the parents. The differential mRNA levels of each segregant relative to the reference are measured using microarray experiments. These relative mRNA levels are then treated as quantitative trait. The segregants mRNA levels of each gene is then split into 2 subsets based on their inherited allele at each chromosomal marker. By performing eQTL analysis the markers for which mRNA levels of the 2 subset is significantly different (upregulated or downregulated) distinguished from those with no significant change (N.S.). The plot on the right is used to present the results. The horizontal and vertical axes represents chromosomal marker locations and genetic location of genes significantly affected by eQTLs, respectively. Red dots indicate upregulation and green downregulation of gene expression. A gene can be affected by multiple eQTLs (horizontal dashed line), by distal eQTL (off diagonal dots) or local eQTL (diagonal dots). Figure from Jansen (2003).

A shortcoming of this approach is that it is not capable of detecting hotspots that marginally regulate the expression of large number of genes. Our method is able to detect such loci. In our method we calculate a single χ^2-statistic that integrates the genome-wide differences in the expression levels of the two segregant subset. It does not need to consider whether the genes are upregulated or downregulated. In summary, our method is able to identify eQTL hotspots based on the collective regulatory effect of each locus on a large number of genes.

One way to validate the detected QTLs is an experimental approach based on allele replacement between the two parental genotype background. Each detected hotspot marker usually contains between 10 to 100 genes. Hypothetically for each of the genes located in a detected chromosomal marker region one could swap the two alleles between the two parents and measure the genome-wide expression profile. By analyzing the effect of the incorporated allele, the causal gene or genes could be identified.

The power of the any eQTL detection analysis depends on the genotype map resolution and population size. In other words, the higher the density of single nucleotide polymorphism (SNPs) locations, and the larger the population, the more precisely the eQTLs are detected. This means pinpointing the causal gene will be simpler. However, there exists a biological limitation known as linkage disequilibrium. Linkage disequilibrium refers to the dependence of inheritance of alleles at nearby loci. It can be overcome to some extent by a larger number parental crossings in non-human populations, which will increase the rate of cross-over between these loci.

3.2 Methods

3.2.1 Experimental Data Used

We performed this analysis for two different organism: *Saccharomyces cerevisiae* (baker's yeast) and *C. elegans* (roundworm).

For yeast we used differential mRNA expression data for two parental strains: a standard laboratory strain (BY) and a wild isolate from a California vineyard (RM), as well as 112 segregants from a cross between these parental strains, collected during mid-log phase growth in rich (YPD) media performed by Brem *et al.* (2002). Expression levels were measured for 6215 genes using a two-color microarray experiment, with the same BY mRNA sample being used as a reference for all experiments (i.e. $\log_2$(sample/BY)). For each sample two experimental replicates were performed with dyes swapped. We used the average of the two log-ratios for each gene. Following Brem *et al.*, we excluded ORFs rejected by (Kellis *et al.*, 2003). We also averaged log-ratios for 13 ORFs that were spotted twice. Finally, we normalized each array by subtracting the mean log-ratio across all genes. Genotyping of the segregants was performed using oligonucleotide arrays at a total of 2957 independent markers along 16 chromosomes by (Brem *et al.*, 2002). **Figure 3.2** displays the experimental cross between the two strains.

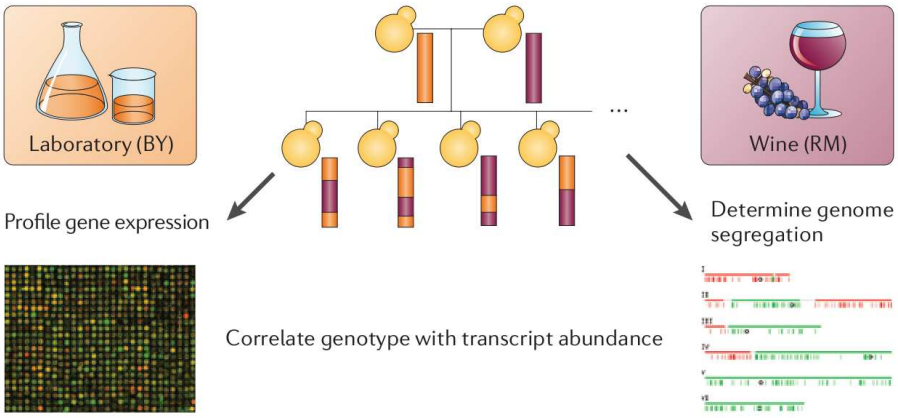

Figure 3.2: Experimental Data for Yeast eQTL Hotspots Analysis. Here two strains of yeast, a laboratory strain (BY) and a wild isolate from a vineyard (RM), were crossed to obtain 112 segregants. The differential mRNA levels for each segregant relative to the BY parental strain were obtained using microarray experiments. Genotyping was performed at 2957 markers along the 16 chromosome of yeast. Figure from Rockman and Kruglyak (2006).

For worm we used mRNA differential expression data collected by (Rockman *et al.*, 2010). The data included RNA abundances of synchronized young adults of 214 recombinant inbred advanced intercross lines (RIALs) from a biological cross between

a laboratory strain (N2) and a wild isolate from Hawaii (CB4856). All microarray experiments were done with the same reference pool (mixed growth stage and mixed N2 and CB parental strains) for a total of 14,792 distinct genes. The genotypes were mapped for 1455 chromosomal markers along chromosomes 1 through 5 and X. Using RIALs supposedly will reduce linkage disequilibrium and increase the recombination rate (Darvasi and Soller, 1995). To develop RIALs, first the parental lines were crossed to create an F_1 generation consisting of both male and hermaphrodite progeny. F_2 individuals were obtained by performing the four possible crosses of F_1. The same procedure was used to generate F_3 progeny from F_2. After these steps, F_3 gametes were crossed by eight breeding designs (Rockman and Kruglyak, 2008). Random mating was continued until F_{10}. To expand the population, two lines were derived from each plate containing tenth-generation hermaphrodites. Each of the lines was then propagated by selfing a randomly selected hermaphrodite for each of 10 generations, for a total of 20 generations starting from the parental strains. **Figure 3.3** depicts this procedure.

3.2.2 Pre-Processing of the Expression Data

We calculated z-scores of expression data. For each gene, we calculated the mean and standard deviation among segregants. We then calculated z-score using the equation below.

$$z_{gs} = \frac{A_{gs} - \mu_g}{\sigma_g} \tag{3.1}$$

Here A_{gs} refers to the mRNA $\log_2$-ratio of gene g for segregant s, and μ_g and σ_g stand for the mean and standard deviation of the mRNA levels of gene g among the segregants, respectively. This transformation is necessary because our eQTL detection is based on χ^2-statistic (as explained in the next section).

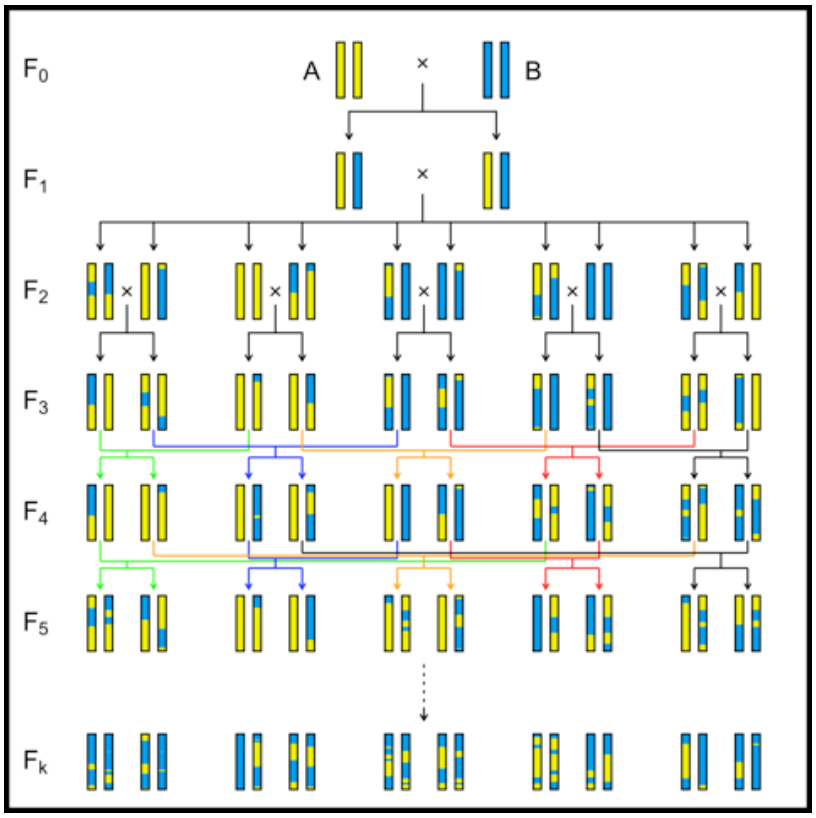

Figure 3.3: Recombinant Inbred Advanced Intercross Lines for (RIALs). This procedure was used for C.elegans data starting from two parental lines represented as A and B. Random mating was performed among the individuals of F_i to obtain F_{i+1} generation. This was done until reaching F_{10} generation. From then on until F_{20} the population was propagated by self-fertilization. Figure from Genetics Society of America website (GSA).

3.2.3 χ^2-statistic Analysis

In this section we will first discuss the χ^2-distribution and then its application for our eQTL analysis.

The χ^2 distribution derives its importance from its relation to the sum of squares of standard Gaussian distributed random variables ($\mu = 0$ and $\sigma = 1$). Given N independent standard Gaussian random variables x_i the sum of squares

$$Q = \sum_{i=1}^{N} x_i^2 \tag{3.2}$$

has a χ^2-distribution with N degrees of freedom. This makes a χ^2-distribution a gamma distribution with $\theta = 2$ and $\alpha = N/2$. Here the null hypothesis is that x_i's are sampled from N random normal variables. Note that the expected value $\langle Q \rangle$ is equal to N, the degrees of freedom.

We want to use χ^2-statistic to test whether the z-scores show a pattern of coherent variation among segregants. To this end, let us formulate the quantity that we calculate χ^2-statistic for. At each marker m, we split the segregants based on the allele (A or B) inherited at that position. We then calculate the sum of the difference of z-scores between the two subsets, normalized by square-root of sum of the population of the two subsets. Note that the denominator is not necessarily equal to total population size, since there can be missing data points. The equation below defines ΔZ for each marker and each gene.

$$\Delta Z_{gm} = \frac{\sum_{\{s\}_{A@m}} z_{gs} - \sum_{\{s\}_{B@m}} z_{gs}}{\sqrt{N_{A@m} + N_{B@m}}} \tag{3.3}$$

Then Q_m for each marker is:

$$Q_m = \sum_{g=1}^{G} \Delta Z_{gm}^2 \tag{3.4}$$

with G equal to the total number of genes. We calculate Q_m independently for each marker. To check for statistical significance, we calculate a p-value for each Q_m-value. To do this we use the χ^2 distribution function with G degrees of freedom as shown in **Figure 3.4**. Here we have a multiple testing situation with M markers. To correct for it, we use the Bonferroni correction and only accept those markers as significant that have p-values smaller than $0.01/M$.

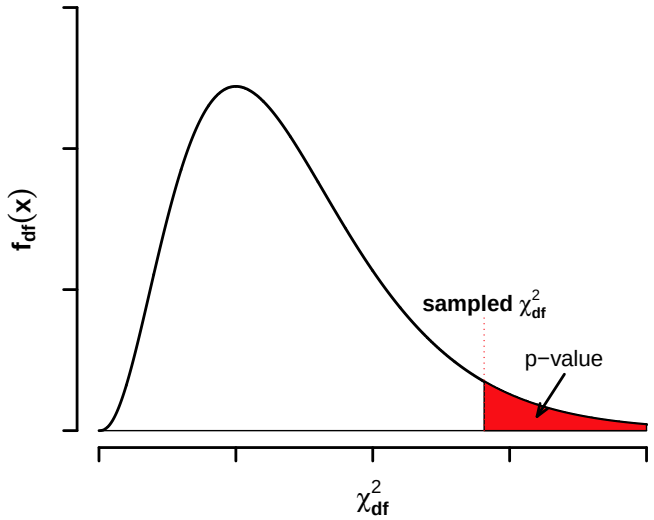

Figure 3.4: p-value Calculation of a Sampled χ^2 Value. Here $f_{df}(x)$ represents χ^2 probability distribution function with df degrees of freedom. The p-value of a given χ^2 is equal to the area under the upper tail of the distribution colored in red.

3.2.4 Forward Selection of Peaks for χ^2 Profile

Due to linkage disequilibrium, fine mapping of causal QTLs is difficult. It requires a higher genetic marker recombination rate in the population to break the association of closely linked markers. To detect the causal markers, we developed an algorithm to iteratively seek the significant markers by removing the effect of selected markers at each iteration. The steps are shown in **Algorithm 1**. Starting from the χ^2 vector Q_m across all markers, we selected the marker m^*_{new} with the largest χ^2 value. This marker is added to the set of selected markers $\{m^*\}$, which is initially empty. We then

remove the effect of this marker from all the other markers by calculating the residuals of a linear regression across all genes of ΔZ_{gm} on ΔZ_{gm^*}. Using the residuals e_{gm}, we calculate a new Q value for each marker m to be used in the while loop condition at the beginning of the next round. At each iteration, we calculate the residuals of the original ΔZ values by performing a multiple regression on the $\Delta Z_{g\{m^*\}}$ values for the set of selected markers $\{m^*\}$. This procedure is repeated until Q_m is no longer significant for any marker. We used Bonferroni correction to determine the significance threshold at an expected number of positives equal to 0.01.

Algorithm 1 Algorithm for Forward Selection of Significant χ^2 Peaks

while $\max(\chi^2) >$ threshold **do**
  $\{m^*\} = \{\{m^*\}_{\text{previous}},\ m^*_{\text{new}} = \underset{m}{\text{argmax}}\ \chi^2_m\}$
  for all m **do**
    $(C^*, F^*) = \underset{C,F}{\text{argmin}}\ (\sum_g (\Delta Z_{gm} - C - F \cdot \Delta Z_{g\{m^*\}})^2)$
    $e_{gm} = \Delta Z_{gm} - C^* - F^* \cdot \Delta Z_{g\{m^*\}}$
    $\chi^2_m \leftarrow \sum_{g=1}^{G} e^2_{gm}$
  end for
end while

3.2.5 Gene Ontology Enrichment Analysis on Selected Peaks

By using GO enrichment scoring it is possible to relate the ΔZ's to the underlying transcriptional program, cellular state or cellular component (see **Section 2.2.6**). We used GO categories to test whether the ΔZ's of selected markers are associated with a specific biological pathway or not. We performed GO enrichment analysis on the ΔZ_{gm}'s for the selected markers $\{m\}$. For each selected marker, we apply the non-parametric Mann-Whitney-Wilcoxon test to determine whether the ΔZ values for genes within a particular GO category have a different distribution than the ΔZ's for all other genes. We performed a Bonferroni correction on the resulting p-values accepting only categories with p-values smaller than 0.01/N where N is the number of unique GO categories with at least 10 genes. We used an iterative procedure for removing the effect of redundant nested GO categories which was implemented

originally in the T-profiler algorithm (Boorsma *et al.*, 2005). We also used Student's t-test to verify whether genes enriched in GO categories were significantly upregulated or downregulated for a specific parental allele based on the t-value sign.

To perform the GO enrichment analysis, we downloaded packages *GO.db*, and then *org.Sc.sgd.db* for yeast and *org.Ce.eg.db* for worm from the Bioconductor website (`http://www.Bioconductor.org`) within the R statistical programming environment.

3.2.6 Correlation to Known Transcription Factors Binding Specificities for Yeast

Besides GO enrichment analysis, we used another information source to understand the biological significance of the selected genetic loci. The idea is that if a polymorphism at a marker with significant eQTL signal affects the transcriptional regulation by a specific DNA-binding transcription factor (TF), then the ΔZ for that marker is expected to correlate to that TF's promoter binding preference across all genes. To do this, we sought for significant correlation between the ΔZ_{gm} of the selected markers and the promoter specificities of the transcription factors.

In order to calculate the promoter specificities, we first obtained the genetic sequence of the parental strains BY from *Saccharomayces cerevisiae* genome database (SGD; `http:://www.yeastgenome.org`). We used 600 base pairs upstream sequences of coding region start site of each gene. Next, we obtained a collection of 124 position weight matrices (PWM) representing binding preferences of yeast TFs from MacIsaac *et al.* (2006). The elements in the PWMs represent the information about the nucleotide frequencies at each position in the set of target DNA binding sites. Together with the promoter sequences specific to each segregant-genotype at selected markers, we calculated sequence specificities. This step is explained in detail in (see **Equation 2.7**). Finally we used a multiple regression on all genes between each selected marker ΔZ and the specificities of the set of TFs.

We used the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to correct

for multiple testing at 1% false discovery rate. The Benjamini-Hochberg method tries to estimate the expected fraction of false positives based on the data size and p-values. Within a group of results ranked by increasing p-value: $p_1 \leq p_2 \leq \cdots \leq p_m$, for a desired false discovery (FDR) rate at q, k is the largest i for which

$$p_i \leq \frac{i}{m} q \tag{3.5}$$

In our case m is the number of chromosomal markers and q is equal to 0.01. All results with p-value larger than p_k were considered insignificant. This method is less stringent that Bonferroni multiple testing correction.

3.3 Results

We applied our eQTL hotspot analysis to two organisms, *Saccharomyces cerevisiae* (baker's yeast) and *Caenorhabditis elegans* (roundworm). For yeast, we used genome-wide mRNA expression data for 5423 genes from a study performed by Brem *et al.* (2002). The dataset included the expression data for two parental strains, a laboratory strain (BY) and a wild isolate from a vineyard in California (RM), as well as segregants from the parental genetic cross. The data included 6 biological replicates for BY strain, 12 for RM strain and 1 replicate for each of 112 segregants. The microarray measurements were done relative to BY strain mRNA levels. The same study also provided the genotype map of 2956 markers for all of the 130 yeast samples. For worm we used expression data from Rockman *et al.* (2010). The data contained the RNA abundances of synchronized young adults of 214 recombinant inbred advanced intercross lines (RIALs) from a biological cross between a laboratory strain (N2) and a wild isolate from Hawaii (CB4856). All microarray experiments were done with the same reference pool (mixed growth stage and mixed parental strains) for total of 14,792 distinct genes. The same study also determined the genotype at 1455 markers for each of the lines.

Figure 3.5 depicts an overview of our approach. Our method requires the expression data and genotype data as inputs. The expression data is represented as a matrix whose rows correspond to genes and whose columns contain the genome-wide mRNA expression profile for one of the segregants. The genotype data is displayed as a matrix containing binary information, whose rows correspond to genetic markers and whose columns provide the inherited parental allele for particular segregant. The first step involves normalization of the expression data for every gene (i.e. row) across the segregants. We will explain the importance of this normalization later in this section. For each marker, we split the segregants based on the genotype profile of that marker and calculate the difference of the sums of the normalized expression data at that marker for the respective alleles. Once we have built the matrix of differences for each gene and each marker, we can calculate the χ^2-statistic for every marker by summing the square of elements of this marker over all genes (see Methods). A large value of the χ^2-statistic for a particular marker indicates the presence of differential expression that is coherent across a subset of the genes, without the need to specify any specific genes or gene sets. Whenever the χ^2 value derived from the signature of differential expression between the two sub groups of segregants defined by a split based on the genotype data at a specific marker is statistically significant, it implies a broad trans-acting effect across many genes driven by the allelic variation at that locus.

The initial step, which involves normalization of the mRNA expression levels for each gene across the segregants, is important in two regards. First, since our method uses χ^2-statistic to test the significance of each marker as a *trans*-acting locus, we are mathematically required to sum the squares of standard normalized random variables. More importantly, we are interested to detect the loci that regulate gene expression through the added value of genotype segregation. So the eQTL detection is not based on the comparison of the expression levels between the genes, but rather it is the comparison of the expression levels between the segregants. Thus, it is important to set the variance of the expression level distribution of each gene among the segregants to one.

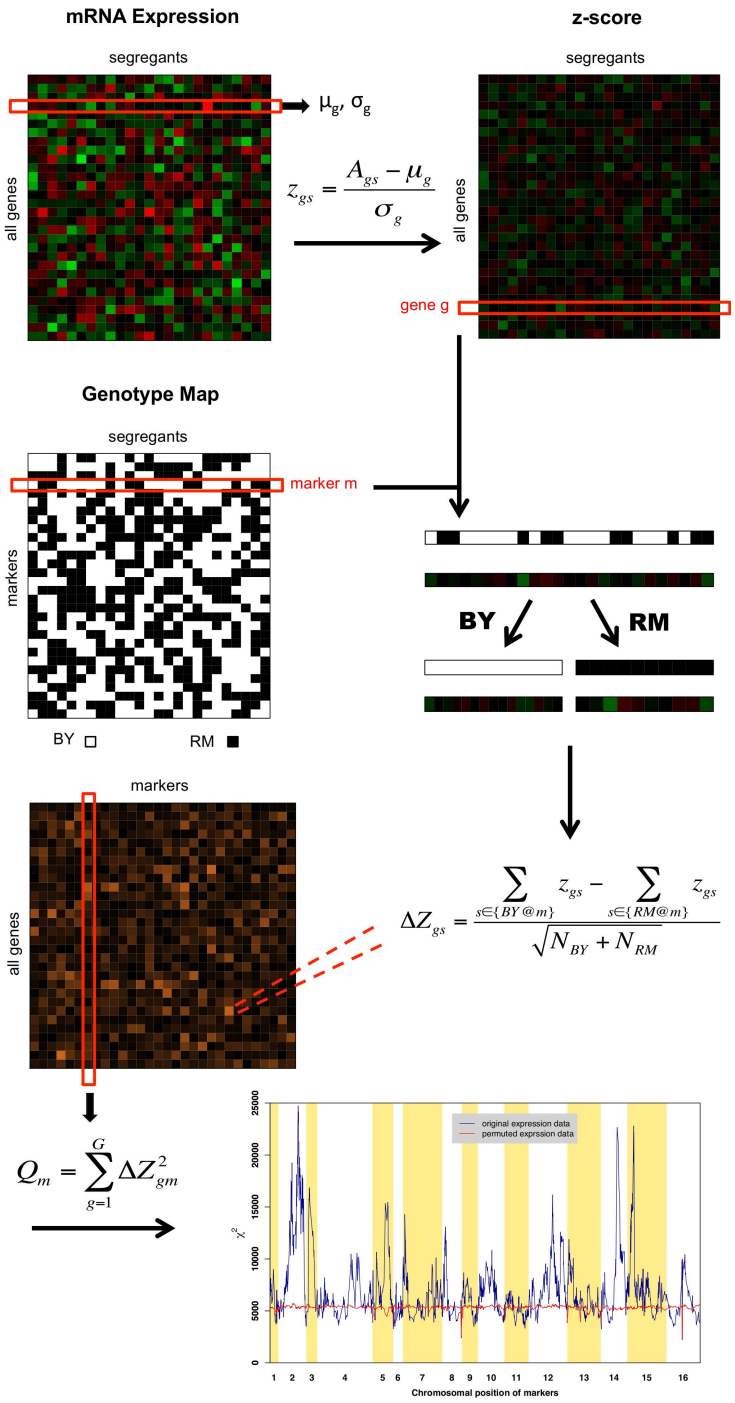

Figure 3.5: Overview of Our eQTL Hotspot Detection Approach. First the expression data is converted to z-scores per each gene (i.e. row). Next for each chromosomal marker m, the zscores of each gene g is split based on the inherited allele at that marker among all segregant into two subsets. The sum of the difference of the zscore of the two subsets is calculated for each combination of $(g,\ m)$, which builds the matrix of ΔZ. The last step involves the calculation of Q_m at each marker m and plot it along the markers chromosomal position. Q_m is expected to behave like

Figure 3.5: (Cont. Caption) χ^2-distribution. Markers with large Q_m value indicate that the cumulative effect of the genotype at m on the gene expression is significant.

The final step involves the detection of influential markers. Linkage disequilibrium reduces the resolution of any QTL methods. We performed an iterative procedure that is based on forward selection (see Methods). In the next sections we will discuss the results of our analysis applied to yeast and worm.

3.3.1 Recovering 70% of the Previously Reported And Discovering a New eQTL Hotspot for *Saccharomyces cerevisiae*

Figure 3.6 presents the χ^2-statistic profile for the 2956 markers along the 16 chromosomes of yeast both for the real expression data (blue) and null data (red). The null data was generated by shuffling the expression levels of each gene among the 112 segregants. We observe that in the case of the randomized data, the χ^2 values oscillate around the expected value equal to the degrees of freedom (df) of the χ^2 distribution. In our case df is the number of genes equal to 5423. However, when the original expression data are used, we observe a distinct profile compared to the null data. The peaks represent putative eQTL hotspots.

We applied a forward selection procedure to detect the influential loci. **Figure 3.7** displays the iterations of our marker selection procedure. Linkage disequilibrium reduces the resolution of marker detection by widening the significant peaks. It can also manifest itself between distal loci. To improve the QTL hotspot detection accuracy, our method attempts to reduce the influence of the detect significant loci at each iteration round (see Methods). The black arrows in **Figure 3.7** mark the location of the newly selected significant marker at each iteration. We detected a total of 11 markers as significant with Bonferroni correction at 1% corresponding to χ^2-threshold $\sim$ 5904. Detailed information for the selected peaks is summarized in

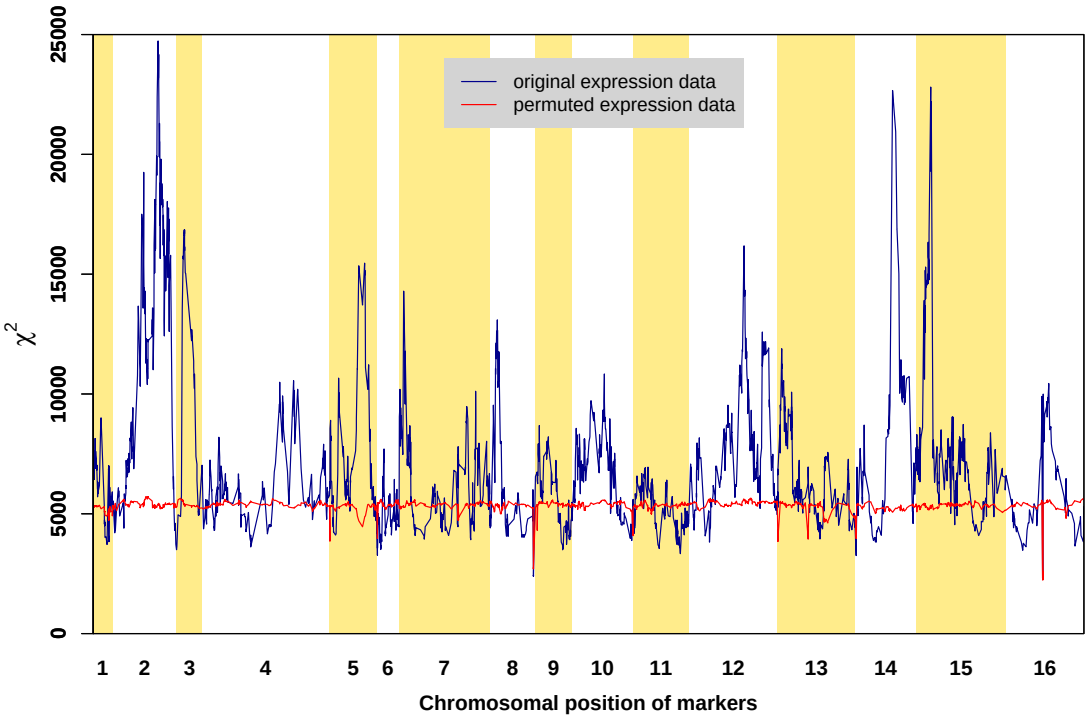

Figure 3.6: eQTL Hotspots Peak Profile for Yeast. We calculated the allele specific difference of the sums of expression data z-scores at each chromosomal marker (ΔZ) and then calculated χ^2 for all the 2956 markers. The blue plot resulted from using expression data and the red was for the case where we permuted the expression vector of each gene randomly among the segregants while keeping the genotype map intact. Then we calculated ΔZ using the permuted expression data z-scores. It serves as the null distribution for the χ^2 values. It oscillates around the expected value $\sim$ 5,400 (i.e. the number of genes).

Table 3.1.

To validate our findings, we compared our results to the results from two independent studies, Brem *et al.* (2002); Zhu *et al.* (2008), where two different eQTL approaches were applied to the same dataset employed in our analysis. The first study applied the traditional eQTL method, which treats the expression data directly as a quantitative trait and detects the significant eQTLs based on analysis of variance (ANOVA) F-statistic at each marker and finally clustering the results to identify the eQTL hotspots. The second study is based on Bayesian network construction to identify the hotspots. **Figure 3.8** presents the results comparison. Our method was able to capture 70% of the reported loci by one or both of the mentioned studies. This success establishes the strength and accuracy of our method for eQTL hotspot detection. It also increased our confidence to apply our approach to another organism as well as incorporate it our final project explained in Chapter 4. In addition, our method was able to detect a novel locus on chromosome V that was not reported by either of the two studies.

3.3.2 Assessment for Possible Regulatory Roles for the Detected Yeast eQTL Hotspots

In order to understand the biological function underlying these eQTL hotspots we employed two strategies, correlation to transcription factors (TFs) binding preferences and gene ontology (GO) enrichment analysis. We assessed whether the difference of expression levels between the two subgroups of segregants, ΔZ, for the detected loci are associated with any known TFs binding preferences. If the ΔZ's at one of the candidate eQTL hotspot markers significantly correlate with the binding specificities of a TF or a group of functionally-related TFs, then this implies the possibility of the involvement of these TFs in the regulation of the expression of the genes linked to this locus. For the second assessment, we checked whether the ΔZ's are enriched for a specific biological process, molecular function and cellular component based on

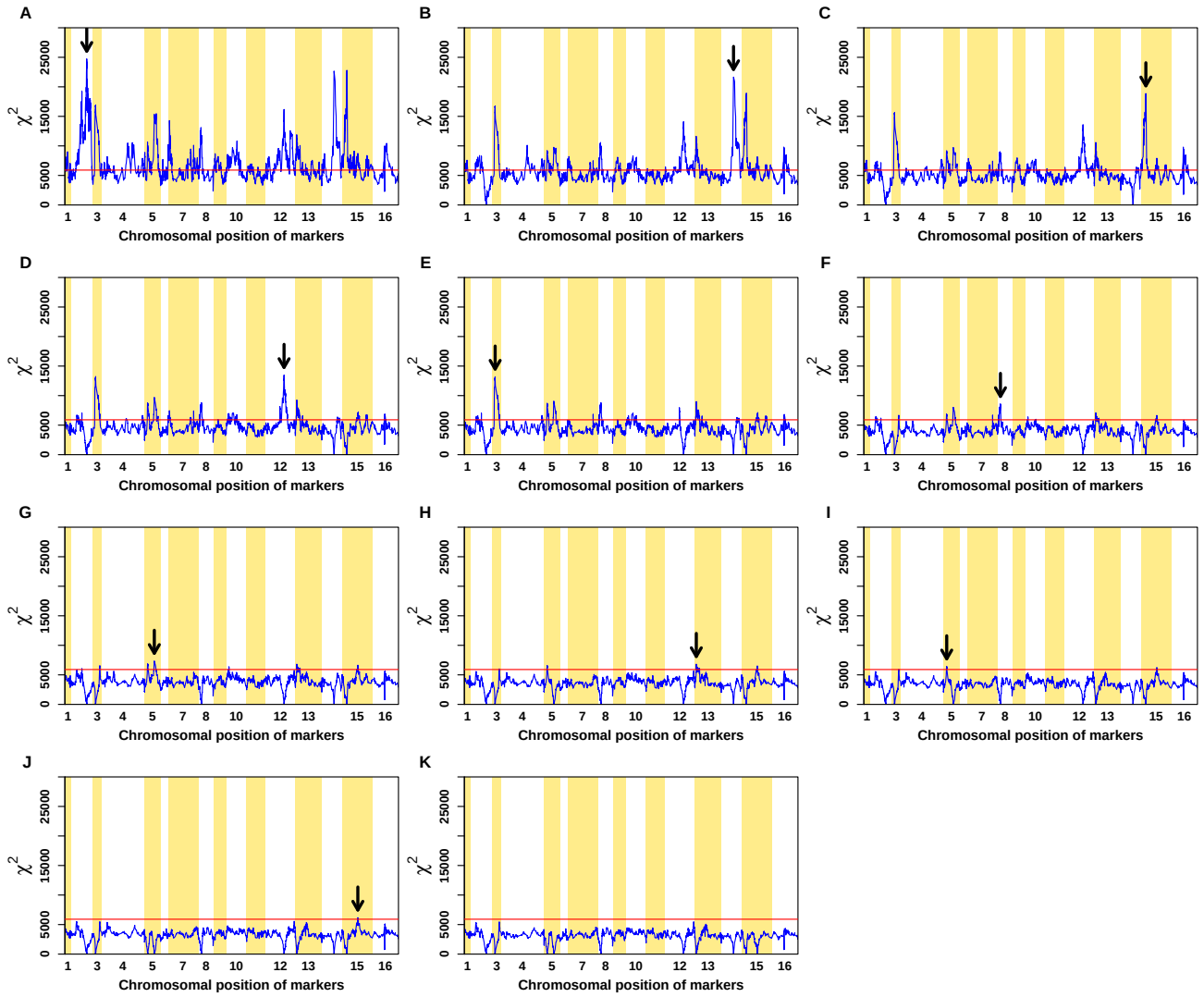

Figure 3.7: Forward Selection of Peaks for Yeast. The red line represent the significant χ^2 threshold Bonferroni corrected at 1%. The black arrows mark the newly selected marker at each round. We Captured total of 10 significant markers.

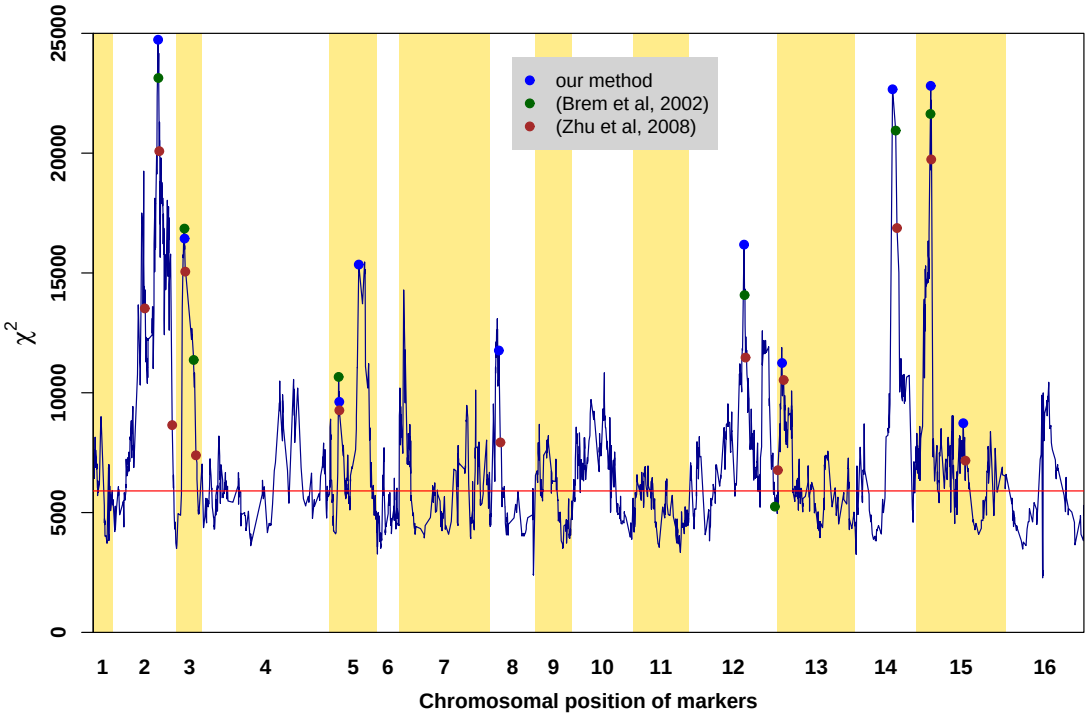

Figure 3.8: eQTL Hotspots Peak Profile for Yeast Comparing Selected Peaks with Two Independent Studies (Brem *et al.*, 2002; Zhu *et al.*, 2008). Our method was able to recover 70% of the peaks reported in these studies. In addition, we were able to discover a new eQTL region on chromosome V. The red horizontal line represents the significant χ^2 threshold for Bonferroni corrected at 1%.

the Gene Ontology (GO) categories.

We analyzed the multiple correlation between the ΔZ of each of the selected markers and a collection of binding preferences for 123 transcription factors from MacIsaac *et al.* (2006). The results are presents in the heatmap of **Figure 3.9**. Each column corresponds to one selected marker. Only those correlations that were significant at a 1% FDR level are colored. For a marker, when the expression of genes enriched for a particular TF's binding site is higher for segregants inheriting the BY allele than RM allele at one of the detected marker, we have positive correlation (indicated in yellow). However, when these genes are relatively more upregulated for segregants inheriting RM allele than BY, we have negative correlation (indicated in cyan). The significant threshold for the p-value is about 2.5×10^{-4} corresponding to |t-value| $\sim$ 3.7. The rows were clustered based on the similarity of the correlation profiles for each TF among the selected markers. To measure for the enrichment for each GO categories, we used non-parametric Mann-Whitney-Wilcoxon test p-values with Bonferroni correction to control for false positives at 1% level.

We will now discuss both analyses results for each detected marker among chromosome I to XVI. The forward selection order of the detected markers is included in **Table 3.1**. We will first focus on the loci on chromosomes III, V, VIII and XII, which have been previously reported by Brem *et al.* (2002); Zhu *et al.* (2008).

For the marker on chromosome III, we did not detect any significant positive correlation of the ΔZ to the TFs specificities. On the other hand, the negatively correlated TFs are more significant including Gcn4p, Leu3p, Uga3p and Bas1p with t-values equal to -10.24, -7.52, -4.71 and -3.90 and -3.49 respectively. Gcn4p, Leu3p and Bas1p are all transcriptional activators of amino acid biosynthesis, with the two later involved in synthesis of specific amino acids, namely leucine, isoleucine, valine and histidine (Arndt *et al.*, 1987; Friden and Schimmel, 1988), whereas Gcn4p is the general transcription activator in amino acid starvation (Hinnebusch and Fink, 1983). Uga3p is active in the presence of GABA and activated genes required for GABA[1]

[1]gamma-Aminobutyric acid, GABA, is a type of amino acid.

utilization as a nitrogen source (Talibi *et al.*, 1995). The top enriched GO category for BY allele is structural constituent of ribosome 5.71 x 10^{-49} and most enriched for RM allele is amino acid biosynthetic process 4.49 x 10^{-30}. The appearance of this locus was expected. This is due to the experimental design of the RM parental strain. The RM strain, used in the genetic crossing, lacks LEU2 and URA3 genes. The transcription of LEU2 is repressed in the presence of leucine and the protein encoded by it acts as catalyser in the leucine amino acid biosynthesis pathway (Kohlhaw, 1988). LEU2 gene is located on chromosome III within the significant region detected by our method. The segregants that inherit this region form the RM parental strain are also auxotrophic for leucine and can not survive media depleted in this amino acids (see **Section 1.5.5** for auxotrophic effect). Even though, the growth medium of the segregants contained leucine and Leu2p was not required, it seems that the lack of LEU2 in the RM strain has caused a significant perturbation in the amino acid metabolism network compared to the BY strain.

For the peak located on chromosome V, only Cha4p and Gcn4p passed the significance threshold with t-value equal to 3.81 and -5.13 respectively. Go analysis ranked the structural constituent of ribosome category (p-value=1.18×10^{-88}) as the most positively correlated and organic acid metabolic process (p-value=5.00×10^{-18}) as the most negatively correlated category. As mentioned in the previous paragraph, the RM strain that was used in the experiment by Brem *et al.* (2002) had been engineered to not possess the URA3 gene. URA3 encodes for an enzyme that catalyzes a step in the uracil biosynthesis pathway, more specifically, it is involved in the metabolism of uridylic acid. Uracil starvation can induce URA3 gene expression about 5 fold (Lacroute, 1968). So the fact that the most negative GO category is the organic acid metabolic (i.e. uridylic acid) supports that this the linkage at this locus is due to RM lacking URA3 gene.

The marker located at the beginning of chromosome VIII, the only TFs that significantly correlated to the ΔZ at this locus were Ste12p (t-value=10.50) and Cha4p (t-value=-6.83). Ste12p is a transcription factor that induces the genes involved in

mating and pseudohyphal/invasive growth pathways (Dolan *et al.*, 1989; Liu *et al.*, 1993). The most positively and negatively enriched GO categories were the regulation of RNA metabolic process and gene expression with p-values equal to 1.86×10^{-13} and 4.72×10^{-78} respectively. Two genes involved in mating response pathway are located near this marker: MAT**a**/MAT$\boldsymbol{\alpha}$ and GPA1. Yeast have two mating types, **a** and $\boldsymbol{\alpha}$. The mating type identity of a cell is determined based on having either MAT**a** or MAT$\boldsymbol{\alpha}$ alleles (Marsh *et al.*, 1991). Two haploid yeast cells mate only if one cell has **a** and the other cell possesses $\boldsymbol{\alpha}$ allele of the MAT gene. So by default the BY and RM parental stains that are involved in a genetic cross have to have a allelic variation at MAT locus. The second gene in this region that is involved in mating response pathway is GPA1. The protein encoded by GPA1 is involved in dampening of the mating-induced signal (Dietzel and Kurjan, 1987). This gene is located closer to the detected marker than the MAT gene. Since this region contain important genes that are involved in mating pathway, it is not surprising that the promoter specificity of Ste12p correlates to this loci.

For the peak located on chromosome XII, we only found one slightly significant positive correlation: Msn4p with t-value of 4.04. Msn4p is a protein activated in the stress conditions such as heat shock, glucose starvation, oxidative shock (Martinez-Pastor *et al.*, 1996). The top positively enriched GO category was the catabolic process (p-value=1.92×10^{-19}). As for the negatively correlated TFs, Hap1p and Cha4p stood at the top of the list with t-values equal to -7.86 and -6.35, respectively. Hap1p is involved in gene expression regulation in response to levels of oxygen and heme in the cell environment (Keng, 1992). The most negatively enriched GO category was the ribosome biogenesis with p-value=6.91×10^{-54}. Gaisne *et al.* (1999) found that the coding region of HAP1 in BY strain carries a Ty1 insertion. This insertion is absent in the RM strain (Brem *et al.*, 2002). The existing mutation within HAP1 gene between BY and RM supports the possibility of HAP1 as the regulator located within the selected region on chromosome XII.

We expected to detect significant linkage at these four loci. The two genes, LEU2

and URA3, were deliberately deleted in the RM parental strain. So these two auxotrophic mutations were generated based on the experimental design and served as positive controls for the linkage (Brem *et al.*, 2002). The linkage to the MAT locus on chromosome VIII is also due to the existing allele variation between the mating BY and RM strains. Finally, the detected linkage on chromosome XII, is caused by the natural mutation in Hap1 gene in the BY strain. These allelic variations served as positive controls for our method and we were able to recover all of them. Now we will discuss the rest of the detected loci.

The peak on chromosome II is the most significant detected eQTL hotspot. Based on the heatmap, the ΔZ's at this marker displays a significant positive correlation to Cha4p (t-value= 11.25) and correlate less significantly to Gln3p (t-value= 4.28). Cha4p is the activator of transcription of CHA1 gene, which encodes a protein that enable yeast cells to survive on L-serine and L-threonine as nitrogen source (Holmberg and Schjerling, 1996). Gln3p is also related to nitrogen utilization regulation (Minehart and Magasanik, 1991). This protein positively regulates the expression of the genes that were suppressed in low level of nitrogen. It seems that the regulatory pathway influenced by this marker has interaction with nitrogen catabolization[2]. On the negative side of the correlation spectrum, we have Adr1p with t-values of -5.73. Adr1p is required for transcription of genes required for ethanol, glycerol and fatty acid utilization (Tachibana *et al.*, 2005). GO analysis for this maker scores ribosome biogenesis with p-value= 1.13×10^{-107} as the most enriched for BY allele category and mitochondrial part with p-value= 2.82×10^{-44} as the most enriched for RM allele.

The peak on chromosome XIII is located about 50k base pairs into the beginning of the chromosome. The TFs that passed the significant correlation threshold with their corresponding t-values are as follows: Zap1p 3.87, Leu3p 4.24, Nrg1p 4.31, Bas1p 4.91 and Hap4p -4.64. Zap1p induces transcription of its target genes in the presence of zinc and represses the transcription of some genes in the low zinc levels (Zhao *et al.*,

[2]The metabolic processes that break down molecules into smaller units and release energy.

1998). Nrg1p is required for glucose repression of some genes (Zhou and Winston, 2001) and also is involved in the repression of FLO11 gene, which is required for invasive and pseudohyphal growth (Kuchin *et al.*, 2002). The top positively enriched GO categories is the amino acid metabolic process (p-value=1.34×10^{-16}). This category is related to the function of Leu3p and Bas1p. Hap4p is the subunit of the complex that activates the transcription of genes involved in cellular respiration (Forsburg and Guarente, 1989). Cellular respiration is a set of catabolic reaction that uses oxygen to convert nutrients into useful energy. The top negative GO category is again related to ribosome: cytosolic ribosome (p-value=3.43×10^{-37}). However the rest of significant negatively enrich GO categories are related to transmembrane transportation. There are two genes located in this region that are involved in transmembrane transport. They are NUP188 and SMA2 with 9 and 2 coding SNPs, respectively. The NUP188 gene encodes a protein that is a subunit of the nuclear pore complex (NPC). Nup188p is involved in the structural organization of NPC and the nuclear envelope permeability. Sma2p is a meiosis-specific prospore membrane protein.

The locus detected on chromosome XIV is the most significant locus after the peak on chromosome II. Correlation to Fkh1p is slightly significant with t-value=3.90. This protein is involved in regulation of the expression of G2/M phase genes. Rpn4p 4.31 and Gcn4p -4.03. RPN4 encodes a protein that induces the degradation of unneeded or damaged proteins. The significant negatively enriched GO category is mitochondrial part (p-value=1.40×10^{-74}), mitochondrial translation (p-value=2.67×10^{-33}), sulfur compound biosynthetic process (p-value=7.07×10^{-7}) and phosphorylation (p-value=5.5×10^{-7}). Positively enriched GO categories are as follows: cytosolic[3] part (p-value=4.37×10^{-27}), actin[4] binding (p-value=5.37×10^{-12}) and nucleosome assembly (p-value=1.12×10^{-6}). This region contains many genes encoding mitochondrial proteins and also several proteins involved in actin organization.

Peak 9 is located on chromosome XV and the genome-wide linkage to this locus was

[3]Cytosol is the intracellular or cytoplasmic fluid.

[4]Actin is a network of proteins that form microfilaments and are responsible for cytokinesis and cell morphogenesis.

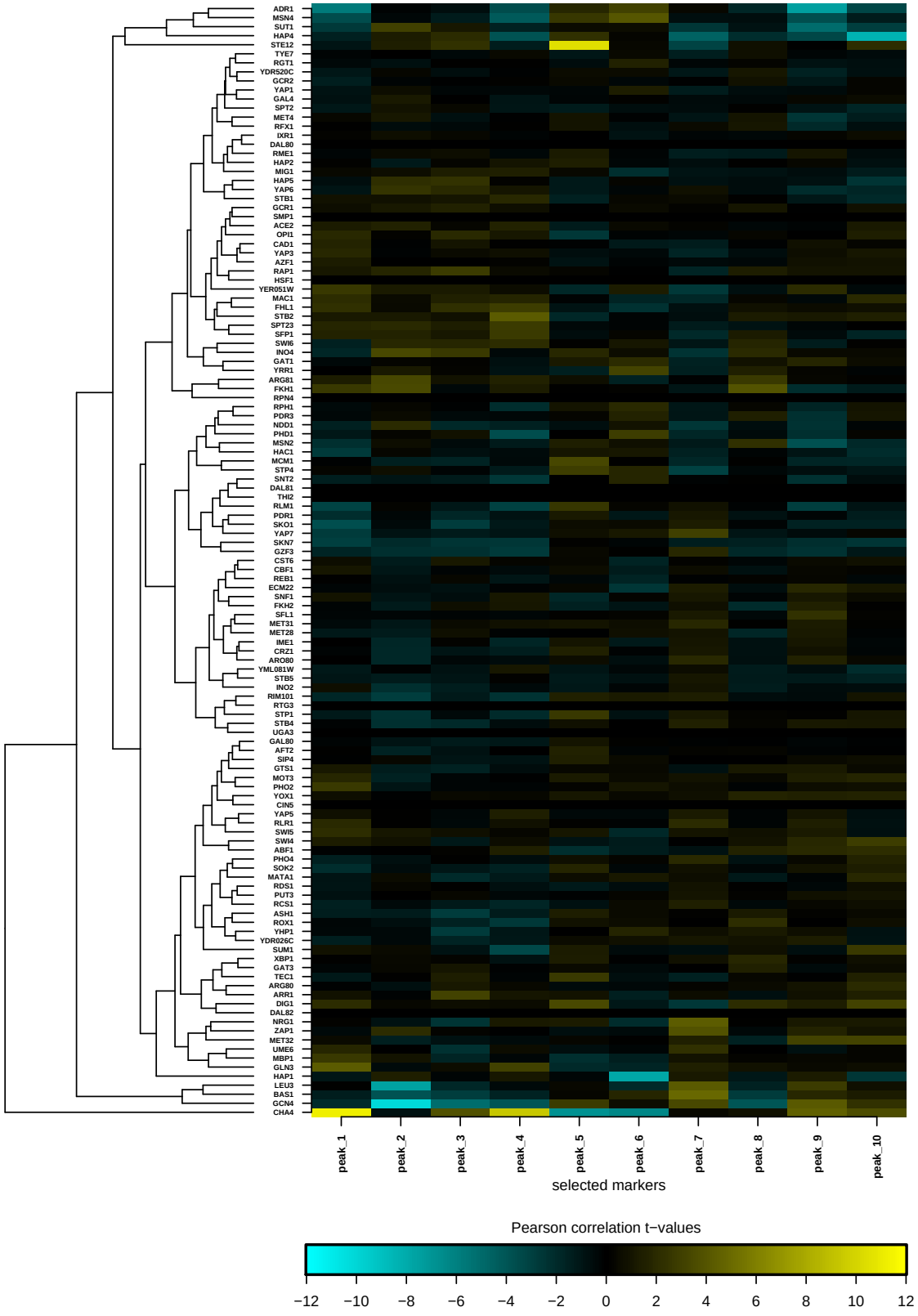

Figure 3.9: Pearson Correlation Heatmap for ΔZ of Selected Markers and the

Figure 3.9: (Cont. Caption) Transcription Factors *in vitro* Binding Specificities. Here ΔZ of each marker was independently regressed by a multiple linear fit against the affinity scores of 123 TF PSAMs from MacIsaac *et al.* (2006). We plotted these correlation t-values using a color scale heatmap clustering the rows. We also calculated correlation coefficient p-values and used them to get the significant threshold at 1% false positive discovery rate. To correct for the multiple testing, we used the approach by Benjamini and Hochberg (1995). We only colored those correlations that were significant at this threshold |t-value|=3.4.

highly significant. Here is the list of the TFs that significantly correlated to this locus: Gcn4p (t-value=4.19), Cha4p (t-value=4.54), Adr1p (t-value=-7.46), Sut1p (t-value=-5.10) and Msn2p (t-value=-3.81). Sut1p is involved in sterol[5] uptake and in the induction of genes required for growth in low oxygen levels (Ness *et al.*, 2001). Msn2p is a transcriptional activator that is active in stress conditions and binds DNA at stress response elements of its target genes. Msn2p and Msn4p are paralog. The significant positively enriched GO categories were: nitrogen compound metabolic process (p-value=1.11×10^{-37}), amino acid biosynthetic process (p-value=4.22×10^{-11}), and cellular localization (p-value=1.28×10^{-10}). Negatively enriched categories are: mitochondrial part (p-value=1.82×10^{-45}), generation of precursor metabolites and energy (p-value=1.26×10^{-12}, and nucleobase and nucleotide metabolic process (p-value=5.79×10^{-9}). The positive GO category are related to the function of Cha4p and Gcn4p.

The 10th peak, detected also on chromosome XV, is the least significant locus among the 10 detected loci. This locus was significantly correlated to Hap4p (t-value=-8.36). The most positively enriched GO category was amino acid metabolic process (p-value=3.38×10^{-17}). Besides the top negatively correlated GO category, which was cytoplasmic part (p-value=5.64×10^{-26}), the rest of significant categories were related to mitochondria and ion transmembrane transportation. Since Hap4p is involved in cellular respiration that takes place in mitochondria, the result from TF correlation and GO enrichment analysis both indicate the linkage between this locus and cellular respiration.

[5]Sterols are essential lipid components of eukaryotes cellular membranes.

The second peak on chromosome V is the novel locus detected by our method. Cha4p and Stb2p are positively correlated to this locus with t-values 9.38 and 4.28 respectively. Stb2p has been shown to interact with the Sin3p-Rpd3p histone deacetylase complex (Kasten and Stillman, 1997). Sin3p is involved in activation and suppression of many cellular processes, including mating-type switch, cell growth, maintaining telomere[6] length and chromatin integrity. The most positively enriched GO category is the ribosome biogenesis category (p-value=1.05×10^{-101}). The significant negative correlations include: Msn4p, Gcn4p and Hap4p with t-values equal to -4.25, -3.98, -3.89 respectively. All these 3 TFs are found to coordinate the transitions between phases of the metabolic cycle of yeast (Rao and Pellegrini, 2011). Interestingly, we found the most negatively correlated GO category is catabolic process (p-value=6.31×10^{-15}) and energy derivation by oxidation of organic compounds (p-value=1.19×10^{-8}). The genes with roles related to these TFs and GO categories are as follows: SWI4, a transcriptional activator that regulates transcription of cyclins and genes required for DNA synthesis and repair, also interacts with Sin3p; DOT6, involved in rRNA and ribosome biogenesis and subunit of the histone deacetylase complex, involved in telomeric gene silencing and filamentation; ALD5, mitochondrial component, involved in regulation or biosynthesis of electron transport chain components and acetate formation; RGI1, involved in energy metabolism under respiratory conditions, induced upon intracellular iron depletion; CEM1, possible role in fatty acid synthesis and is required for mitochondrial respiration.

[6]Telomeres are repetitive nucleotide sequences located at the end of each chromosome and protect them from degradation.

Table 3.1: Yeast eQTL Hotspots Results, obtained by forward selection with significant threshold with Bonferroni correction at 1%.

No.	**Selection Order**	**Max χ^2**	**Position of Selected Marker**	**No. of Genes in Selected Region**	**Putative Regulator**
1	1	24.7×10^3	Chr2: 548,401	327	?
2	5	16.4×10^3	Chr3: 91,977	122	LEU2
3	9	9.6×10^3	Chr5: 117,048	65	URA3
4	7	15.3×10^3	Chr5: 350,744	149	SWI4, DOT6, ALD5, RGI1, CEM1
5	6	11.8×10^3	Chr8: 111,690	73	GPA1, MAT
6	4	16.2×10^3	Chr12: 662,627	161	HAP1
7	8	11.2×10^3	Chr13: 49,903	83	NUP188, SMA2
8	2	22.7×10^3	Chr14: 449,639	186	Mitochondial & actin genes
9	3	22.8×10^3	Chr15: 174,364	151	GPD2, SKM1, DDR2
10	10	8.7×10^3	Chr15: 563,943	103	AZF1, IDH2

3.3.3 Recovered and Novel eQTL Hotspots for *C.elegans*

Figure 3.10 presents the χ^2-statistic profile for the 1455 markers along the chromosomes of worm for the real expression data (blue) and null data (red). The null data was generated by shuffling the expression levels of each gene among the 214 RIALs. We observe that in the case of the randomized data, the χ^2 values oscillates around the expected value equal to the degrees of freedom (df) of the χ^2 distribution. In our case df is the number of genes equal to 13,500. However, when the actual expression data was used, we observe a distinct profile compared to the null data. The peaks represent putative eQTL hotspots.

A region on chromosome I has a χ^2-statistic significantly below the expected value both for the expression data and the null data. By looking at the allelic composition of markers within this region, we realized that the samples' allele is almost entirely inherited from the Bristol parental strain. We are not sure whether it is a experimental artifact or there is a biological reason for it. It could be that the worm, which acquired Hawaii parental allele at these marker, did not survive the experimental medium in combination with the allelic composition of other markers. As we mentioned earlier, we had normalized the expression data for each gene among the 214 samples. This means that at markers with mostly one type of the genotype, the ΔZ's are almost close to zero and resulting in a small χ^2-statistic.

Like we did for yeast analysis, we applied a forward selection procedure to detect the influential loci. **Figure 3.11** displays the iterations of our marker selection procedure. The black arrows in the figure mark the location of newly selected significant marker between each iteration. We detected a total of 6 markers as significant with Bonferroni correction at 1% corresponding to χ^2-threshold $\sim$ 14,213. The detailed information for the selected peaks are summarized in **Table 3.2**.

We compared our results to the results from Rockman *et al.* (2010), where the mRNA expression levels of each gene was treated as quantitative traits and a nonparametric interval mapping approach and clustering were used to detect the significant eQTLs

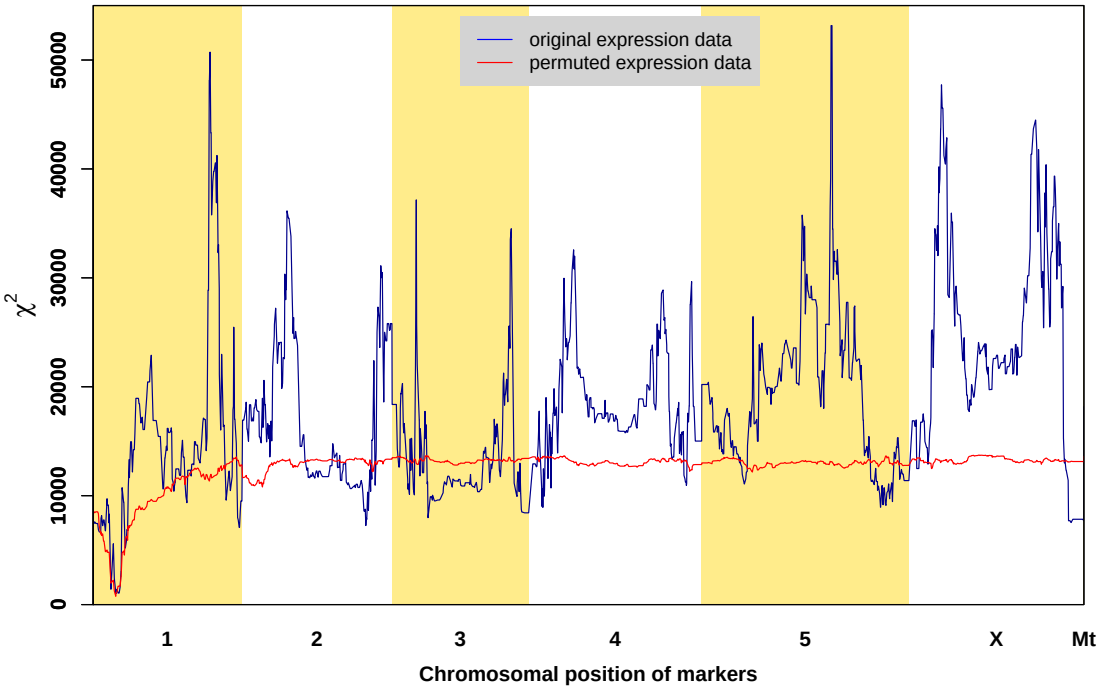

Figure 3.10: eQTL Hotspots Peak Profile for Worm. We calculated the allele specific difference of the sums of expression data z-scores at each chromosomal marker (ΔZ) and then calculated χ^2 for all the 1455 markers. The blue plot resulted from using expression data and the red was for the case where we permuted the expression vector of each gene randomly among the segregants while keeping the genotype map intact. Then we calculated ΔZ's using the z-scores of the permuted expression data. It serves as the null distribution for the χ^2 values. It oscillates around the expected value $\sim$ 13,500 (i.e. the number of genes).

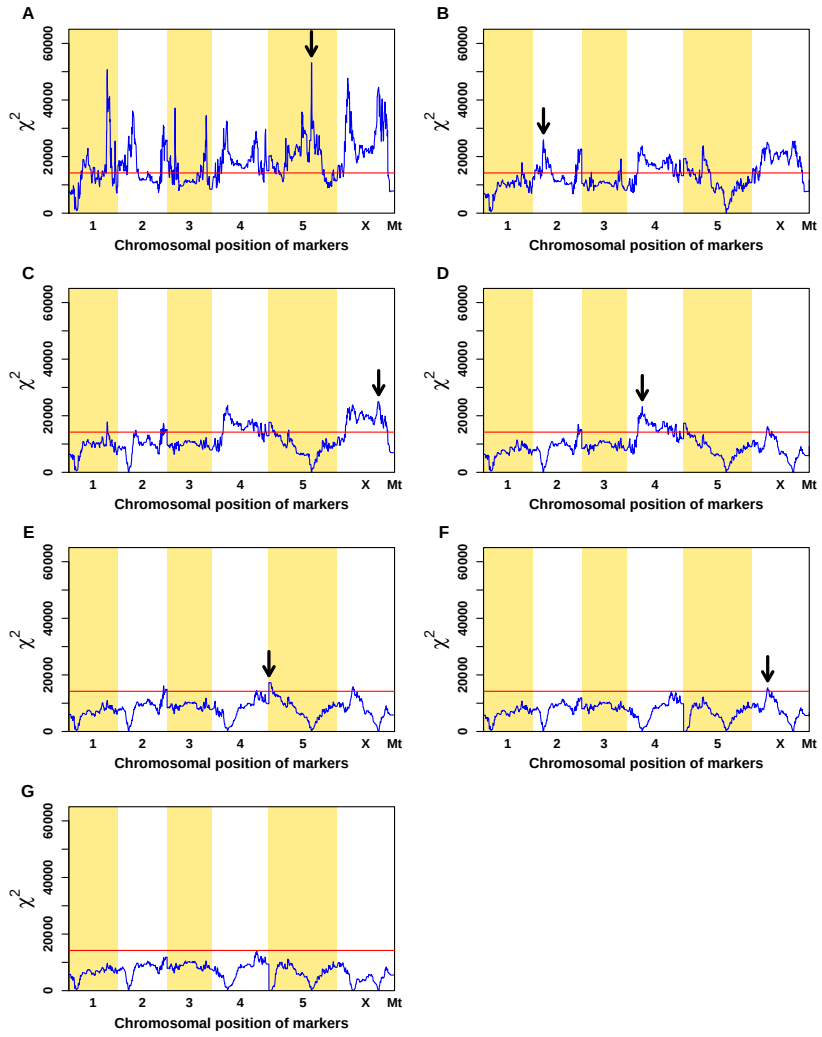

Figure 3.11: Forward Selection of Peaks for Worm. The red line represent the significant χ^2 threshold Bonferroni corrected at 1%. We captured a total of 6 peaks.

hotspots. The interval mapping method uses the likelihood analysis to tests the linkage at each marker with considering its two neighbouring markers in the linkage region to estimate the QTL location more precisely (Lander and Botstein, 1989). The nonparametric approach uses the generalized Wilcoxon rank-sum statistics and interval mapping developed by (Kruglyak and Lander, 1995). **Figure. 3.12A** displays the results of our χ^2-statistic analysis, and **Figure. 3.12B** results from Rockman *et al.*. Our method was able to capture 2 of the reported markers on chromosomes IV and X.

3.3.4 Assessment of Possible Regulatory Roles for the Detected Worm eQTL Hotspots

We used GO enrichment analysis to test whether the detected eQTL hotspot are biased toward a specific biological function.

The first peak is located on chromosome II. The GO enrichment analysis at this locus showed relative enrichment for RIALS that inherited the Bristol parental allele for these categories: cell cycle (p-value=5.86×10^{-24}), protein tyrosine phosphatase activity (p-value=2.45×10^{-18}), DNA metabolic process (p-value=3.94×10^{-11}), reproductive cellular process (p-value=2.27×10^{-9}), proteasome complex (p-value=2.13×10^{-8}), structural constituent of cuticle (p-value=6.07×10^{-8}), biopolymer modification (p-value=2.74×10^{-7}) and embryonic development ending in birth or egg hatching (p-value=5.56^{-10}). These categories are involved in a variety of processes and not related to a specific function. However the categories with relative enrichment for RIALs inheriting Hawaii allele are less dispersed: regulation of cellular process (p-value=1.85×10^{-46}), synapse (p-value=2.09×10^{-16}), signal transducer activity (p-value= 1.66×10^{-9}), growth (p-value=3.79×10^{-7}), tetrapyrrole binding (p-value=4.19×10^{-6}) and behavior (p-value=5.73×10^{-6}). Three categories are related to the nervous system of the nematode.

The second peak is located on chromosome IV and it is one of the peaks that was

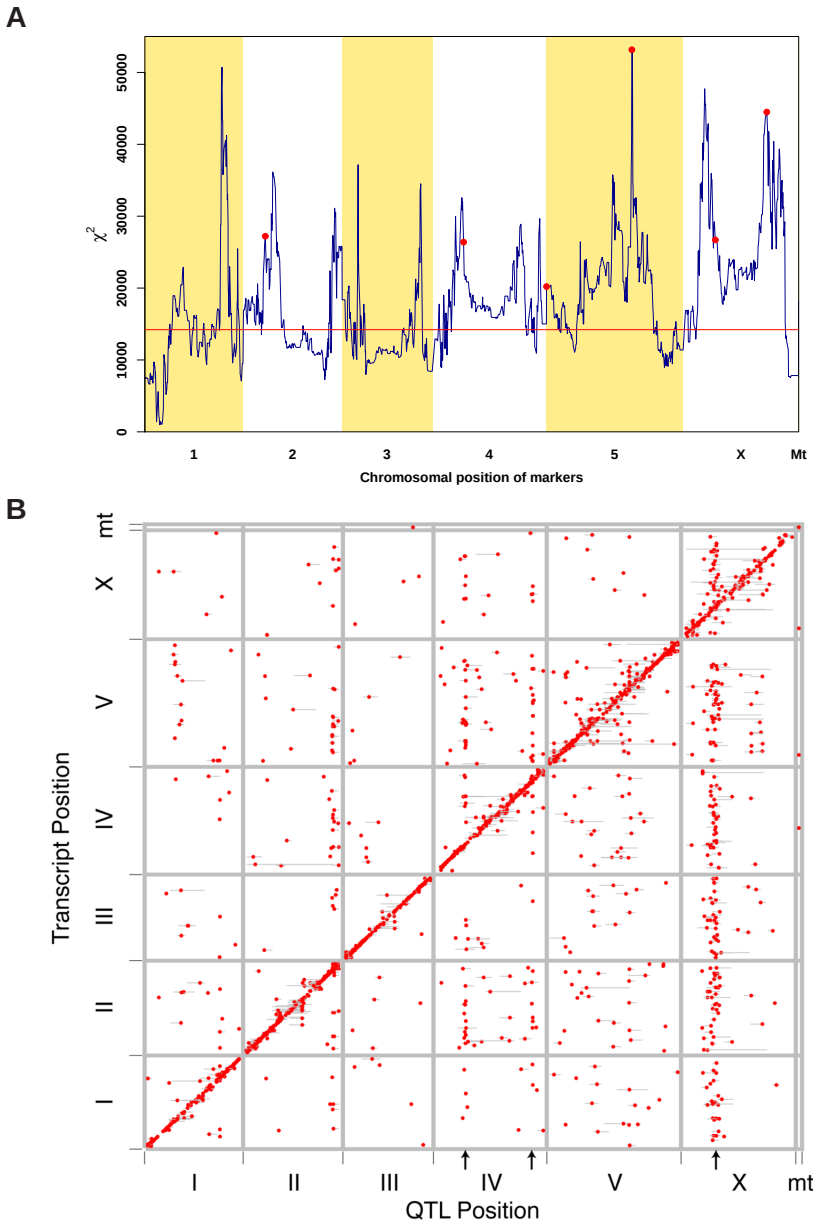

Figure 3.12: eQTL Hotspots Peak Profile for Worm Comparing Selected Peaks with an Independent Study (Rockman *et al.*, 2010). Our method was able to recover two of the peaks reported in that study (A). The red horizontal line represents the significant χ^2 threshold for Bonferroni corrected at 1% level. eQTL hotspots from Rockman *et al.* (2010) detected at 5% FDR rate (B). Three top significant loci are shown with the arrows. Figure (B) from Rockman *et al.* (2010).

detected by Rockman *et al.* (2010). These categories that are enriched for Bristol allele are: embryonic development ending in birth or egg hatching (p-value=9.23×10^{-106}) and DNA metabolic process (p-value=6.76×10^{-13}). The categories enriched for Hawaii allele are mostly related to signal transduction through the cellular membrane: integral to membrane (p-value=7.16×10^{-57}), transmembrane transporter activity (p-value=7.28×10^{-20}), G-protein coupled receptor protein signaling pathway (p-value=4.63×10^{-20}), lipid glycosylation (p-value=7.47×10^{-11}). So it is possible that a SNP within this region regulates the function of the cellular membrane. Other GO categories are: iron ion binding (p-value=7.25×10^{-9}) and peptidase activity, acting on L-amino acid peptides (p-value=2.84×10^{-7}).

The third peak is located several dozen kb into chromosome V. Many enriched GO categories for Bristol allele are related to three processes: metabolic, oxidation-reduction and transportstion. Here is the list of significant categories: oxidoreductase activity (p-value=2.82×10^{-17}), transporter activity (p-value=3.72×10^{-13}), metabolic process (p-value=5.28×10^{-12}), cellular metabolic process (p-value=2.16×10^{-21}), biosynthetic process (p-value=1.40×10^{-16}), primary metabolic process (p-value=2.90×10^{-20}), catalytic activity (p-value=1.55×10^{-19}), ion transport (p-value=8.12×10^{-14}), cellular macromolecule metabolic process (p-value=2.96×10^{-10}), nucleosome assembly (p-value=2.46×10^{-9}), signal transducer activity (p-value=2.22×10^{-9}), oxidation reduction (p-value=1.36×10^{-9}), phosphorylation (p-value=1.44×10^{-7}) and transmembrane transport (p-value=1.96×10^{-7}). There were only two significant GO categories enriched for Hawaii allele: structural constituent of cuticle (p-value=1.85×10^{-21}) and phosphorus metabolic process (p-value=1.38×10^{-12}).

Peak 4 is located on chromosome V. This locus has the largest χ^2 value. The positively enriched significant categories were mostly related to membrane: membrane part (p-value=4.40×10^{-84}), membrane (p-value=9.53×10^{-17}), signal transducer activity (p-value=2.16×10^{-16}) integral to membrane (p-value=6.74×10^{-17}). The most negatively enriched categories was the embryonic development ending in birth or egg hatching category (p-value=1.07^{-144}).

Table 3.2: Worm eQTL Hotspots Results, obtained by forward selection with significant threshold with Bonferroni correction at 1%.

No.	**Selection Order**	**Max χ^2**	**Position of Selected Marker**
1	2	27.2×10^3	Chr2: 3,397,001
2	4	26.4×10^3	Chr4: 4,779,016
3	5	20.2×10^3	Chr5: 16,796
4	1	53.2×10^3	Chr5: 13,031,355
5	6	26.7×10^3	ChrX: 4,900,879
6	3	44.5×10^3	ChrX: 12,750,713

The fifth peak is located on the sex chromosome X and it was also reported by Rockman *et al.* (2010). This locus was positively enriched for the embryonic development ending in birth or egg hatching (p-value=5.73×10^{-116}), nuclear RNA (ncRNA) metabolic process (p-value=3.23×10^{-11}) and cellular catabolic process (p-value=5.97×10^{-9}). The negatively correlated GO categories were highly enriched for categories related to membrane: membrane part (p-value=5.07×10^{-38}), integral to membrane (p-value=4.03×10^{-18}), ion channel activity (p-value=1.08×10^{-13}), G-protein coupled receptor protein signaling pathway (p-value=2.04×10^{-13}), structural molecule activity (p-value=5.98×10^{-13}).

Finally, 6th peak is also located on the sex chromosome X. This locus is also positively correlated to the embryonic development ending in birth or egg hatching (p-value=3.52×10^{-117}). Other positively enriched GO category for this locus were the cellular response to stress category (p-value=1.27×10^{-8}), glycerolipid metabolic process (p-value=7.73×10^{-8}) and ncRNA metabolic process (p-value=8.98×10^{-8}).

3.4 Conclusion

In this chapter, we have presented the method developed for the detection of *trans*-acting genetic loci that regulate the expression of large number of genes. It uses

mRNA expression levels and genotype data to identify these loci. At each marker, we first calculate the difference of the sums of the expression level z-scores between the samples inheriting different genotypes. Then, we calculate a χ^2-statistic for every chromosomal marker position using the difference of the sums of all genes. The traditional approach for identifying the expression quantitative trait loci or "eQTL" hotspots considers the expression levels of each gene as a quantitative trait and uses ANOVA and clustering to find these hotspots. In contrast, our approach uses all of the genes and does not require enforcing any significant threshold on the expression levels.

We applied our method to *S. cerevisiae* and *C. elegans*. We used the expression data and genotype data of 108 yeast segregants generated from a genetic cross between the BY and RM strains (Brem *et al.*, 2002). We identified 10 loci for yeast, of which 9 were reported previously (Brem *et al.*, 2002; Zhu *et al.*, 2008). Four loci appeared because of the lack of LEU2 and URA3 genes in the RM strain due to the experimental design, the mating locus allelic variation between the parental strain, and a natural mutation existing in the HAP1 gene of the BY strain. These four loci served as positive controls for our method. Therefore, the detection of them increased our confidence in the validity of our method. We found a novel locus on chromosome V with a possible role in regulation of mitochondrial respiration and ribosome biogenesis. The putative regulator at this locus and the rest of the detected loci can be validated by allele replacement experiments (see **Section 1.5.5**).

For worm, we used the expression data and genotype data of 214 RIALs developed from the N2 and Bristol strains (Rockman *et al.*, 2010). Our analysis detected 6 loci, with 2 having been discovered previously by (Rockman *et al.*, 2010).

Our method can be applied to a population of segregants/offsprings of any organism whose genome-wide expression data and genotype data are available. In a way, we consider the genome of an organism as a system that has intrinsic resonance modes. Hence, the *trans*-acting loci resemble these modes to which the regulation of a large number of pathways is linked. By using the natural sequence variations within a

related population, we can "excite" these resonance modes. So, our method can be thought of as a way of seeking these "trans-acting resonances" or "nuclear genetic resonances".

Chapter 4

Harnessing Natural Sequence Variation to Dissect Post-Transcriptional Networks in Yeast

This chapter is adopted from a manuscript co-authored by Mina Fazlollahi, Eunjee Lee, Xiang-Jun Lu, Pilar Gomez-Alcala and Harmen J. Bussemaker.

4.1 Introduction

With the advancement of high-throughput sequencing technologies, linkage studies have become a significant tool for deciphering genetic regulatory networks. Linkage studies seek to identify the genetic loci that correspond to an observed phenotype among genetically related individuals. A common approach is to treat the mRNA expression levels as heritable traits and use them to identify expression quantitative trait loci (eQTL) hotspots that regulate the expression of a larger number of genes.

Our approach uses differential mRNA expression data and genotype data of a segregating population together with sequence specificities for RNA-binding proteins (RBPs) as a priori information. The goal is to identify chromosomal loci (i.e. genes) that modulate the activity levels (i.e. traits) of the RBPs under study. We call these loci activity quantitative trait loci or "aQTLs" (Lee and Bussemaker, 2010). Our method consists of two steps. We first use the sequence specificities of RBPs and segregant-specific mRNA expression data to infer the activity levels of RBPs for each segregant. These inferred activity levels are treated as quantitative traits used in the second step. The second step aims to detect the aQTLs based on the significance of the difference between the distributions of the activities when splitting the segregants based on the inherited parental alleles at each chromosomal marker. RNA binding proteins perform post-transcriptional processing of the RNA transcript. Many RBPs are also involved in the stability regulation of the mRNAs by binding to stability associated motifs mostly located within the untranslated regions (UTRs) of the mRNA. Therefore, the significant aQTL that we detect could be involved in the post-transcriptional stability regulation of the mRNAs.

We used the affinity scores of the 25 RBP-region combinations, which we obtained with our motif discovery approach presented in Chapter 2, to infer the activity levels. We used the expression and genotype data of 108 segregants generated from a genetic cross between the BY and RM strains in yeast (Smith and Kruglyak, 2008). Using the estimated affinity scores of the 25 factor combinations together with the mRNA differential expression levels for every segregant in the population, we inferred the activity levels of the RBPs for every individual. We treated these activity levels as quantitative trait to discover loci modulating them.

For the aQTL analysis we recovered a known locus that contains the MKT1 gene, which has been shown to modulate the activity levels of Puf3p (Lee *et al.*, 2009). Interestingly, we found that depending on the interaction of Puf3p with the $5'$ or $3'$ UTRs of its target mRNAs, there are different loci regulating this protein's activity. We also found a locus on chromosome XV containing the IRA2 gene as a putative

activity modulator for Puf4p. We further tested it by using parental IRA2 allele replacement data obtained by Smith and Kruglyak (2008). The difference in the expression between the two mutant strains with reciprocal IRA2 alleles significantly correlated with Puf4p sequence specificities. This indicates the possible role for IRA2 as modulator of Puf4p activity level.

4.2 Methods

4.2.1 Experimental Data Used

For the aQTL analysis, we used genome-wide mRNA expression data for 108 haploid segregants from a genetic cross between two parental strains (BY and RM) (Smith and Kruglyak, 2008). As differential expression values, we used $\log_2$-ratios between segregants and a reference consisting of a mixture of the BY and RM strains. Genotype data for the same segregants at 2956 markers was obtained from the authors of (Brem *et al.*, 2002).

4.2.2 Inferring Segregant-Specific of RNA-Binding Protein Activities

From our RNA-binding proteins (RBPs) motif discovery analysis we obtained 25 independent RBP-region combinations (see Chapter 2). As in the work by (Lee and Bussemaker, 2010), we used the affinity scores of the obtained position specific affinity matrices (PSAMs) as a predictor for mRNA differential expression levels in the low protein concentration region established by (Bussemaker *et al.*, 2007; Foat *et al.*, 2006). We considered the occupancy of a given mRNA region by a particular RBP to be proportional to the total affinity of a desired PSAM for a sliding window along the whole mRNA, $5'$ or $3'$ untranslated region (UTR) or open reading frame (ORF) sequences (see **Equation 2.7**). We performed a genome-wide multiple regression on

the 25 RBP-region combinations presented in Chapter 2 of every segregant mRNA expression $\log_2$-ratios to infer segregant-specific activity levels of the RBPs (see also **Figure 4.2**).

$$y_{gs} = \beta_{0s} + \sum_{\phi} \beta_{\phi s} K_{\phi g} \tag{4.1}$$

where y_{gs} represents the differential mRNA levels of gene g for segregant s relative to the reference. The affinity score the mRNA sequence of gene g is denoted as $K_{\phi g}$ corresponding to a RNA-binding protein ϕ. Here, the regression coefficient $\beta_{\phi s}$ represents the activity level of RBP ϕ for segregant s.

4.2.3 aQTL Mapping

Significant aQTL region were discovered by splitting the multiple regression coefficient between BY and RM at every marker and testing for the significance of the difference between the distributions of the two groups of coefficients using composite interval mapping (CIM) method for maximum resolution (Zeng, 1994). CIM uses multiple regression on multiple markers to obtain a more precise mapping of the QTL. We used CIM implementation in R/qtl package by Broman *et al.* (2003). LOD score, an acronym for 'logarithm of the odds ratio' was calculated to check for the linkage effect. The odds ratio is the probability of observing a specific genotype in the population given linkage at a particular recombination fraction (θ) versus the same probability computed conditional on independent genotype assortment ($\theta = 0.5$) (Chotai, 1984).

$$z(\theta) = \log_{10}[p(r;\theta)/p(r;0.5)] \tag{4.2}$$

For example, a LOD score of 3 means that the probability of observing the linkage considering a random recombination has odds of ~ 0.001. Thus, high values of LOD score favor the linkage hypothesis.

We calculated LOD score to test the linkage of the RBPs inferred activities to each

locus. We performed 200 independent random permutations on the expression data for each gene among the segregants (preserving the genotype data) to get LOD score threshold at 1% FDR level. We obtained this threshold for each factor separately.

To ensure that the detected aQTL regions for the RBPs are modulated by trans-acting factors and also not dominated by a single gene eQTL, once we obtained the significant regions we re-did the analysis after eliminating 3 groups of genes: gene that encode the RBPs, genes fully or partly located within about 10 kb up- and down-stream of obtained aQTL regions and genes with significant eQTL peak located 20 kbp window around detected aQTL marker and have affinity higher than 50% of max affinity score for the RBP under study. To find the last group, we did our QTL analysis using the expression of each gene as a trait and calculated LOD score for every marker using CIM method. We combined these 3 groups of genes and eliminated them for each RBP separately, thus not affecting the activity calculation of each factor by eliminating unrelated genes for it. Same procedure was performed for calculating aQTL profile using the protein levels.

4.2.4 Protein-Protein Interaction Data

We downloaded the latest version (April 2012) of protein-protein interaction from the Biogrid website (`http://thebiogrid.org`) for yeast as of April 2012. We used it to detect any known genetic or physical interaction with the genes located in our detected aQTL regions. The physical interaction refers to the case where the two proteins directly or through a cofactor bind together in order to initiate or inhibit a process. Two common approaches to test the physical interaction between two proteins are: two-hybrid assay and affinity-capture (see **Section 1.5.4**). Whereas, a genetic interaction is inferred when the two proteins do not directly interact but rather are involved in connected pathway or process. For example, one approach to test genetic interaction is when the deletion or mutation of a gene rescues the growth defect of a yeast strain containing a mutation or deletion of another gene. This assay

is known as synthetic rescue. Other types of approaches such as positive (or negative) genetic interaction can also be employed to identify genetic interactions. The positive genetic interaction is based on the observation that mutation or deletion of two genes separately result in a more severe phenotypic defect than expected when compared to the combined mutations or deletions in the same cell. Conversely, negative genetic interaction is detected when the combined effect of mutations or deletions of two separate genes is more severe than expected.

4.2.5 Validation of Predicted Locus-RBP Associations

We used gene expression profiles for two mutant strains growing in glucose medium collected by Smith and Kruglyak (2008) where IRA2 alleles were swapped between the BY and RM strains. We label the strain carrying the RM allele of IRA2 in the BY background as (RM@IRA2) and the strain carrying the BY allele of IRA2 in the RM background as (BY@IRA2). The reference sample used for the gene expression measurements was pooled parental mRNA (BY and RM). To obtain the net effect of the IRA2 allele replacement on the genome-wide mRNA levels, we subtracted the mean log-ratio of the related background of each mutated strain (shown RM@IRA2 strain below).

$$y_g^{\mathrm{BY}\rightarrow\mathrm{RM@IRA2}} = \log_2\left(\frac{[\mathrm{mRN}A_g](\mathrm{RM@IRA2},\mathrm{glucose})}{[\mathrm{mRNA}_g](\mathrm{pool})}\right) - \log_2\left(\frac{[\mathrm{mRN}A_g](\mathrm{BY},\mathrm{glucose})}{[\mathrm{mRNA}_g](\mathrm{pool})}\right) \tag{4.3}$$

We performed multiple regression between the above data vector and the affinity scores of 25 RBP-region combinations. Similarly, we calculated the relative mRNA expression for the RM strain when IRA2 was swapped with the BY allele and the RM background and applied multiple regression analysis. To capture the effect of the IRA2 allele swap between the two backgrounds, we subtracted the regression coefficients between the two cases for all the 25 combinations. We then permuted the two y vectors for all genes 1000 independently to calculate the statistical significance

threshold at 1% FDR level ($|y| > 2.7$).

4.3 Results

4.3.1 Genetic Linkage Analysis

Boorsma *et al.* (2008) experimentally validated that predicting the modulation of a transcription factor (TF) activity at post-transcriptional level is possible by scoring mRNA differential expression levels of putative targets of the TF. As for the activity levels, it has been shown that they vary among members of a population of an organism and can be treated as a quantitative trait for genetic linkage analysis to capture polymorphisms that modulate the activity of the transcription factors (Lee and Bussemaker, 2010). We have applied a similar approach to detect the post-transcriptional and translational network and identify transacting-loci that control the activity of RBPs in yeast both at mRNA expression levels and protein levels respectively.

The segregants generated from the genetic cross between the BY and RM strains each have a distinct allelic combination of the parental genotypes. Therefore, the gene expression levels of every segregant is uniquely perturbed relative to the reference, compared to the rest of the segregants. We used the segregant-specific genome-wide mRNA expression levels to predict RNA-binding protein (RBP) activity levels for each of the 108 segregants (Smith and Kruglyak, 2008). **Figure 4.1** displays the steps involved in the analysis. To calculate the RBP affinity scores, we used the 25 RBP-region combinations that we obtained by our motif discovery approach (see **Figure 2.7**). We first scored all genes by calculating the affinity of the PSAM using the chosen sequence (whole mRNA, UTRs or ORFs) and **Equation 2.7**. Then we performed multiple regression on all 25 factor affinities of each segregant mRNA expression data. We considered the coefficients of the regression as the quantitative representation of the RBP activities. We used composite interval mapping (CIM) (Zeng, 1994) to map aQTLs for each RBP-region combination. To correct for multiple test-

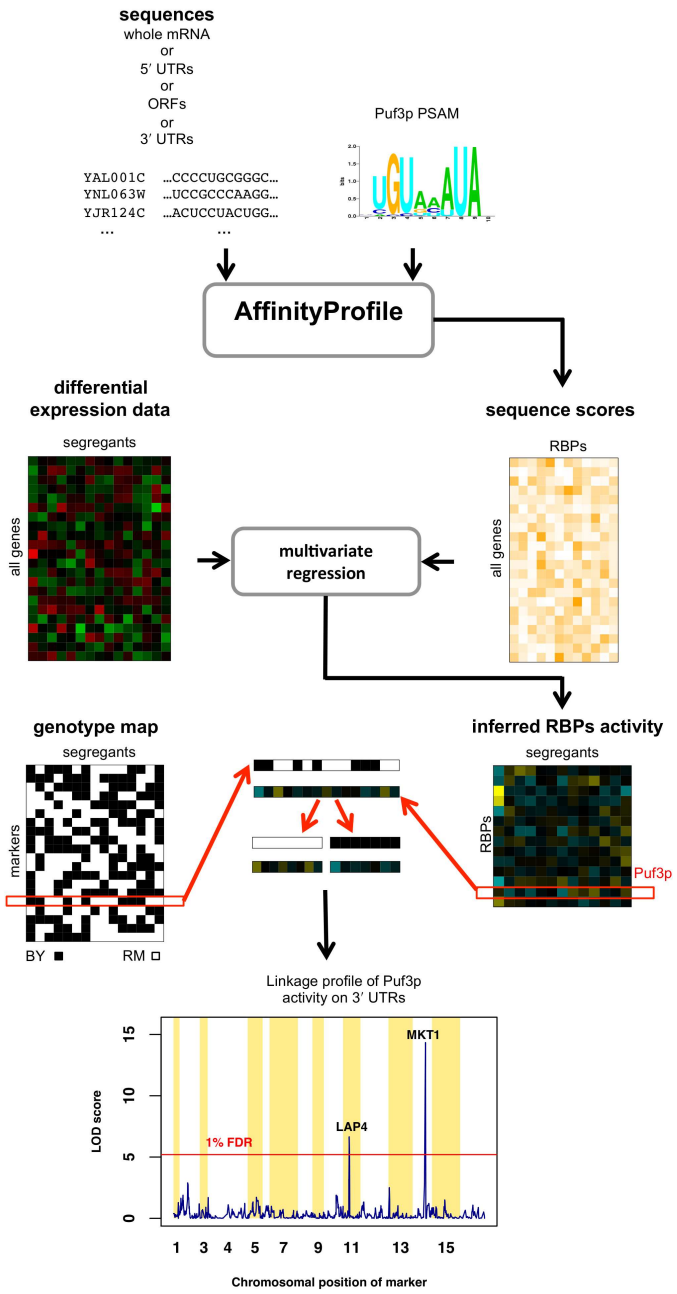

Figure 4.1: Overview of the aQTL Approach. Genome-wide affinity scores were calculated using position-specific affinity matrix (PSAM) and sequences. The affinity scores were used to infer segregant-specific RBP activities. The activities were obtained by multiple linear regression on differential mRNA expression levels to the affinity scores. The regression coefficients represent the RBP activity levels for each segregant. For linkage analysis, the activities were treated as quantitative traits. Whenever the distribution of the inferred activity levels of a RBP depends on the genotype variation of a specific chromosomal marker, we would obtain a high LOD score at that marker and it indicates the presence of an aQTL (at 1% FDR level).

ing, we calculated the LOD score thresholds corresponding to a 1% false discovery rate (FDR) by performing 200 permutations (see Methods).

Figure 4.2 illustrates the genome-wide linear regression of mRNA differential expression of a particular segregant on the 3′ UTRs affinity scores for Puf3p (A) and Puf4p (B). In each case, the slope represents the activity level of the RBP under study. In practice, we used multiple linear regression of differential mRNA levels of a segregant on all 25 RBP-region combinations mentioned above (see **Equation 4.1**). The inferred activities were treated as a quantitative phenotype for our aQTL analysis.

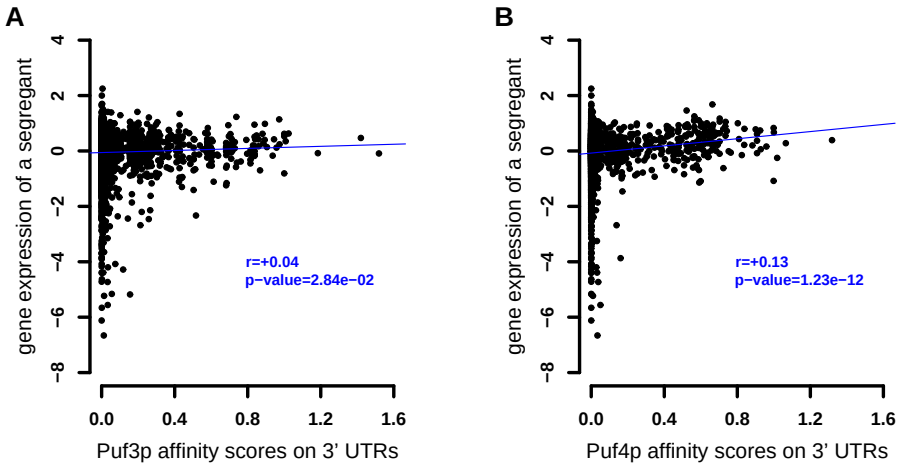

Figure 4.2: Scatter Plots for Activity Calculation for a Particular Segregant. Activity level is the slope of the linear fit on the gene expression levels to the 3′ UTRs affinity scores for Puf3p (A) and Puf4p (B). For this particular segregant, Puf4p is significantly active whereas Puf3p is slightly active.

Figure 4.3A shows the clustered heatmap for the RBP activities Pearson correlation and **Figure 4.3B** shows the affinity scores Pearson correlation. We observed that even though the affinities of most of the factors are uncorrelated, most RBP activity levels are negatively or positively correlated. This could reflect that at the mRNA expression levels, some of the RBPs activities are modulated by the same upstream mechanism even though the RBPs have different target sets as shown schematically in **Figure 4.3C**. An example for supporting this idea are Puf4 (3′ UTR) and Sik1p (ORF). The affinity correlation between these two factors is 0.04 (p-value = 0.01),

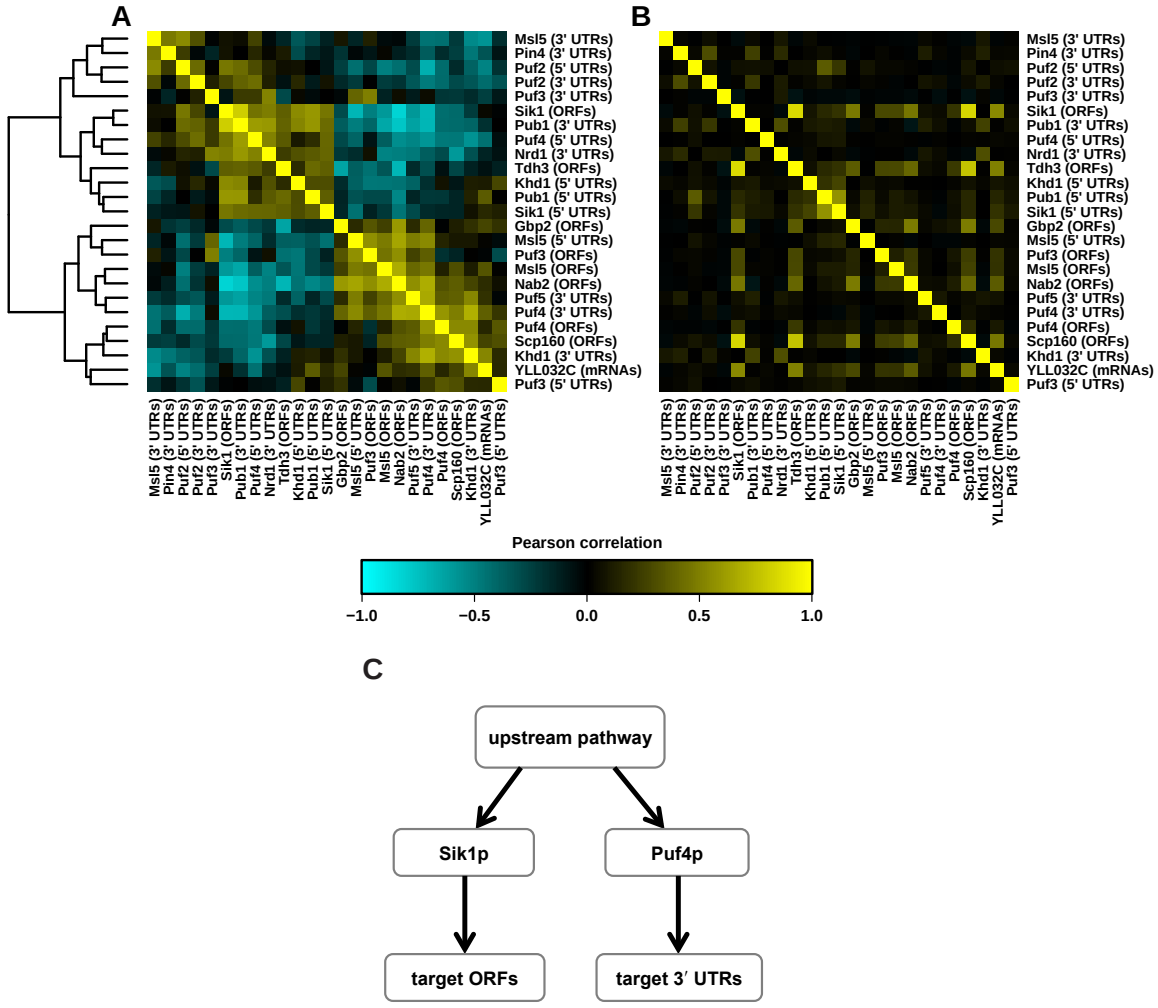

Figure 4.3: (A) Clustered heatmap of Peasrson correlations calculated for the inferred activity levels of 25 factors. The activities were calculated using multiple regression on the genome-wide expression levels to the RBP affinities on selected mRNA regions. (B) Affinity Pearson correlations. For the comparison convenience, we have kept the same order for the affinity correlation heatmap as in the clustered heatmap of panel (A). (C) Possible scenario for observing more significant correlations between inferred activity levels of the RBPs. Panel (A) Shows that there is a significant negative correlation between Puf4p activity on 3′ UTRs and Sik1p activity on ORFs (Wilmes *et al.*, 2008); However, Puf4p and Sik1p have distinct set of target genes (weak correlation between their affinities as observed in panel (B)).

which indicate they little target set overlap; However, their activities are highly correlated with correlation of -0.72 (p-value $< 1 \times 10^{-16}$). In Wilmes *et al.* (2008) by high-throughput measurements, they found that Sik1p and Puf4p have negative ge-

netic interaction (see **Section 4.2.4**), which confirms our negative activity correlation between the two proteins. Both proteins are involved in ribosome biogenesis.

4.3.2 Decoupling of Activities of Two PUF Protein Family: Puf3p and Puf4p

We also focused on Puf3p ad Puf4p activity to check their correlation in more detail **Figure 4.4**. The biding site for these two PUF family members differ in the length of the gap between UGUA and AUA and they have distinct target sets (as shown in the affinity correlation heatmap of **Figure 4.3B**. Foat *et al.* (2005) studied the activity levels of several proteins including Puf3p and Puf4p using expression data for many different stress conditions. Their finding indicated that Puf3p and Puf4p, activity level is highly anti correlated when cells were exposed to different sugar sources. Using expression data by (Gasch *et al.*, 2000), we calculated the activity levels between each stress condition and the affinity of 25 factors. **Figure 4.4A** shows the activity levels for Puf3p (3′ UTR) and Puf4p (3′ UTRs). There is a significant negative correlation (r = -0.67, p-value $< 1 \times 10^{-16}$). Our results confirm that their activity levels among the segragants are not correlated (r = +0.047, p-value= 0.63) shown in **Figure 4.4B**. This suggests that their activity modulation is linked to different genetic loci and different pathways are responsible for activating or suppressing these two proteins. Thus we can decouple the aQTL for these two proteins using genetic differences naturally occurring in the segregants, which is not observed simply by exposing the cells to stress conditions.

Figure 4.5 displays the aQTL profile for Puf3p whose activity were inferred based on affinity score of 3′ UTRs. The significant LOD score threshold at 1% FDR is indicated with the red horizontal line. Two loci are marked with "*". The LOD score of the locus on chromosome IV is insignificant. The strip chart corresponding to the split of the inferred activity levels of Puf3p at this marker is shown above it with a black arrow. There is no significant difference in the distribution of the activities of

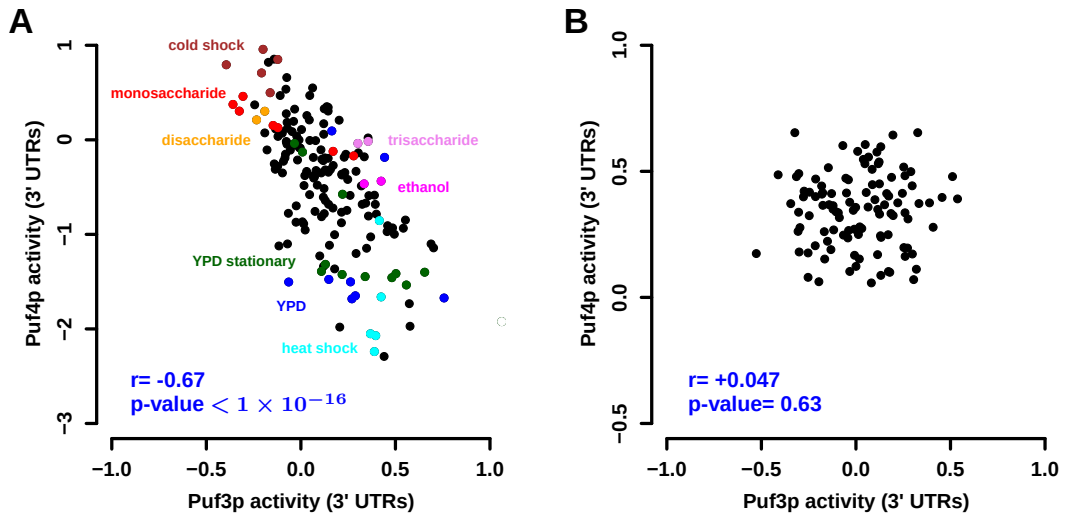

Figure 4.4: Decoupling of Puf3p and Puf4p activity in Segregants in Comparison to Stress Conditions. (A) Scatter plot of inferred activity levels for (Gasch *et al.*, 2000) stress conditions shown for Puf3p (3′ UTRs) and Puf4 (3′ UTRs). To infer the activities, we performed a multiple regression on the mRNA levels of each experimental condition to all 25 factor affinity scores. (B) Scatter plot of inferred activity among 108 segregants for the same two factors.

segregants inheriting the BY allele and segregants inheriting the RM allele at this marker. On the other hand, the split of the activities based on the genotype at the marker on chromosome XIV are significantly different and results in a large LOD score. This indicates the presence of an aQTL.

To make sure that the detected aQTLs are *trans*-acting, we eliminated the genes that encode the RBPs from the expression data and also the group of genes that are located within about 10 kb upstream and downstream of the detected aQTL. Furthermore, to eliminate any effect caused by eQTL of only a few genes at the detected aQTL, we obtained the set of genes with significant eQTL LOD score within a region of about 20 kb around the aQTL locus and among them we eliminated the genes with affinity higher than 50% of the maximum affinity score for the RBP under study. **Figure 4.6** presents the motivation for eliminating these three set of genes. For example, the peak on chromosome XII for Puf3p (5′ UTRs) becomes insignificant when we eliminate the PUF3 gene located within this locus (**Figure 4.6A**). Also, the

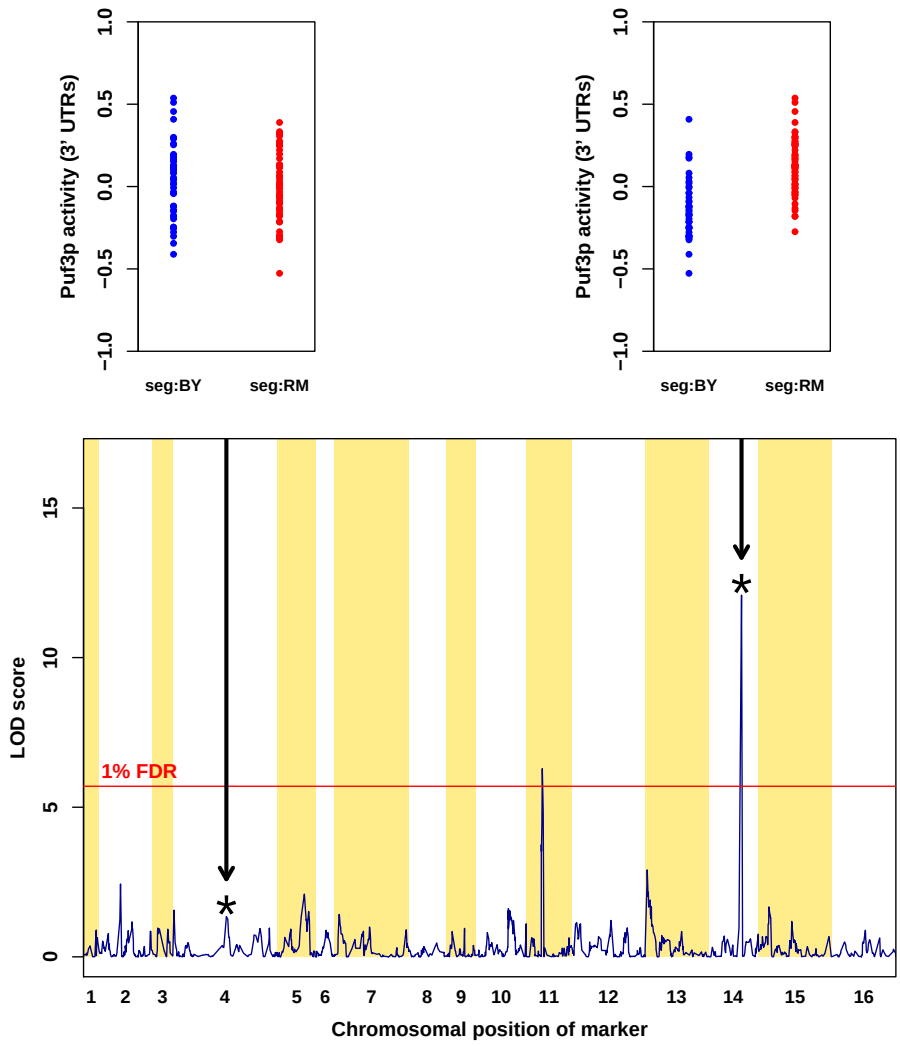

Figure 4.5: LOD Score Profile for Puf3p Using 3′ UTR Affinity Scores. The red line indicates the significant threshold at 1% FDR. The LOD score for the marker on chromosome IV is insignificant because the activity level distribution of the two subset of segregants based on the genotype at this marker are not significantly different. Whereas the two distributions are significantly different for the activity split based on the genotype of the marker on chromosome XIV.

elimination of the YER124C gene, located on chromosome V, causes the significant peak on chromosome II to disappear for Msl5p (3′ UTRs). The expression of this gene is linked to this locus on chromosome II and its 3′ UTR scores more than 50% of the maximum affinity score for Msl5p (**Figure 4.6**B). Therefore, we eliminated the outliers by using such strategy.

By eliminating these three groups of genes, we are confident that the observed varia-

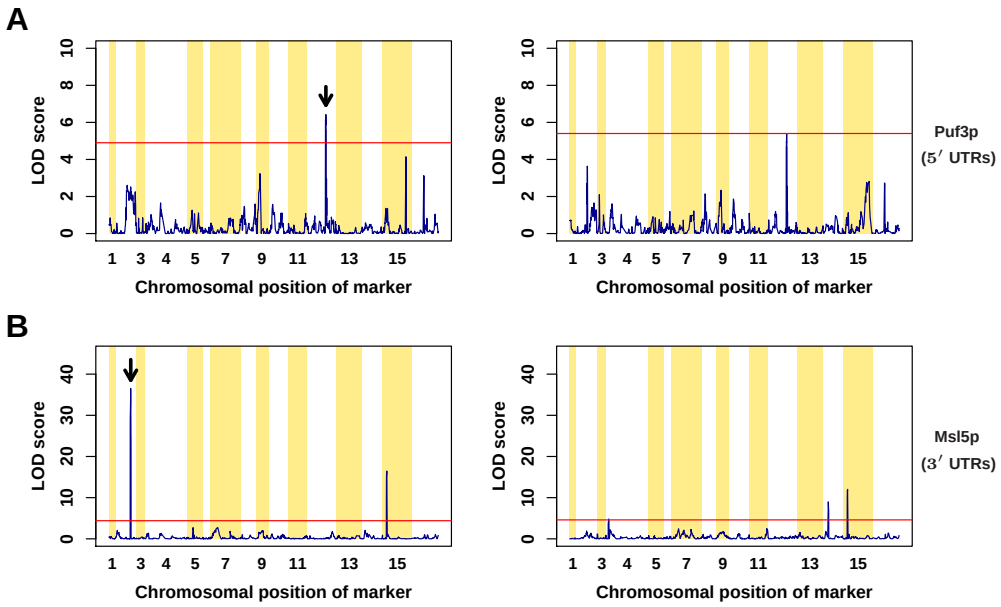

Figure 4.6: Elimination of Outliers for aQTL. For each of the 25 combinations, we eliminated the gene encoding the RBP, genes located within the 20 kb around the detected aQTL, and genes with significant eQTL peak within the 20 kb region of aQTL and estimated affinity score more than 50% of the max score for that RBP. These three groups of outlier genes are eliminated from the expression data and affinity scores for each RBP.(A) The peak on chromosome XII becomes insignificant after elimination of the PUF3 gene form the expression data and affinity scores. (B) The removal of YER124C, located on chromosome V, causes the peak on chromosome II to vanish for Msl5p.

tion in the activity levels of RBPs are not due to a very few local or distal polymorphisms that effect the expression of few genes. We eliminated these three groups of genes from the expression data and affinity scores for each RBP separately. If a new peak emerged after this elimination step, we again checked and eliminated the genes in the vicinity of that new peak to make sure our analysis is self-consistent. **Figure 4.7** displays the summary of significant markers, indicated by blue color at the marker coordinates, for the first round of the analysis where all genes were included and the last round where the three groups of genes were eliminated consistently.

Detailed information for the identified aQTLs is summarized in **Table 4.1**. It lists any known physical or genetic interactions of the RBPs and other genes or gene products (also see **Figure 4.8** and **Figure 4.10**).

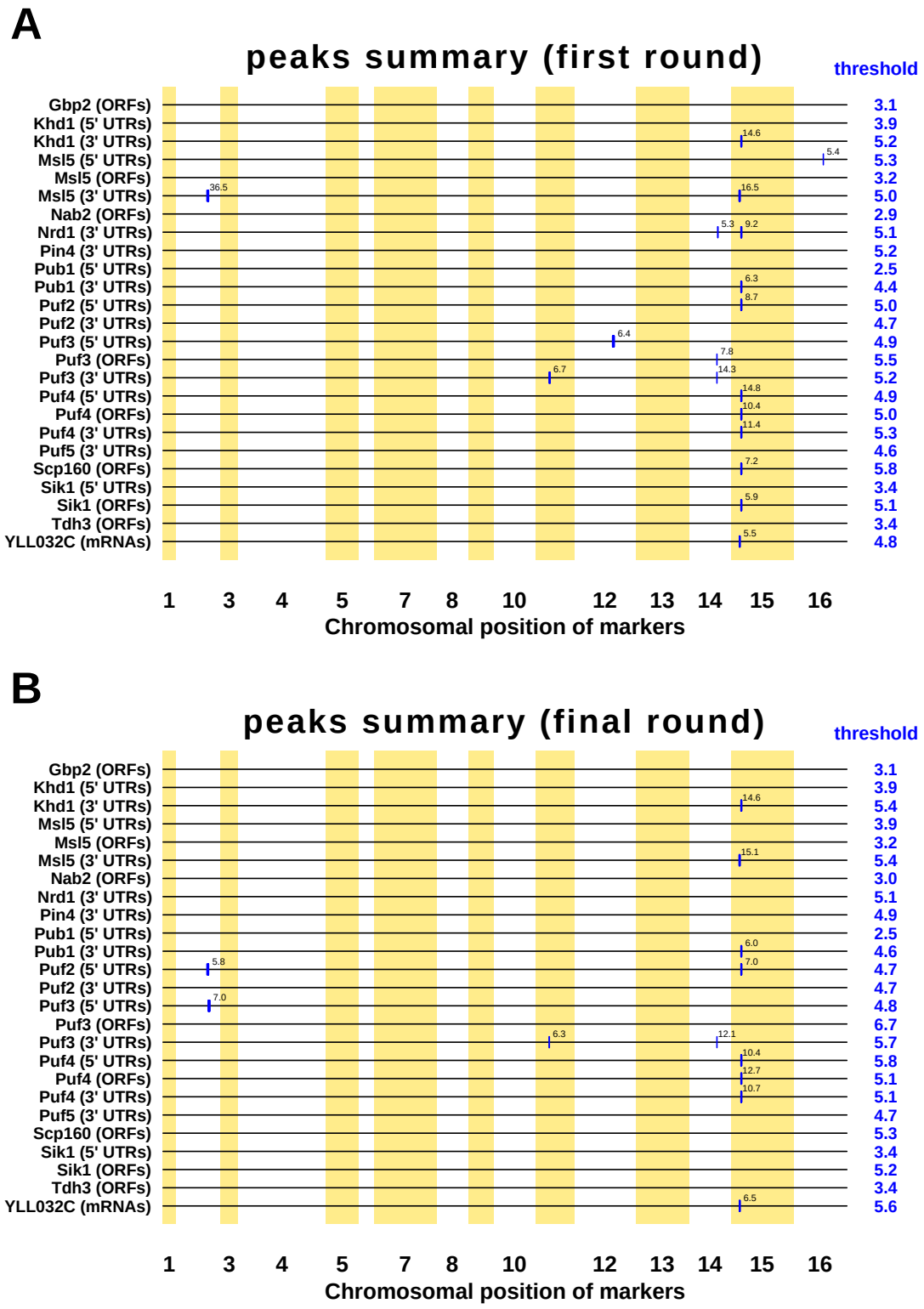

Figure 4.7: aQTL results for all of the 25 accepted RBP-mRNA-region combinations. (A) shows the significant peaks obtained by composite interval mapping (CIM) method when all genes were included (first round), and (B) shows the results after eliminating neighboring genes of the peaks, genes with significant eQTL at these peaks and genes encoding the RBPs (last round).

Table 4.1: Yeast RNA Binding Proteins aQTL Results, Based on 200 Permutations on mRNA Expression Data at 1% FDR Level

RBP	**mRNA Region**[a]	**aQTL Region**	**Max LOD-Score**	**Direct Interaction**	**Interaction Type**	**Reference**
Khd1	3′ UTRs	Chr15 154,310-193,910	14.6			
Msl5	3′ UTRs	Chr15 136,328-170,944	15.1			
Pub1	3′ UTRs	Chr15 154,310-193,910	6.0			
Puf2	5′ UTRs	Chr2 533,269-567,220	5.8			
Puf2	5′ UTRs	Chr15 154,310-193,910	7.0			
Puf3	5′ UTRs	Chr2 555,788-592,862	7.0	POP7	GI[b]	(Wilmes *et al.*, 2008)
Puf3	3′ UTRs	Chr11 229,053-247,943	6.3	LAP4	PI[c]	(Breitkreutz *et al.*, 2010)
Puf3	3′ UTRs	Chr14 449,640-502,315	12.1	MKT1	GI	(Lee *et al.*, 2009)
Puf4	5′ UTRs	Chr15 154,310-193,910	10.4			
Puf4	ORFs	Chr15 154,310-193,910	12.7			
Puf4	3′ UTRs	Chr15 154,310-193,910	10.7			
YLL32C	mRNAs	Chr15 141,634-170,944	6.5			

[a] The affinity scores calculated for the indicated mRNA region, which is used for inferring the activity levels.
[b] Genetic Interaction
[c] Physical Interaction

4.3.3 Recovered aQTL for Puf3p

Figure 4.8 presents the aQTL results for Puf3p and the corresponding line plots for the inferred activity distribution of the segregants based on their inherited allele at the significant aQTL markers (blue and red dots). Our method recovered a locus on chromosome XIV for Puf3p when we calculated its activity using 3′ UTRs sequences **Figure 4.8C,F**. This locus was previously discovered computationally and experimentally by (Lee *et al.*, 2009). They have suggested that MKT1 regulates p-body abundance, which consequently regulates Puf3p target abundance. Lee *et al.* (2009) tested the effect of the MKT1 deletion on Puf3p target mRNAs in the RM strain. The genome-wide mRNA expression profile of the MKT1Δ strain demonstrated that Puf3p targets are significantly down-regulated.

We used BLAST (Altschul *et al.*, 1997) to align the protein sequences of MKT1 for RM strain and S288c[1] strain, an isogenic strain to BY strain. We identified two amino acid mutations between the two strains at position 30, the glycine amino acid (G) in RM is switched to aspartic acid (D) in S288c, and at position 453, the arginine in RM (R) is replaced with a lysine (K) in S288c.

Besides MKT1, there are 29 other genes located in this region of which one has a role related to Puf3p. TOM7 is a Component of the translocase of outer membrane (TOM) complex responsible for recognition and initial import steps for all mitochondrially directed proteins and it promotes assembly and stability of the TOM complex (Meisinger *et al.*, 1999) and (Model *et al.*, 2001). However, we did not identify any coding or non-coding polymorphisms for TOM7 mRNA including the UTR regions between S288c and RM. This makes MKT1 a more plausible candidate as the modulator of Puf3p activity when bound to 3′ UTRs of its targets. See **Figure 4.8F** for the distribution of the inferred segregant-specific activity of Puf3p for a split based on the inherited alleles at the MKT1 locus.

[1]S288c is a laboratory yeast strain that is isogenic to the BY strain with only about 39 SNPs occuring between the their genomes.

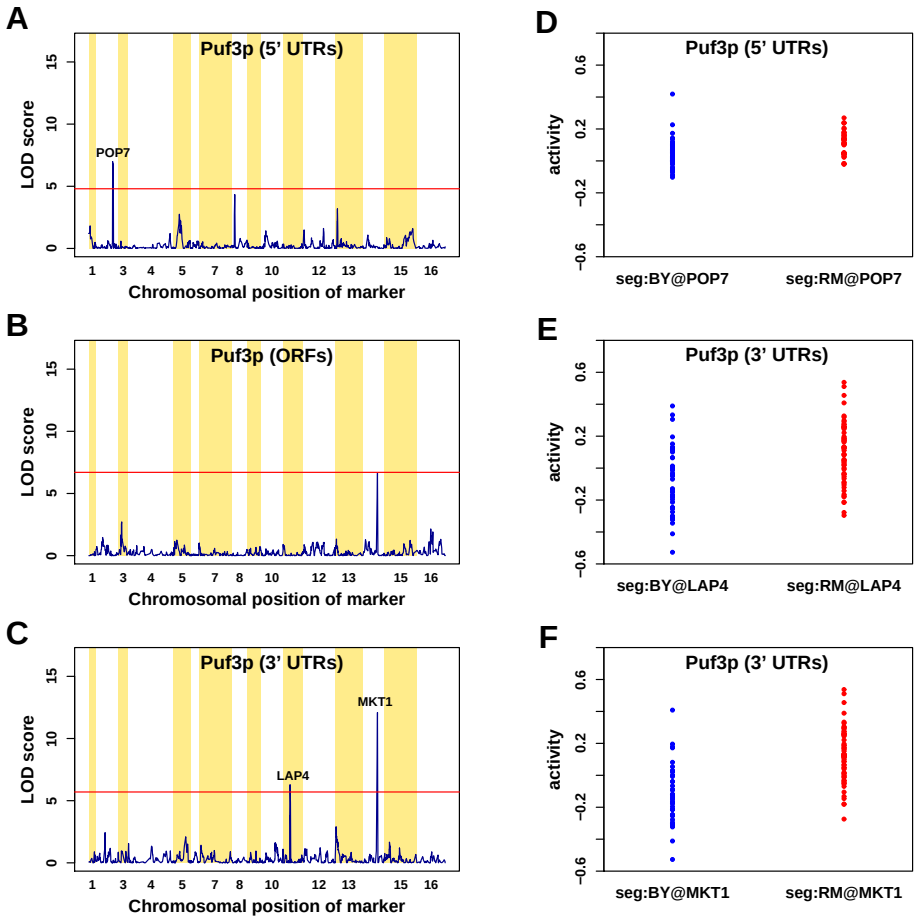

Figure 4.8: Puf3p aQTL Profile. Results for the *trans*-acting genetic modulators of Puf3p activity, mapped using our aQTL method. The significant threshold at 1% FDR level are calculated using 200 independent permutations of the expression data (red horizontal lines). We obtained a separate aQTL profile for Puf3p when using affinity scores on the 5′ UTRs (A), ORFs (B), and 3′ UTRs (C). The significant aQTL peaks survived after filtering out for the three groups of genes mentioned earlier. We identified POP7 as a putative modulator of Puf3p activity levels when inferred from the 5′ UTRs for the locus on chromosome II (A). The corresponding split of the activity levels at this marker is shown in panel (D). We detected two possible modulators, LAP4 on chromosome XI and MKT1 on chromosome XIV, for Puf3p activity levels when inferred from the 3′ UTRs (C). (E) and (F) present the activity level splits at these two loci.

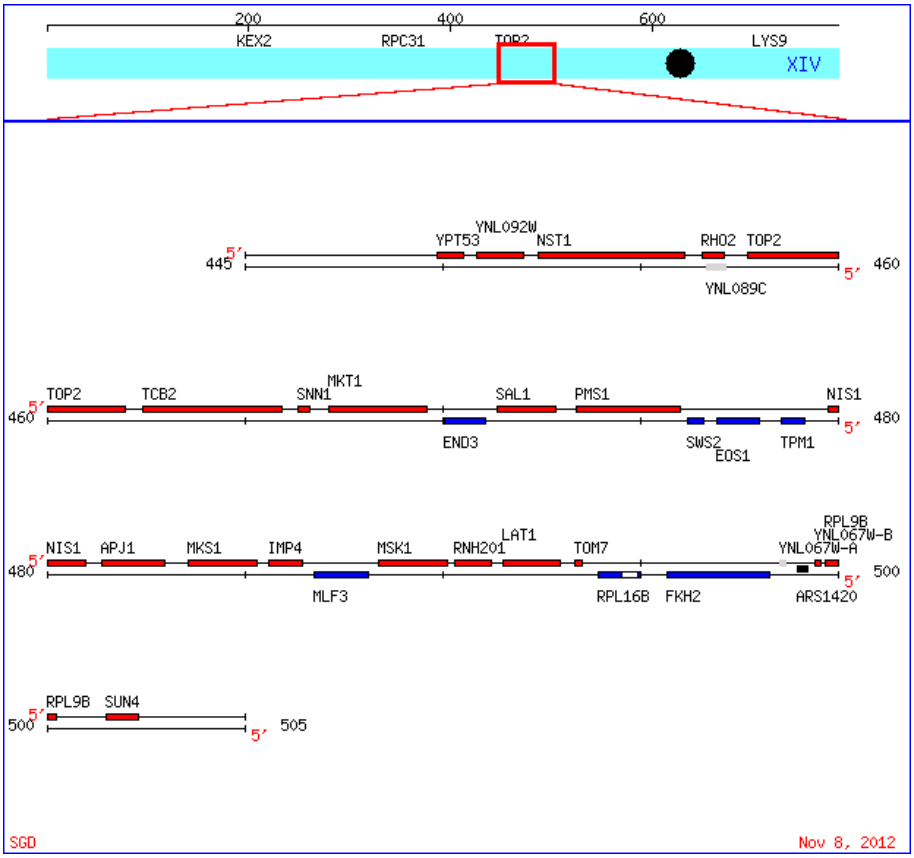

Figure 4.9: Region Around the MKT1 Gene on Chromosome XIV. We found that at least one SNP in a gene in this region might be significantly effecting the activity levels of Puf3p on (3′ UTRs) between the BY and RM strains. Two related genes are located in this region: MKT1 and TOM7. The MKT1 gene has two coding SNPs between the RM and S288c strains. Figure generated by SGD website (`http://www.yeastgenome.org`).

4.3.4 Puf3p Activity Modulated Through Different Loci Depending on Binding to 5′ UTRs or 3′ UTRs

Previously it was mostly known that Puf3p interacts with the 3′ UTRs of its targets and no evidence of functional interaction with the 5′ UTR has been reported. As mentioned above, Puf3p activity is modulated through a locus on chromosome XIV when considering its binding to 3′ UTRs of its targets. However considering Puf3p binding to 5′ UTRs, we were able to link its activity level variation to a locus on Chromosome II (see **Figure 4.8A,D**). This region contains POP7, which is reported to have positive genetic interaction with Puf3p (Wilmes *et al.*, 2008). Alignment of POP7 gene between RM and S288c strains revealed a coding polymorphism at amino acid position 58 on the Pop7p sequence. The histidine (H) in RM strain is replaced by a glutamine (Q) in S288c. Pop7p is the subunit of both RNase MRP and nuclear RNAse P; RNase MRP cleaves pre-rRNA, while nuclear RNase P cleaves tRNA precursors to generate mature 5′ ends and facilitates turnover of nuclear RNAs (Chamberlain *et al.*, 1998) and (Houser-Scott *et al.*, 2002). This region also contains CDC28 and also several mitochondrial related genes such as ETHA1 and FZO1. ETHA1 is involved in ethyl ester biosynthesis and is localized to the mitochondrial outer membrane. Perhaps FZO1 has a function closer to Puf3p. Fzo1p is an integral membrane protein involved in mitochondrial outer membrane tethering and fusion and it has a role in mitochondrial genome maintenance. These findings indicate that the activity modulation of Puf3p is linked to different genomic locations and networks depending on where it binds to the mRNA.

Besides MKT1 locus, we identified a second locus as a putative modulator of Puf3p activity levels when inferred from 3′ UTRs affinity scores. This locus, which is marginally significant, contains the LAP4 gene on chromosome XI (**Figure 4.8C,E**). Lap4p contains 4 coding polymorphism between RM and S288c.

4.3.5 Independence of Puf4p Activity Modulation to the Motif Location on its Target mRNAs

Puf4p aQTL profile is shown in **Figure 4.10**. Whether Puf4p binds to 5′ UTRs, ORFs or 3′ UTRs of its targets, its activity regulation is controlled by a locus on chromosome XV. This locus contains REX4 and BRX1. Both of them are involved in pre-rRNA possessing and ribosome assembly. The coding region of the REX4 gene contains three coding SNPs between the RM and S288c strains as follows: the asparagine (N) at position 34, phenylalanine (F) at position 155, lysine (K) at position 248 in RM are mutated to lysine (K), leucine (L), and arginine (R) in S288c, respectively. There is a single non-coding polymorphism at position 243 within the coding region, the thymine in RM is mutated to cytosine in S288c.

It is known that Puf4p interacts with mRNAs encoding nucleolar rRNA-processing factors. The inferred activity distribution of the segregants based on their inherited parental allele at the locus on chromosome XV is shown in **Figure 4.10D-F**. It is interesting to note that the sign of Puf4p activity levels in segregants inheriting BY and RM alleles at this locus switches between 5′ UTRs and 3′ UTRs.

4.3.6 Validation of Detected Loci with IRA2 Allele Swap Data

To pinpoint more precisely the putative modulators located within the detected loci responsible for the variation in activity levels of the RBPs, further investigation by narrowing down the significant aQTL regions and then using allele swap data for the genes located in the detected regions is required. Using the segregant expression data from (Smith and Kruglyak, 2008), we found that the activity variation for 6 of the 25 RBP-region combinations are linked to the same region on chromosome XV. One of the genes located in this region is IRA2, which encodes a GTPase-activating protein that negatively regulates Ras proteins and controls intercellular cAMP levels (Tanaka

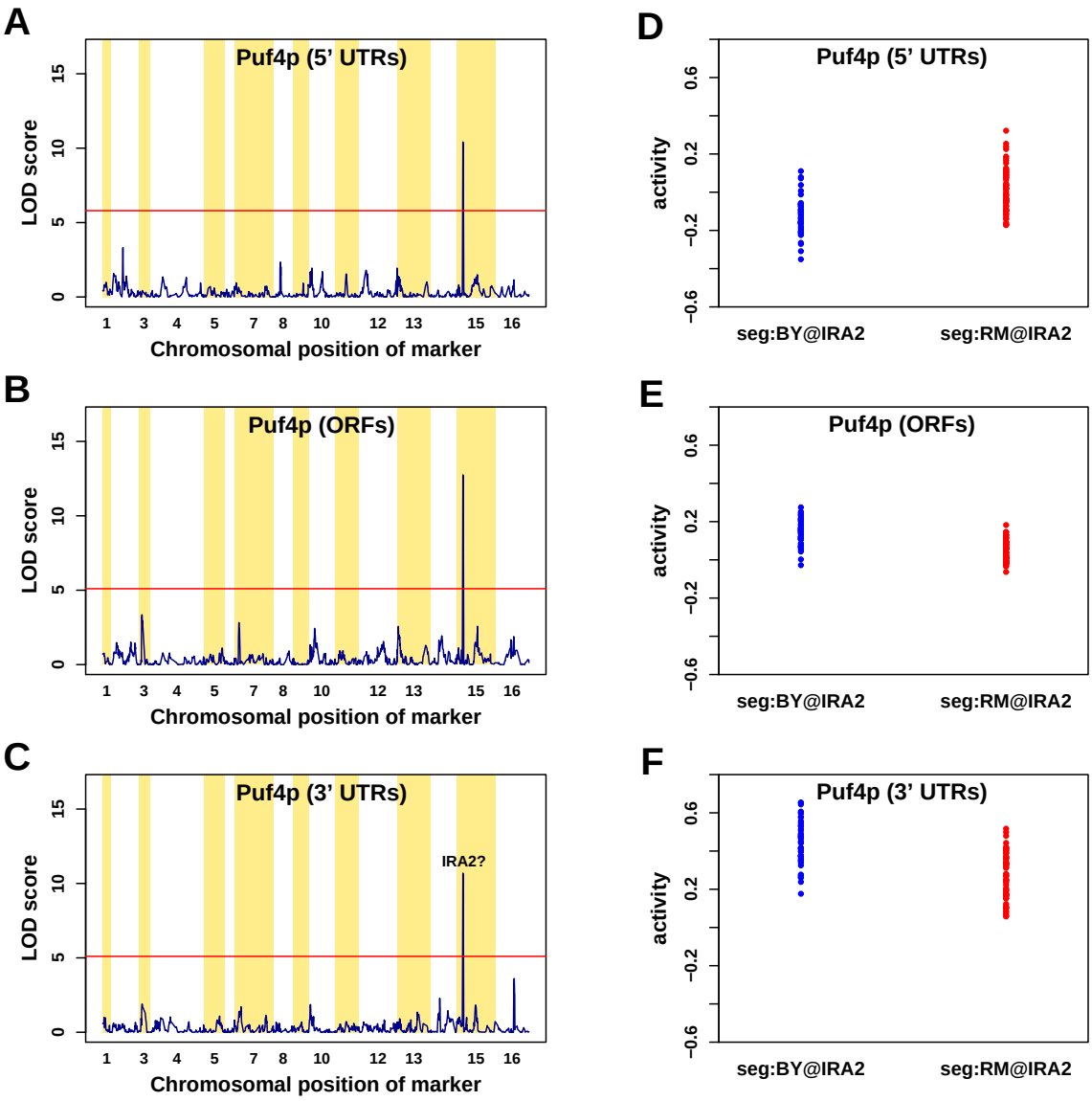

Figure 4.10: Puf4p aQTL Profiles. Results of the *trans*-acting genetic modulators of Puf4p activity levels mapped using our aQTL method. The significant thresholds at 1% FDR level were calculated using 200 independent permutations of the expression data (horizontal red lines). The peaks on chromosome XV survived after filtering out for the 3 groups of genes mentioned earlier. (A-C) show aQTL profile for Puf4p activity inferred from 5′ UTRs, ORF and 3′ UTRs affinity scores, respectively. Puf4p activity showed a significant linkage to a locus on chromosome XV irrespective of which mRNA region we used for affinity calculation. This locus includes the IRA2 gene. More interestingly, the sign of Puf4p activity levels in segregants inheriting BY and RM alleles at this locus switches between 5′ UTRs and 3′ UTRs.

et al., 1990). The mRNA expression levels for two stains were measured by (Smith and Kruglyak, 2008): a strain carrying RM IRA2 allele in BY background and the other strain carrying BY IRA2 allele in RM background. With our analysis we found that activity of Khd1p on 3′ UTRs, Puf2p on 5′ UTR and Puf4p on 3′ UTR are correlated to mRNA expression difference between the two mutant strains (Pearson t-values= -2.7, +4.4 and -5.1) at 1% FDR level (t-value= 2.5). The signs are for the case of subtracting the effect of RM allele from BY allele. This further validates the observation of the aQTL at this chromosomal location for Puf4p is caused by polymorphism is IRA2 and the role of the IRA2 gene as a putative modulator of the activity levels of Puf4p.

FLO gene family encodes cell-wall glycoproteins that regulate cell-cell and cell-surface adhesion (Guo *et al.*, 2000). When FLO11 is expressed, diploid cells form pseudohyphal filaments; whereas when FLO11 is silent, pseudohyphal differentiation and invasive growth is abolished (Lambrechts *et al.*, 1996). Khd1p represses FLO11 at the transcriptional level through its inhibition of ASH1 and at the post-transcriptional level by directly repressing translation of FLO11 gene (Wolf *et al.*, 2010)(Note: Khd1 binds repetitive pattern in ORF of FLO11). In a study on genetic regulation of the FLO gene family by (Halme *et al.*, 2004) they showed experimentally that mutations in IRA2 gene could potentially cause increase in FLO11 expression.

The second factor, which its activity was linked to this region on chromosome XV, is Puf2 (5′ UTR). It is reported that Puf2p preferentially interacts with mRNAs encoding membrane-associated proteins (Gerber *et al.*, 2004).

The third factor which has the highest correlation to IRA2 locus is Puf4p (3′ UTR). As mentioned previously, there are 2 genes with function related to ribosome are located in this region.

4.4 Conclusion

We have presented a method for identifying *trans*-acting genetic modulators of gene expression, which uses mRNA expression and genotyping data from a segregating population. We used this method to detect activity QTL (aQTL) of RNA-binding proteins (RBPs). The activities are inferred from RBP binding preferences and the expression data. The inferred activity levels of the RBPs are treated as quantitative traits and where mapped to the chromosomal marker using genotype data. Our method aims to identify post-transcriptional regulatory mechanism underlying genetic variation in gene expression levels.

We applied our aQTL method to a data set for 108 segregants from a genetic cross between two yeast strains (Smith and Kruglyak, 2008). We used RBP sequence specificities obtained by our motif discovery approach presented in Chapter 2. We calculated the affinity scores for the 25 RBP-region combinations. We detected 12 locus-RBP linkages out of which one was previously reported. We recovered the MKT1 locus on chromosome XIV as a putative modulator of Puf3p activity inferred from 3′ UTRs (Lee *et al.*, 2009). Interestingly, we found different loci as modulators of Puf3p when using the 5′ and 3′ UTRs, respectively. We also found IRA2 as a possible modulator of Puf4p activity.

Chapter 5

Modulators of Connectivity Between Transcription Factors and Their Target Genes

5.1 Introduction

This study focused on the detection of genetic loci whose allelic variation modulates the *in vivo* regulatory connectivity between a transcription factor (TFs) and its target genes. We call these loci connectivity QTLs or "cQTLs". Our method for discovering aQTLs incorporates the binding preferences of the TFs and mRNA levels of the individuals in a segregating population of yeast to infer activity levels of the TFs. We further use the activities to calculate the genotype-specific susceptibilities of each gene to TF activity variation existing between the individuals in the population. This variation occurs due to the allelic variation naturally inherited between the segregant of a genetic cross between two different yeast strains. Finally we used these genome-wide susceptibilities to TFs to construct a χ^2-statistic to identify the cQTLs.

The model used in this study is illustrated in **Figure 5.1**. Transcription factors bind to the promoter region of their target genes to initiate or repress DNA transcription

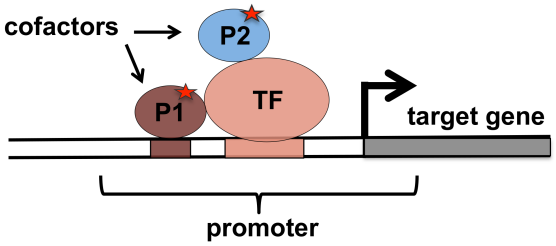

Figure 5.1: Connectivity Quantitative Trait Loci (cQTL) Model. Many TFs require cofactor proteins to tether effectively to the promoter region of their target genes (cofactor shown as P1). In these cases the cofactor proteins usually contain a DNA binding domain in their amino acid sequence allowing them to bind to DNA. In some other cases, cofactors may assist the recruitment or blocking of the binding of the transcription machinery to the transcription start site (cofactor shown as P2). This could be achieved by chromatin remodeling to make the RNA polymerase complex binding site accessible. Another situation (not shown) could be that TF activity is suppressed by a protein (i.e. inhibitor) where the detachment of the two will allow the binding of the TF to the promoter regions. Mutations (shown by red starts) in the amino acid chain of these cofactor could effect the efficiency of the transcription of the target genes by the TFs. We are interested to identify genetic loci (i.e. the chromosomal locations of these cofactors) whose allelic variation modulates the *in vivo* connectivity between the transcription factors (TF) and their target genes. We call these loci connectivity QTL or "cQTL".

to RNA. In yeast, promoter regions are segments of DNA typically about 600 base pairs long that are located upstream of open reading frame of each gene. These regions contain *cis*-regulatory elements through which TFs are able to regulate their targets.

Many TFs require cofactor proteins to tether effectively to the promoter region of their target genes (cofactor shown in **Figure 5.1** as P1). In these cases the cofactor proteins usually contains a DNA binding domain in their amino acid sequence allowing them to binding to DNA as well. An example for this case in yeast is the recruitment of Met4p to the promoter of genes involved in sulfur metabolism pathway (i.e. MET genes) with the help of Cbf1p (Thomas *et al.*, 1992). Met4p has been identified as the main transcriptional activator of this pathway and the expression of most, if not all, MET genes is strictly depends on Met4p recruitment to their promoters. It has been shown that several other cofactors besides Cbf1p enhance this recruitment including

Met28p, Met31p and Met32p (Blaiseau and Thomas, 1998; Siggers *et al.*, 2011).

In some other cases, cofactors may assist the recruitment or blocking of the binding of the transcription machinery to the transcription start site (cofactor shown in **Figure 5.1** as P2). Multiprotein bridging factor 1 (MBF1) is shown to mediate Gcn4p-dependent transcriptional activation by bridging the DNA-binding domain of Gcn4p and subunit of RNA polymerase II complex (Takemaru *et al.*, 1998). Gcn4p is the transcriptional activator of genes involved in amino acid synthesis. TFs can also facilitate the binding of transcription machinery by chromatin remodeling. A very well known case is yeast transcriptional initiation of galactose metabolism, a type of monosaccharide sugar. GAL genes are induced when galactose is present in the cell medium . Upon detection of transcription activation signal, Gal4p will facilitate the binding of RNA polymerase II to the GAL1 promoter by replacing the nucleosome blocking the TATA box on the promoter (Axelrod *et al.*, 1993).

Another situation (not shown in the figure) is when the TF is suppressed by a complex inhibiting it from binding to the promoter of its targets. In this case the cofactor act as inhibitor of TF activity. The detachment of the two will allow the recruitment of the TF to the promoter regions. Yeast mating response activation serves as an example for this case. Ste12p, activator of mating response in yeast, is bound by Dig2p in the absence of mating pheromones. This inhibits the binding of Ste12p to its target genes. In the presence of pheromone, Dig2p gets phorphorylated and detaches from Ste12p. This allows Ste12p to bind to the mating genes and initiate their transcription.

In all these cases, mutations (shown by red stars in the figure) in the amino acid chain of these cofactor could affect the efficiency of the transcription of the target genes by the TFs. Our analysis aimed to identify the causal mutations (cQTL). We mapped the DIG2 locus on chromosome IV as a cQTL for the transcription factor Ste12p. Dig2p is indeed a known inhibitor of yeast mating response activator Ste12p. The coding region of the DIG2 gene contains a single non-synonymous mutation (T83I). We are experimentally testing the functional impact of this mutation in allele

replacement strains. We also identified the TAF13 locus as a putative modulator of GCN4p connectivity.

5.2 Methods

5.2.1 Experimental Data Used

We analyzed genome-wide mRNA expression data collected by Smith and Kruglyak (2008). The data included mRNA levels of two strains of yeast: a laboratory strain (BY) and a wild isolate from a vineyard in California (RM) and 108 segregants produced from a BYxRM genetic cross. All samples were grown in two conditions: 2% glucose and 1% ethanol medium. The reference mRNA pool for microarray hybridization consisted of equal amounts of mRNA from both parents (BY and RM) grown in both glucose and ethanol conditions. For our analysis we used $\log_2$(sample/reference) for samples grown in glucose condition. We also used the genotype map of 2956 markers along yeast 16 chromosomes identified with oligonucleotide microarray performed by Brem *et al.* (2002).

5.2.2 Representation of Transcription Factors Promoter Binding Preferences

We used binding preferences for 124 transcription factors (TFs) in the form of position weight matrix (PWM) from MacIsaac *et al.* (2006). The elements in the PWMs represent the information about the nucleotide frequencies at each position in the set of target DNA binding sites. In that study, PWMs were trained on chromatin immunoprecipitation (ChIP) binding data for 172 TFs. The TFs binding data are collected by Harbison *et al.* (2004).The two motif finding algorithms used by MacIsaac *et al.* were designed to find evolutionarily conserved motifs among a set of genes co-regulated by a specific TF based on expectation-maximization (EM) algorithm and

local sequence alignment.

We used the convert2psam utility from the REDUCESuite v2.0 software package (see `http://bussemakerlab.org`) to convert each PWM to a position-specific affinity matrix (PSAM) (Foat et al, 2005, 2006; Bussemaker et al, 2007). The base counts at each position within the PWM were divided by that of the most frequent base to get an estimate for the relative affinity associated with each point mutation away from the optimal-binding sequence (Foat *et al.*, 2006). The resulting PSAM affinity scores on promoter sequences of genes were used to infer segregant-specific changes in TF activity levels.

5.2.3 Calculation of Segregant-Specific Promoter Affinity

For our analysis we only used 123 PSAMs from the study mention above. We excluded Hap3p PWM due to exact similarity to Hap5p PWM. To calculate TF specificities, we first downloaded 600 base pair nucleotide sequences upstream of open reading frame of every gene from *Saccharomayce cereviciae* Genome Database (SGD; `http://www.yeastgenome.org`) for the BY strain. This database is based on genetic information for the S288c yeast strain, a strain that is isogenic to BY with total of 39 single nucleotide polymorphisms (Schacherer *et al.*, 2007). We obtained the genetic sequences of the second parental strain from the Broad Institute website (`http://www.broadinstitute.org`) for the RM11-1a strain. We then used the Bioperl interface for the BLAST software (Altschul *et al.*, 1997) to identify pairs of orthologous genes between BY and RM by aligning the coding sequences of the two strains. We used 600 base pairs upstream sequences of each orthologous pair to define BY and RM specific promoter sequence. Using the genotype map we located the allele type of every gene for each of 108 segregants. This step was first implemented by Lee and Bussemaker (2010). Affinity scores (K) were calculated based on

Equation 2.7.

$$K_\phi(s) = \sum_{i=1}^{L_s - L_\phi + 1} K_{\phi i} = \sum_{i=1}^{L_s - L_\phi + 1} \prod_{j=1}^{L_\phi} w_{\phi j b_{i+j-1}}(s) \tag{5.1}$$

Here, ϕ is an index over proteins, s is for sequences, L_s is for the length of the sequences, L_ϕ is the binding site lengths, w stands for weight matrix elements and b denotes the nucleotide type at position $i + j - 1$. We first used the genotype map and genetic chromosomal coordinates from SGD to build an allele map for genes-segregants. Using this map, we then calculated the promoter affinity scores for each TF for every segregant using the corresponding parental promoter sequence.

5.2.4 Inferring TFs Activity Levels

As demonstrated by **Equation 2.7**, the *in vivo* mRNA steady state differential levels between one of the sample segregant and parental reference pool are proportional to the sequence affinity. The occupancy itself is proportional to the *in vitro* promoter affinity considering a low protein density regime. In this regime we can define a protein's activity level as correlation between predicted promoter affinities and the genome-wide transcriptional response of on of the segregants as shown in **Figure 5.2**. The slope represents the activity level of the protein. For the expression levels in the plot, the promoter affinity explains only 2% of the variance in the signal ($r^2 \sim 0.02$).

Extending this model to include genotyphic variation between the segregants in the population is shown in the equation below (From (Lee and Bussemaker, 2010)).

$$\begin{aligned} \log_2([\mathrm{mRNA}_g]_{\mathrm{sample}}) - \log_2([\mathrm{mRNA}_g]_{\mathrm{ref}}) &\propto N_{\phi g,\mathrm{sample}} - N_{\phi g,\mathrm{ref}} \\ &\approx [\phi]_{\mathrm{sample}} K_{\phi g,\mathrm{sample}} - [\phi]_{\mathrm{ref}} K_{\phi g,\mathrm{ref}} \\ &= ([\phi]_{\mathrm{sample}} - [\phi]_{\mathrm{ref}}) K_{\phi g,\mathrm{sample}} \\ &+ [\phi]_{\mathrm{ref}} (K_{\phi g,\mathrm{sample}} - K_{\phi g,\mathrm{ref}}) \end{aligned} \tag{5.2}$$

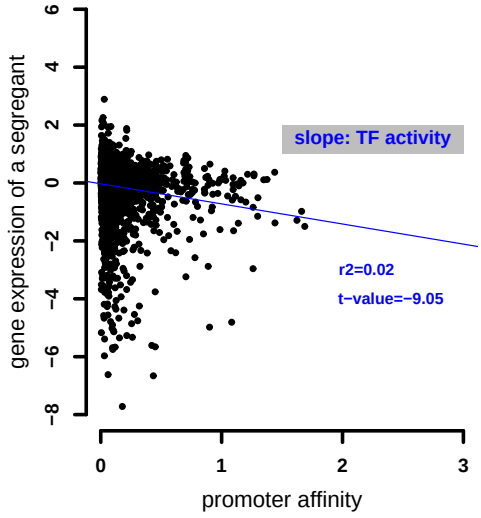

Figure 5.2: Inferring Transcription Factor Activity Level From Predicted Promoter Affinity and Genome-Wide Regulatory Response. In low protein regime, one can assume a linear relation between mRNA expression levels and predicted promoter affinity for a TF. The slope represent the activity level, which is a function of the protein's concentration. Each point in the scatter plot represents a gene.

$[\phi]_{\text{sample}}$ and $[\phi]_{\text{ref}}$ are the protein concentrations in the sample and the reference pool respectively. Note that the first term captures all the *trans*-acting modulators that cause differences in the activity level of the protein. These effects could be mutations in the nucleotide sequence of the gene encoding protein ϕ which can result in mutations in protein's amino acid sequence, Polymorphism in the protein's cofactors amino acid sequence resulting in variation in their interaction efficiency and finally dissimilarity in chromatin content between the sample and the reference pool. The *cis* effects are absorbed by the second term in the above expression. It accounts for the differences in the nucleotide sequence of the preferred binding site on the promoter region of target gene g of protein ϕ. We can rewrite the last equation using *trans* and *cis* terms (Lee and Bussemaker, 2010). This regression is illustrated in **Figure 5.4B**.

$$y_{gs} = \beta_{0s} + \sum_{\phi} \beta_{\phi s}^{\text{trans}} K_{\phi gs} + \sum_{\phi} \beta_{\phi s}^{\text{cis}} (K_{\phi gs} - \langle K_{\phi g} \rangle_{\text{ref}}) \tag{5.3}$$

Here y_{gs} is the $\log_2$-ratio of the mRNA levels of gene g between the sample and the reference pool. Since the reference pool is a mixture of equal amounts of parental strains, the term $\langle K_{\phi g} \rangle_{\text{ref}}$ is equal to the average of BY and RM promoter affinities.

Several of the TFs in the MacIsaac *et al.* (2006) collection are involved in the same or interacting complexes such as Met4p and Met32p, both part of sulfur metabolism pathway, or Msn2p and Msn4p, both involved in the yeast stress response pathway. This means that the proteins related together through a protein complex or genetic pathway, regulate the same target genes and their affinities are highly correlated. To circumvent this multicolinearity issue among the promoter affinities, we used multiple ridge regression (Hoerl and Kennard, 1970) instead of a multiple linear regression to calculate $\beta_{\phi s}^{\text{trans}}$ and $\beta_{\phi s}^{\text{cis}}$. Ridge regression minimizes simultaneously the residual sum of squares and a penalty term for a parameter λ to estimate the model parameters. In this case, the predictors are slightly biased but more precise (variances are smaller than with the Least-Squares method).

$$\{\{A\}, b\} = \text{argmin}(\|y - b - AX\|^2 + \lambda \|A\|^2), \qquad \lambda > 0 \tag{5.4}$$

We used $\lambda = 0, 0.002, 0.004, ..., 1$ and chose the λ that resulted in minimum cross-validation. We only used $\beta_{\phi s}^{\text{trans}}$ representing inferred activity levels for all further analyses.

As we will explain in the next section, we used the inferred activity levels together with mRNA expression data to calculate susceptibilities (X) to every protein ϕ activity variation for each gene g. Using expression data of all genes in activity calculation step, will result in circularity between susceptibilities and mRNA expression level of each gene. To avoid this situation, we eliminated one gene at a time from the mRNA expression data and affinity promoter data sets. We label this new gene set with $\{-g\}$, indicating the elimination of gene g. We then performed ridge regression to obtain $\{\beta^{\text{trans}}\}_{\{-g\}}$ and $\{\beta^{\text{cis}}\}_{\{-g\}}$ on the expression data and affinity data for the rest of the genes. As we will discuss later in the result section, this step was crucial in our analysis.

5.2.5 Calculation of Genome-Wide TF Susceptibilities

We used the mRNA expression data and the activity levels to infer the susceptibilities (X). $X_{\phi g}$ is a measure of connection or responsiveness of gene g to the variation in activity levels of protein ϕ. In other words, it is the partial derivative of the expression level of gene g with respect to the activity level of protein ϕ:

$$X_{\phi g} = \frac{\partial y_g}{\partial \beta^{\text{trans}}_{\phi s,\{-g\}}} \tag{5.5}$$

Since the explicit form of y_g as a function of the activity is not known, we assume a linear relationship as a first order approximation. We applied multiple ridge regression between mRNA levels of gene g and activity levels among the segregants. This regression is illustrated in **Figure 5.4C**.

$$\{\{X_{\phi g}\}, b_g\} = \text{argmin}(\sum_s (y_{sg} - \sum_\phi (\beta^{trans}_{s\phi,\{-g\}} X_{\phi g}) - b_g)^2 + \lambda \sum_\phi X^2_{\phi g}) \tag{5.6}$$

We used the same range of values for the λ parameter as reported in the previous section. The regression coefficients represent the inferred susceptibilities. As we explained earlier, the susceptibility of each gene g to factor ϕ is obtained from activity levels that were calculated independent of mRNA levels of that same gene. This step was necessary to avoid any circularity in calculation of the susceptibilities.

5.2.6 Selection Criteria for TFs Based on the Inferred Susceptibilities

Out of 123 TFs, we accepted only those for which susceptibility $X_{\phi g}$ was highly correlated to their estimated promoter affinity scores. If we assume that the susceptibilities represent functional connection and promoter affinity scores represent biophysical connection to protein ϕ activity levels, one would expect high correlation between the two. For promoter affinities, we used the average affinity of the BY

and RM strains for each gene. We then calculated t-values for Pearson correlation between the susceptibilities and the promoter affinities. We accepted a TF only if this correlation was above 1% FDR level and was the highest t-value compared to correlation to the affinities of all the other 122 factors (See **Figure 5.6**).

The factors, which passed the criteria described above, also showed exclusive correlation of susceptibilities to affinities when multiple ridge regression was replaced with a univariate linear regression. Since Ridge implementation of R first normalizes the columns of the independent variable (i.e. inferred activities), we decided to use the normalized activities for the case of univariate regression to be consistent when comparing the results of the two cases.

5.2.7 Functional Validation of Selected TFs

As validation for our selection process, We used genome-wide mRNA levels over a time course of controlled over-expression of some of the factors from (McIsaac *et al.*, 2013). The measurement is based on transcriptional activator complex Gal4p-DBD.ER.VP16 (GEV). GEV is constructed by fusing the DNA-binding domain (DBD) of Gal4p into the human estrogen receptor (ER) and portion of herpes simplex virus protein VP16. The GEV construct is active only in the presence of β-estradiol. This hormone results in localization of inactive cytoplasmic GEV into the nucleus. GEV then binds to the Gal4p consensus upstream activation sequence (UAS_{GAL}), which was infused in the promoter region of the desired genes and causes the over-expression of its mRNA within minutes following hormone addition to the cells culture (McIsaac *et al.*, 2011). We calculated the correlation of inferred susceptibilities to the genome-wide mRNA levels at different time points, starting from time=0 when the over-expression of factor ϕ was induced. A different construct was used for the over-expression of Gcn4p, where a zinc-finger binding domain was fused to the ER.VP16 complex (ZEV) and the promoter region of the GCN4 gene was modified to contain the ZEV-binding sites (McIsaac *et al.*, 2012).

We also performed Gene Ontology (GO) enrichment analysis on the genome-wide inferred susceptibilities to each TF by a univariate linear regression. We calculated Wilcoxon-Mann-Whitney p-value and Pearson t-value for GO categories with at least 10 members. We used Benjamini and Hochberg (1995) method to correct for multiple testing at 1% FDR level as significance threshold. The Benjamini-Hochberg method tries to estimate the expected fraction of false positives based on the data size and p-values. We first sorted the p-values for all genes in increasing order: $p_1 \leq p_2 \leq \cdots \leq p_m$. For a desired false discovery rate (FDR) at q, k is the largest i for which

$$p_i \leq \frac{i}{m} q \tag{5.7}$$

In our case, m is the number of GO categories equal to 1891 and q is equal to FDR level of 1%. All GO categories with p-value larger than p_k were considered insignificant. We used an iterative procedure for removing the effect of redundant nested GO categories that was implemented originally in the T-profiler algorithm by Boorsma *et al.* (2005) (see **Section 2.2.6**).

5.2.8 Defining positive and negative target sets for the TFs

We used the susceptibilities of the selected factors to define the set of positive and negative target sets. For each gene and protein ϕ, we first obtained the p-value of the regression coefficient explained in the previous section. For corrected for multiple testing based on the method from Benjamini and Hochberg (1995). In this case the number of tests was equal to the number of chromosomal markers and FDR level was set to 1%. We grouped the significant targets into positive and negative sets based on the sign of the susceptibilities to be used in the cQTL discovery analysis.

5.2.9 Calculation of Genotype-Specific Susceptibilities to TFs

Once we had filtered the TFs based on the criteria discussed above, we calculated genotype-specific susceptibilities to the selected TFs, X^{BY} and X^{RM}, for every gene g at each marker location m. To do this, we first split the segregants based on their genotype at a marker m. Then we separately calculated the univariate linear regression coefficient between mRNA levels of gene g and activity levels of protein ϕ for each segragant subset, as indicated in **Figure 5.3**.

$$\{\{X^{\text{geno}}_{\phi gm}\}, b\} = \text{argmin}(\sum_{\{s\}_{\text{geno@}m}} (y_{sg} - \beta^{trans}_{s\phi,\{-g\}} X^{\text{geno}}_{\phi gm}) - b)^2) \tag{5.8}$$

Here geno refers to the parental genotype: BY or RM. Note that the activity levels were inferred from gene set missing gene g as explained earlier. We performed this step for every gene g at every marker m.

5.2.10 cQTL Discovery Using χ^2-statistic

We used χ^2-statistic to check whether the susceptibilities to factor ϕ are significantly different when splitting the segregants based on their genotype at marker m. We first calculated $\Delta t_{\phi gm}$ for every gene at every marker as given by the equation below. SE stands for the standard error of the slope from the univariate regression explained in the previous section.

$$\Delta t_{\phi gm} = \frac{X^{BY}_{\phi gm} - X^{RM}_{\phi gm}}{\sqrt{(SE^{BY}_{\phi gm})^2 + (SE^{RM}_{\phi gm})^2}} \tag{5.9}$$

Next we calculated the χ^2-statistic for each marker m by squaring the Δt's and summing them for all genes given by

$$\chi^2_{\phi m} = \sum_{g=1}^{N_g} \Delta t^2_{\phi gm} \tag{5.10}$$

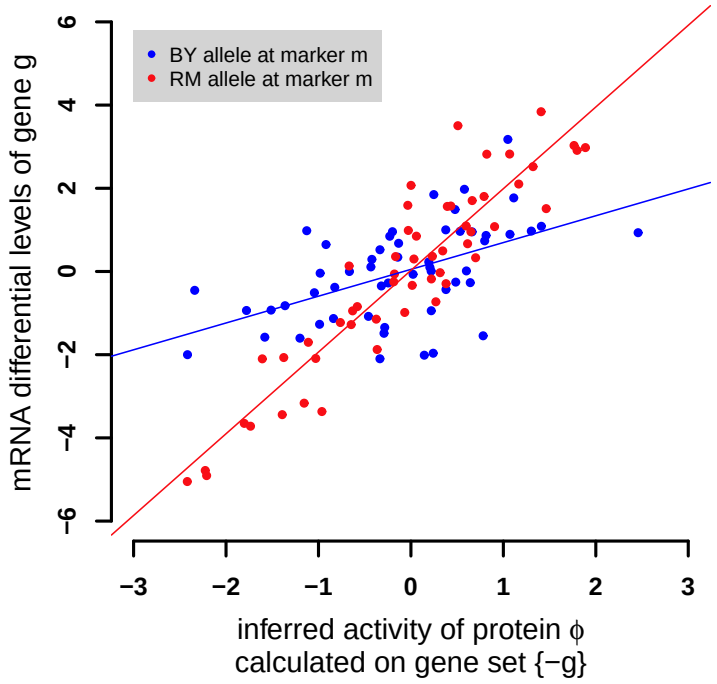

Figure 5.3: Calculation of Genotype-Specific Susceptibilities X on Synthetic Data. The susceptibilities are calculated at each marker first by splitting the segregants based on their genotype: BY or RM at that marker. Then $X^{BY}_{\phi gm}$ and $X^{RM}_{\phi gm}$ are calculated by a univariate linear regression of mRNA levels of a gene g to the activity levels of protein ϕ on the two segregants subset separately. Note that the activity levels were calculated on the gene set missing the gene g ($\{-g\}$). Each point in this plot represents one of the 108 segregatns.

where N_g is the total number of genes. If there is no significant linkage of the susceptibilities to the loci m for one of the segregant subsets, then $\Delta t_{\phi gm}$ is expected to behave like a standard normal random variable. In other words, with no linkage the value of χ^2-statistic is expected to equal to N_g. So by calculating χ^2-statistic at this marker, we can have a single measure to test the significance of this loci for modulation of connection between the protein ϕ and its targets.

Finally, we performed a forward selection to extract significant markers similar to method explained in **Section 3.2.4**. It tries to iteratively select the loci that contribute significantly to the χ^2-statistic marker profile. At each iteration the effect of the previously selected markers are removed from Δt's and the residuals are used for the next iteration. We performed this selection rounds until all residual χ^2 values

fell below a significant χ^2 threshold. To calculate this threshold, we used Bonferroni correction at 1%. To define the significant cQTL region, we extended the region around each selected marker in both direction until hitting the significant χ^2 threshold value.

For each TF, ϕ, this entire procedure was done for three gene sets: all 4482 genes, positive targets and negative targets.

5.2.11 Protein-Protein Interaction Data

To identify putative causal genes, or quantitative trait genes (QTGs), within a cQTL, we downloaded yeast protein-protein interaction dataset from the Biogrid website (`http://thebiogrid.org`) as of April 2012. We only looked into interactions that involved physical interaction between the TF and other proteins (i.e. cofactors). We considered only those cofactors where the gene encoding them were located in the detected significant cQTL regions. We identified novel putative modulators of the TF-target connectivity for two factors, Ste12p and Gcn4p. We will discuss these findings in the results section of this chapter.

5.2.12 Gene Ontology Analysis on Δt of Detected Loci

Finally we performed Gene Ontology (GO) enrichment analysis on $\Delta t_{\phi gm}$ for each selected marker m. This step was applied to Ste12p and Gcn4p only. We calculated Wilcoxon-Mann Whitney p-value and Pearson t-value for GO categories with at least 10 member genes. We used the Benjamini-Hochberg method to correct for multiple testing at 1% FDR level as significance threshold. Again, we used an iterative procedure for identifying the enriched GO categories.

5.3 Results

The goal of the analysis presented in this chapter is the detection of proteins modulating the strength of the functional connection between a transcription factor protein and its target genes. We surveyed genome-wide mRNA expression data for a collection of yeast strains measured by (Smith and Kruglyak, 2008). The data included differential mRNA expression levels of 108 segregants produced by a genetic cross between two yeast strains mating: a laboratory strain (BY) and a wild strain from a vineyard in California (RM). The mRNA abundances were measured with DNA microarray relative to a reference pool consisting equal amount of both parental strains grown both in glucose and ethanol. For our analysis we only used the expression data of segregants grown in glucose. We also use the segregants genotype map for 2956 chromosomal marker location (Brem *et al.*, 2002).

5.3.1 Inferring Segregant-Specific TFs Activity

Our method is illustrated in **Figure 5.4**. As prior knowledge, our method uses the transcription factors' (TFs) binding preferences to *cis*-regulatory element on the promoter regions of genes **Figure 5.4A**. To predict the binding specificities we used a compendium of weight matrices representing binding preferences for 123 TFs (MacIsaac *et al.*, 2006) and 600 base pair upstream sequence of each gene. We calculated the upstream affinity scores by summing the scores of a sliding window along the upstream sequence of each gene. We calculated the genotype-specific promoter affinities for every segregant based on their inherited allele (see Methods).

Figure 5.4B depicts the step for inferring the segregant-specific transcriptional activity levels. This step uses the matrix of expression data whose rows correspond to genes, and its columns contain the genome-wide differential mRNA levels of one of the 108 segregants. Using the expression data and the predicted *in vitro* binding affinities, we calculated the activity levels by performing a linear regression between the two. The TF activity levels of a particular segregant was obtained by calculating

the slope from regressing the expression data on the promoter affinities. As will be clear, using these inferred activities for estimating the genome-wide susceptibilities creates a substantial circularity. To avoid this, we eliminated the rows corresponding to a particular gene in the expression data and affinity data. We then regressed this new expression dataset to the new affinity dataset. This way, the activity levels that will be used to estimate the susceptibility of a gene are independent of the mRNA level of that gene. The effect of this elimination step on susceptibilities is shown in **Figure 5.5**. The variation of inferred activity levels between the segregants reflects the transcriptional circuit differences among them. Mutations in the amino acid chain of a TF and/or its cofactor as well as differences in the expression level of the cofactors, and variations in the chromatin state near the binding site can all create the observed variations in the mRNA levels between the segregants (*trans*-effects). This can also be the result of critical mutations in nucleotide sequence of the preferred binding site on the promoter region by the TF (*cis*-effect). We used the portion the activity level of the TFs that are modulated by *trans*-effects, β^{trans}, in our subsequent analyses (see **Equation 5.3**).

Figure 5.4C explains the details of genome-wide susceptibility calculation step of our method. This step calculated the susceptibility of every individual gene to the variation of TF activity level. We performed a linear regression of mRNA levels of a gene to the inferred TF activity levels across all segregants. The slopes from the linear regressions represent the susceptibilities. These susceptibilities in turn were used to select the set of TFs whose functional connectivity (i.e susceptibility) is correlated with their biophysical connectivity (i.e. promoter affinity).

After the selection step, the genotype-specific susceptibilities are calculated by splitting the segregants at each chromosomal marker based on their inherited parental alleles and separately regressing the mRNA levels to the activity levels (see **Figure 5.4D**). The genotype information is represented by a matrix whose rows correspond to one of the 2956 markers, and whose columns contain the allele-type of the loci for a particular segregant. The differences in the susceptibilities of the two subset

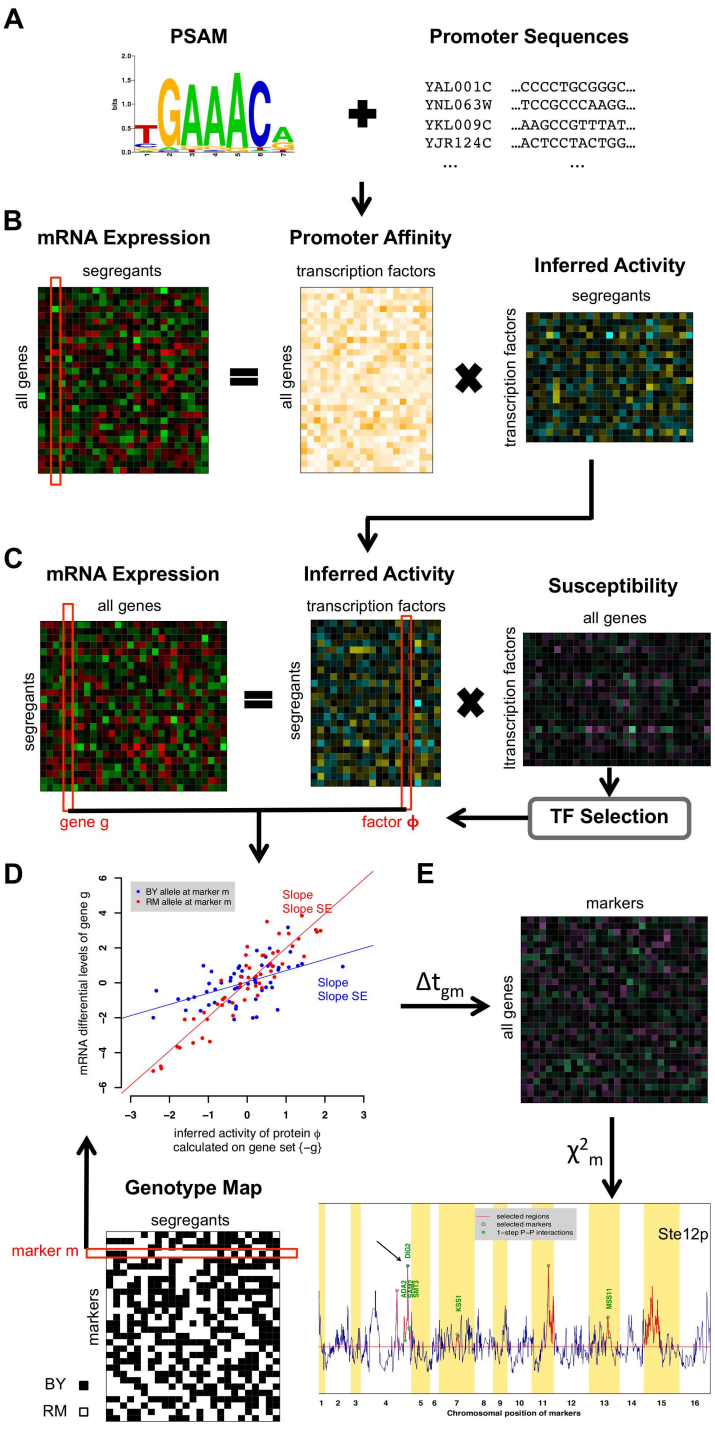

Figure 5.4: Connectivity Quantitative Trait Loci (cQTL) Detection Method. (A) Our method uses binding preferences in the form of position specific affinity matrix (PSAM) of 123 yeast TFs as prior information. We start by calculating the affinity scores of the PSAMs on the promoter region of every gene. We used 600 base pairs upstream of the begining of the coding region of each gene as an estimate for its promoter region. This is done by summing the affinity scores of a window sliding along each

Figure 5.4: (Cont. Caption) upstream sequence. (B) Next, we perform a linear regression of the genome-wide expression data on the estimated promoter affinity scores. We do this regression separately for each segregant. The regression slope is the inferred protein activity level. Also to prevent any circularity in further calculations, we eliminated one gene at a time (i.e. excluding a row from the segregants expression matrix and affinity matrix) and applied the linear regression on the reduced gene set $\{-g\}$. In other words, we obtained the activity matrix for $\{-g\}$ gene sets for G times. Here G is the total number of genes in the gene set. (C) In this step, we infer the susceptibility (i.e. responsiveness) of expression of each gene to the variation in the TF activity levels. We do this by applying a linear regression between the expression data of each gene g and the activity matrix calculated on the gene set $\{-g\}$. The slope of the fit represents the susceptibility of the gene to activity levels of the TF. The inferred susceptibilities are used to select the set of TFs for the next step. The selection is based on correlation of susceptibilities and promoter specificities of the TFs (see Methods). (D) For each selected TF, we calculate the genotype-specific specificity by splitting the segregant based on their inherited parental allele type (BY or RM) at a marker m and performing a linear regression on each segregant subset between the expression levels of each gene g and activity levels of each selected TFs calculated on the gene set $\{-g\}$. We conducted this step for every gene and every marker. (E) Using the genotype-specific susceptibility data, we construct the Δt matrix where each element is a standardized susceptibility difference for each gene at each marker. The last step involves the calculation of a χ^2-statistic for each marker by summing the squared Δt's of all genes at each marker. Significant peaks in the χ^2 profile are identified as cQTLs.

are treated as quantitative phenotype for our method.

The final step involves the calculating of a χ^2-statistic at each marker by summing the squares of z-scores of the susceptibility differences. The χ^2 value quantifies the significance of the global susceptibility differences at a marker. When these differences depend on the inherited allele at a loci for significant number of genes, then it implies that the locus is influential, and indicates the presence of a cQTL (see **Figure 5.4E**).

5.3.2 Selected TFs Based on Their Susceptibilities

As we discussed in the previous section, using all genes to calculate the activity levels creates a circularity in the susceptibility calculations. To avoid this, we eliminate one

gene at a time from the expression data and promoter affinity data and performed a linear regression on the new data sets. The inferred activity levels are independent of the mRNA levels of the eliminated gene. This removes the circularity when using the activity levels for susceptibility calculation. **Figure 5.5** demonstrates the effect of the gene-elimination step on the susceptibility levels. When including all genes to infer the TF activity levels, we observed a significant correlation for the promoter affinity and susceptibility to a particular TF as shown in **Figure 5.5A**. We displayed the correlation levels in a heatmap structure where the rows correspond to the TF affinity scores and columns correspond to susceptibilities to TFs.

Figure 5.6 summarizes the selection step results. The selection criteria are based on the exclusive correlation of functional connectivity (i.e. susceptibilities) and the biophysical connectivity (i.e. promoter affinities) of the TFs. We tested this by calculating the Pearson correlation t-values between the genome-wide susceptibilities and promoter affinities. **Figure 5.6A** shows the correlation results for the case where the susceptibilities were calculated by a multiple ridge regression of the expression level of a gene on the activity level of all TFs. The x-axis represents the susceptibilities to each of the 123 TFs. For a particular TF, if the regulation is exclusive then the genome-wide susceptibilities to its activity variation should most significantly correlate to the promoter affinity scores of that TF (i.e. the red dot should be above blue dots for that TF). The green line represents the t-value threshold at 1% FDR level using Benjamini-Hochberg method for multiple testing correction (t-value $\sim$ 4.22, p-value $\sim 3.9 \times 10^{-6}$). For some of these TFs the susceptibilities were poorly correlated to their promoter affinities indicating that the data and the model were not adequate to capture the correct susceptibility levels. For 12 of the TFs, we measured a high and exclusive correlation between susceptibilities and their promoter affinities. This list includes Chap4p, Gcn4p, Hap4p, Ino4p, Leu3p, Met32p, Msn2p, Pdr3p, Rcs1p, Stb5p, Ste12p and Sut1p. As mentioned earlier, the susceptibilities to these TFs were calculated using multiple ridge regression. The next step requires the calculation of genotype-specific susceptibilities for each gene and every marker. By performing this regression we obtain the regression coefficients for each segregant

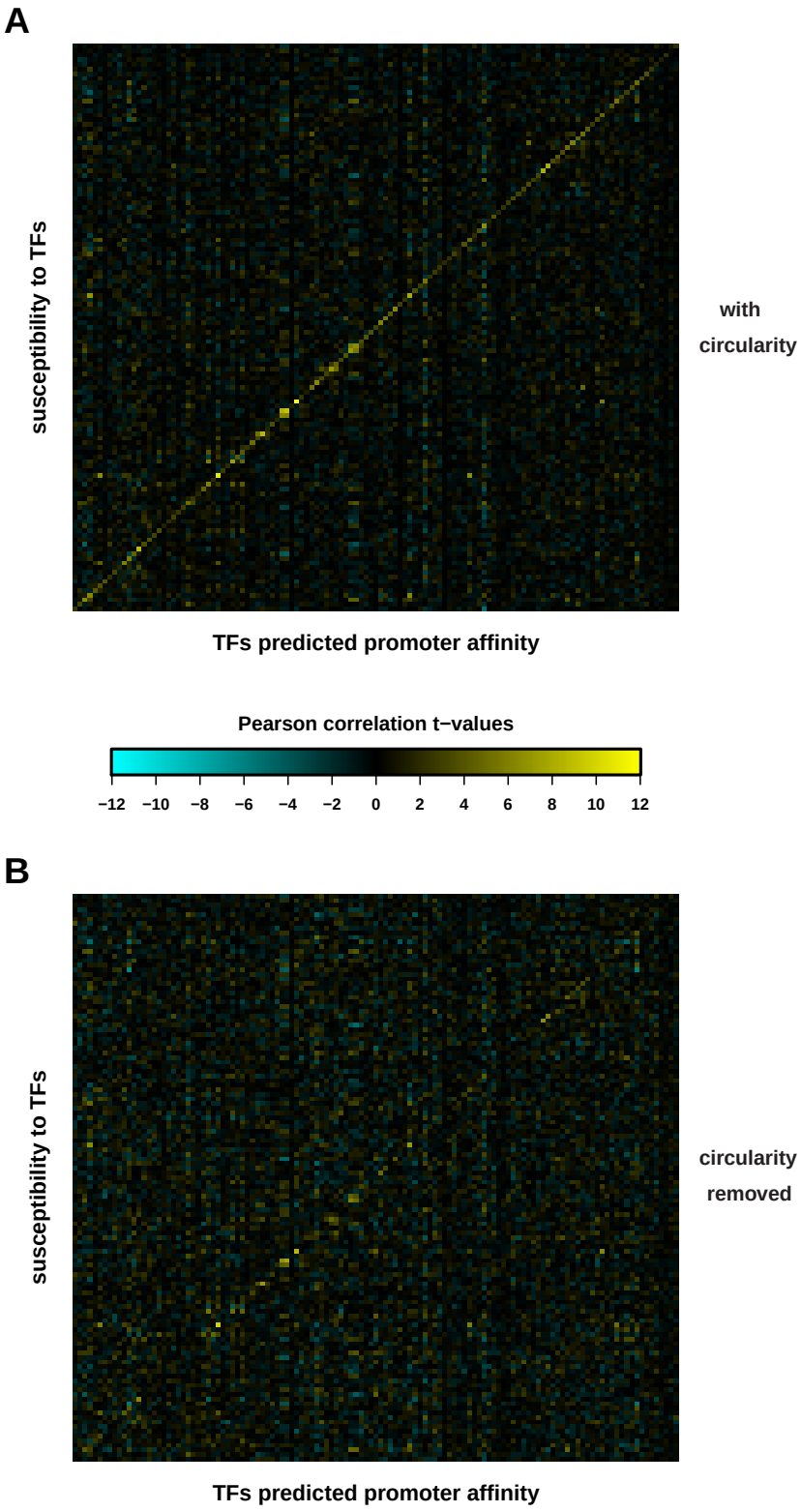

Figure 5.5: Resolving the Circularity for Susceptibility Calculation. Susceptibilities were obtained by a multiple linear regression between the mRNA levels of a gene

Figure 5.5: (Cont. Caption) g and activity levels of a TF. The activity levels of the TFs were inferred using the expression levels of (A) all of the genes and (B) all of the genes excluding a gene g to be used for obtaining susceptibility of that gene.

subset. However we also need the standard error of the slope to calculate the Δt matrix given by **Equation 5.9**. One possible approach to calculate these errors is to use bootstrapping. However doing so is computationally intensive. For this reason, we decided to also obtain the susceptibilities by performing a univariate regression between the mRNA expression of each gene g and the activity levels of a particular TF. These susceptibilities were then used in our selection step. The results are shown in **Figure 5.6**. In this case only 12 TFs passed our criteria: Cha4p, Gln3p, Gcn4p, Ino4p, Leu3p, Mcm1p, Msn2p, Msn4p, Rcs1p, Sip4p, Ste12p and Swi5p. In the case of Msn2p, the TF that correlated the most was Msn4p. However these two proteins are know to be involved in stress response activation and both bind to promoter genes containing stress response element (Martinez-Pastor *et al.*, 1996; Schmitt and McEntee, 1996). Schmitt and McEntee showed that Msn4p can partially compensate for the lack of Msn2p function in MSN2Δ mutant yeast strains.

We accepted a total of 7 TFs that passed our criteria in both cases. This list contains Cha4p, Gcn4p, Ino4p, Leu3p, Msn2p, Rcs1p and Ste12p. Meeting the first criterion (see **Figure 5.6A**) ensures that the significant correlation between the susceptibilities and the affinity scores for each of these 7 TFs is not due to over-fitting. Meeting the second criterion (see **Figure 5.6B**) guarantees that the susceptibilities to each of these TFs can be calculated in a manner independent of the transcriptional regulation by other TFs. In other words, the usage of a univariate linear regression for the calculation of the susceptibilities is acceptable.

Figure 5.7 displays the scatterplot of the susceptibilities and the promoter affinity scores for Gcn4p and Ste12p. In both cases, The two quantities were highly correlated.

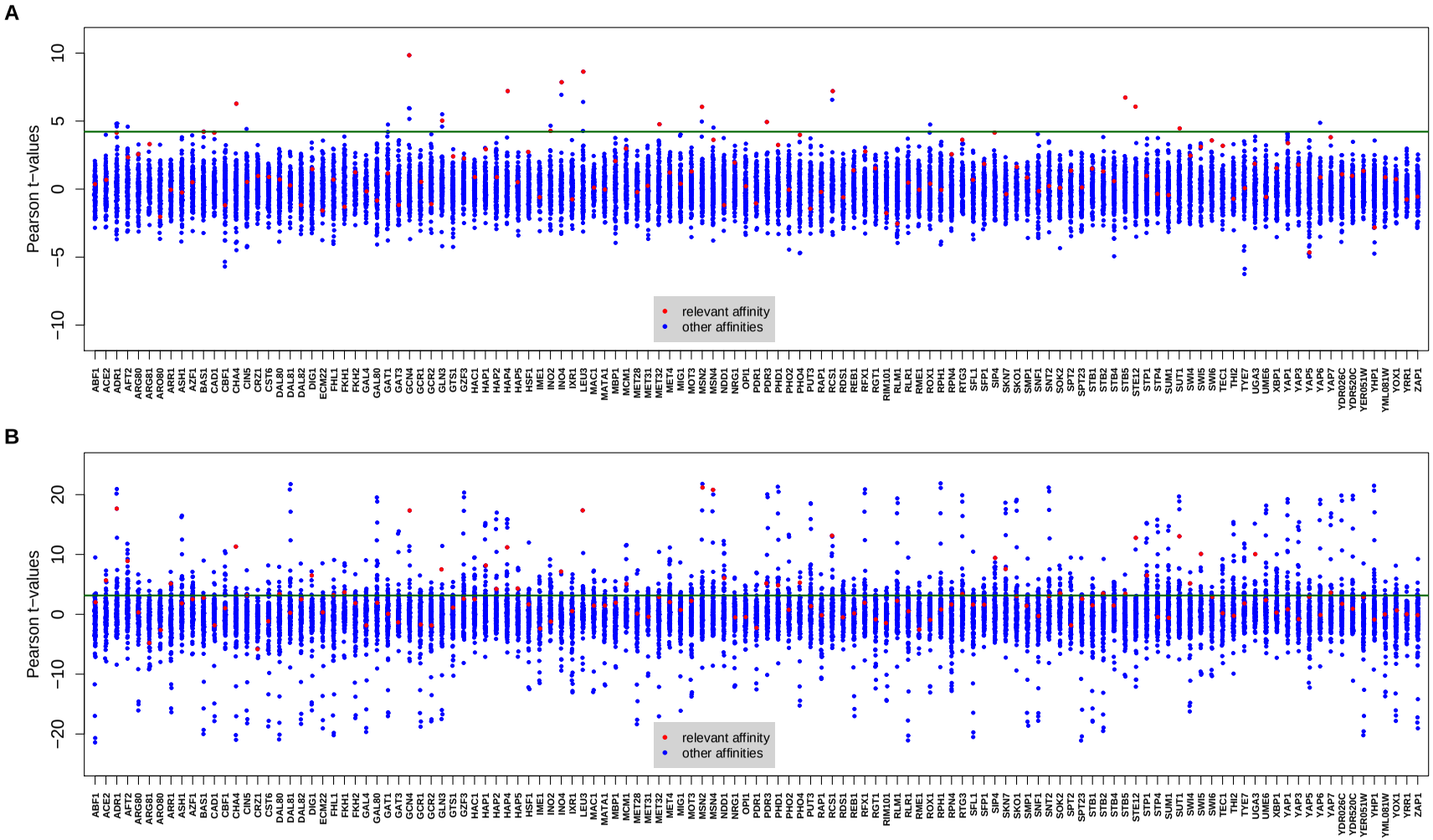

Figure 5.6: Univariate Pearson Correlation t-values Between TFs Susceptibilities and Their Promoter Affinities. The susceptibilities were derived using (A) multiple ridge regression, and (B) univariate regression. The x-axes represent the susceptibilities of each of the 123 TFs from MacIsaac set. For each TF, the t-values of Pearson correlation between the susceptibilities and its promoter affinities (relevant affinity) are pointed out with red dot and to the rest of 122 TFs affinity with blue dots. We accepted the TFs whose susceptibility significantly and exclusively correlated to their promoter affinity (i.e. the red dot stands out from the rest of blue dots). The dark green line represents the significant t-value threshold at 1% FDR level.

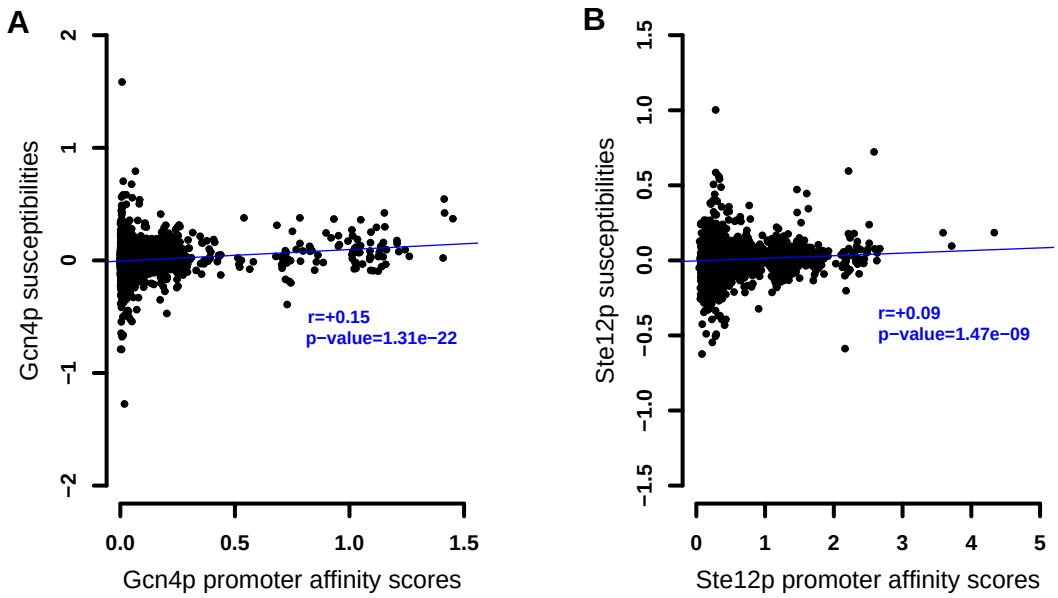

Figure 5.7: Susceptibilities Versus Promoter Affinities Scatter plots for (A) Gcn4p and (B) Ste12p. The susceptibilities were obtained by applying a multiple ridge regression between the mRNA expression levels of each gene and all of the activity levels of the TFs.

5.3.3 Functional Validation of Susceptibilities for the Selected TFs

In this section we present the results of functional validation of the susceptibilities to the selected TFs by using two different approaches.

Correlation to TF Over-Expression Data

We obtained genome-wide mRNA levels of a time series measurement where the mRNA of a particular TF was over-expressed (McIsaac and others; unpublished).In that study, they controlled the over-expression of the mRNA of genes including GCN4, MSN2 and RCS1 based on transcriptional activation constructs GEV and ZEV (see Methods). This construct becomes active with the addition of β-estradiol hormone to the culture and induces the expression of the listed genes. Upon this activation, we expect to observe changes in the direct targets of these TFs and as time goes on

the indirect targets get induced as well. We used this data to validate our inferred susceptibilities to Gcn4p, Msn2p and Rcs1p.

Figure 5.8 presents the correlation of the over-expression data measured at different time points to the affinities/susceptibilities to 123 TFs. The affinities/susceptibilities to the TF that was over-expressed, are indicated in red. We see that in all three cases, Gcn4p, Msn2p and Rcs1p, the correlation of their *in vitro* occupancy (i.e. affnity) to their *in vivo* functionality (i.e. over expression data) is significant and the highest among all other TF correlation (see panel A, C and E). This also justifies that the DNA-binding specificities in the form of PSAMs contain relevant functional information (Gao *et al.*, 2004). For all three TFs, the correlation of the susceptibilities to over-expression data improves at later times. This might be due to the fact that there is a time delay between the activation onset of the GEV (or ZEV) construct and the synthesis of the transcription factor of interest (i.e. translation); Furthermore, only the direct targets of the TF are induced at first and then its indirect targets are influenced. So it takes some time for the system to each its equilibrium state. Whereas the susceptibilities were inferred from a steady-state condition where the system is at equilibrium, the correlation improves for the later time points for all of the 3 cases. We observe that the susceptibilities have significant correlation to the over-expression data for each of the TFs.

Figure 5.9 displays the scatter plots for these 3 TFs at 6 different time points after the addition of the β-estradiol hormone to the culture. As shown also here, in all three cases the correlation between the over-expression data and the susceptibilities is relatively weaker at earlier time points and improves with time. We can conclude that our inferred susceptibilities do capture the functional connectivity of the genes to the activity level variations for the case of Gcn4p, Msn2p and Rcs1p.

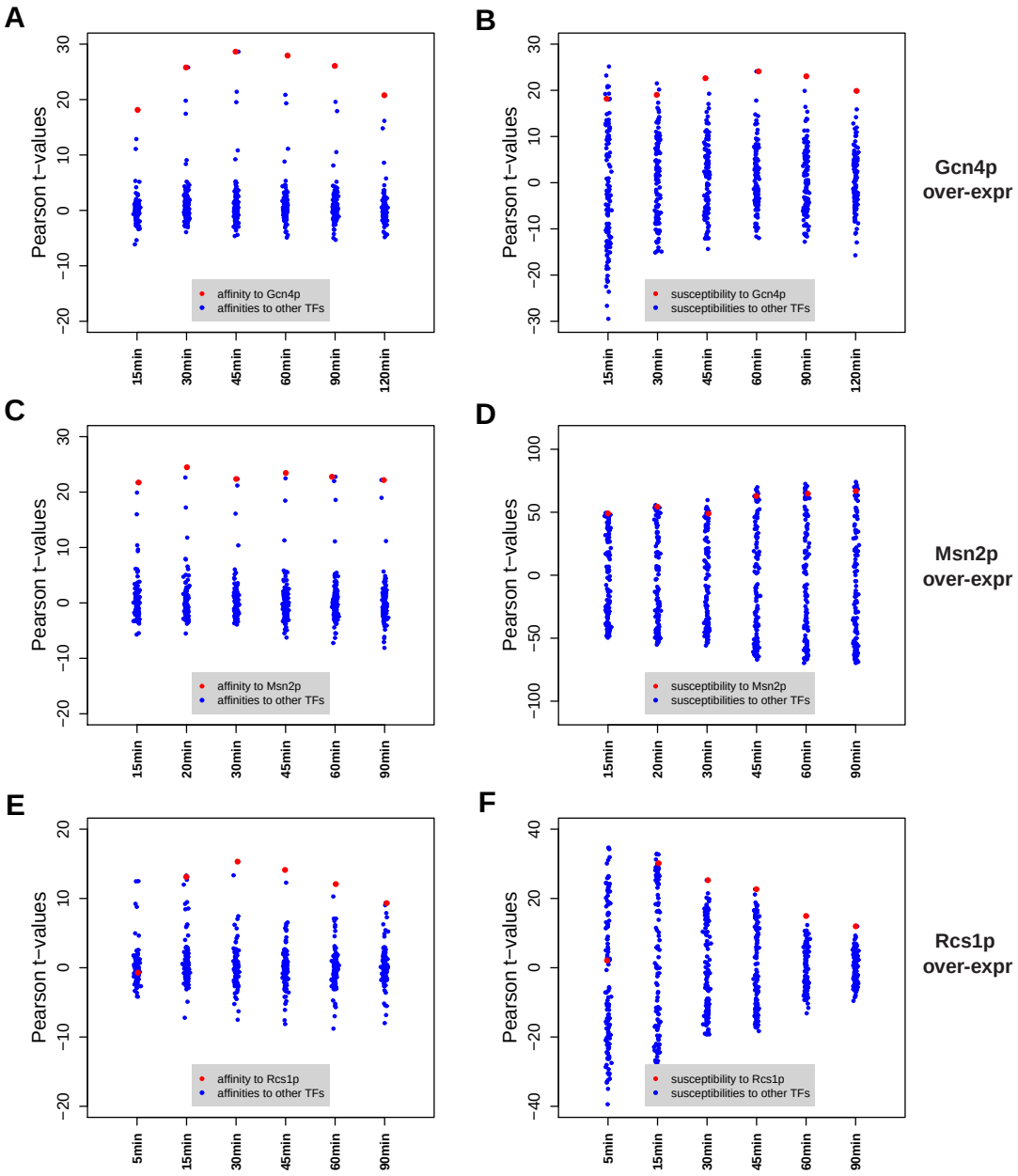

Figure 5.8: Correlation of the Time Series Over-Expression Data to the Affinities and Susceptibilities to 123 TFs. The x-axis represents the over-expression data at different time point. Each point corresponds to the Pearson t-value of the correlation between the affinities/susceptibility to a TF and the mRNA levels in the over-expression experiment. The t-value of the affinity/susceptibilities to Gcn4p, Msn2p and Rcs1p are indicated in red. The affinities to these three TFs are exclusively correlated to the over-expression data. The results in panel (A), (C), and (E) demonstrate that their *in vitro* occupancies correlate significantly to their *in vivo* function. Susceptibilities to Gcn4p and Rcs1p are exclusively correlated to the over-expression data after the first two time points (panel (B) and (F)). The Msn2p over-expression data correlated higher to the susceptibilities to Msn4p, a paralog of Msn2p. At later time points the correlation for Msn2p improves (panel (D)).

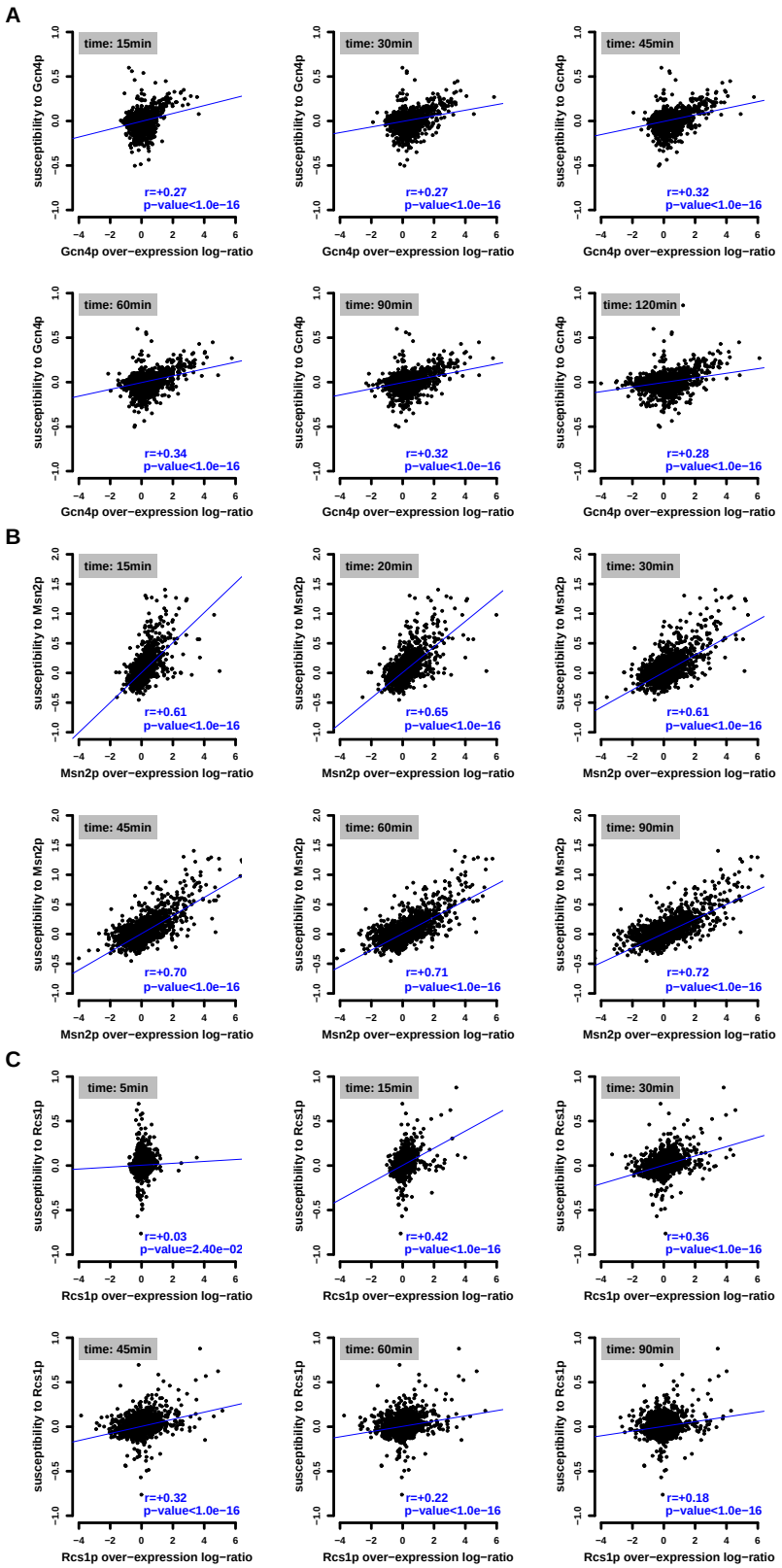

Figure 5.9: Functional Validation of Selected TFs Based on Over-Expression Data. We obtained the genome-wide mRNA levels for controlled over-expression of a group of TFs from McIsaac *et al.* (2013). We calculated how well our inferred

Figure 5.9: (Cont. Caption) susceptibilities to (A) Gcn4p, (B) Msn2p, and (C) Rcs1p are correlated to the over-expression data. We expected a relatively low correlation for early time points and higher correlation for later time points. This plots confirms that the inferred susceptibilities have captured the functional connection to the TFs.

Gene Ontology (GO) Enrichment Analysis

As a second validation, we used the Gene Ontology Consortium (Ashburner *et al.*, 2000) to identify any enrichment in the inferred susceptibilities toward a particular biological function or molecular structure. We expected to detect enrichment toward the GO categories that are related to the TFs function. We considered only the GO categories with at least 10 genes. We applied the Wilcoxon-Mann-Whitney test on the susceptibilities of the selected TFs. We accepted the GO categories with enrichment higher than the significant threshold at 1% FDR level using Benjamini-Hochberg method for multiple testing correction(p-value threshold $\sim 1.0 \times 10^{-3}$).

Cha4p is the transcriptional activator for catabolism[1] of hydroxyamino acids. (Holmberg and Schjerling, 1996). Cha4p regulated transcription of the gene CHA1, which encodes for a protein that allows the yeast to grow on media with L-serine or L-threonine as sole nitrogen source by removing of the amino group NH2 from the amino compounds. We found many significant GO categories among which was the glycoprotein metabolic process category with positive correlation (p-value = 2.02×10^{-13}).

Gcn4p transcriptionally activates genes involved in amino acid biosynthetic in response to amino acid starvation (Hinnebusch and Fink, 1983). Detailed examination of its targets revealed that it also activates genes involved in glycogen homeostasis, mitochondrial carrier proteins, vitamin biosynthesis and autophagy (Natarajan *et al.*, 2001). Also increasing number of studies are connecting Gcn4p to the initial step for nucleosome displacement and recruitment of RNA pol II to the transcription start site (Natarajan *et al.*, 1999). We found total of 43 enriched GO categories among

[1]The metabolic processes that break down molecules into smaller units and release energy.

which was the amino acid biosynthetic process category (p-value = 3.6×10^{-11}).

Ino4p is a transcriptional activator required for genes involved in phospholipid synthesis (Schwank *et al.*, 1995). We did not detect any significant and relevant GO category for this transcription factor.

Leu3p regulates genes involved in branched chain amino acid biosynthesis and ammonia assimilation (Friden and Schimmel, 1988). It can also act as a repressor in high levels of leucine amino acid. Two related GO categories that were significant were cellular biosynthetic process (p-value = 1.9×10^{-11}) and branched chain family amino acid biosynthetic process (p-value = 5.4×10^{-4}).

Msn2p is the transcriptional activator that is active in stress conditions and regulates the general stress response of yeast (Martinez-Pastor *et al.*, 1996; Schmitt and McEntee, 1996). During activation of stress response the expression levels of about 500 genes are effected. The induced genes include those with function in protein folding, protein degradation and energy generation and repressed genes are dominated by functions in translation and protein synthesis, ribosomal proteins, processing of rRNAs and tRNAs and different aspects of cell growth (Causton *et al.*, 2001; Gasch *et al.*, 2000). We measured positive enrichment for categories such as oxidation reduction (p-value = 4.9×10^{-20}) and many relevant negatively enriched categories including ribosome biogenesis (p-value = 7.6×10^{-90}), DNA-directed RNA polymerase III complex (p-value = 1.6×10^{-5}), tRNA modification (p-value = 9.4×10^{-5}).

Rcs1p is a transcription factor that is involved in iron utilization and homeostasis (Yamaguchi-Iwai *et al.*, 1995). Among the enriched categories was transition metal ion transport (p-value = 2.1×10^{-4}).

Ste12p is a transcription factor that is activated by a MAP kinase signaling cascade (Elion *et al.*, 1993; Roberts and Fink, 1994). Upon activation, it induces the genes involved in mating or pseudohyphal/invasive growth pathways (Dolan *et al.*, 1989; Liu *et al.*, 1993). We found two relevant enriched categories site of polarized growth (p-value = 1.7×10^{-10}) and reproductive process (p-value = 1.5×10^{-5}).

The results from GO enrichment analysis for these 7 proteins indicate that the inferred susceptibilities contain relevant functional and biological connection to the TFs.

5.3.4 cQTL Discovery

Our hypothesis is that if a locus influences the strength of connectivity between a TF and its targets, we expect to observe a significant difference between the susceptibilities based on the inherited parental allele at that locus for the target genes. If there is no significant linkage to that locus, then the susceptibility differences are expected to behave like a standard normal random variable. So by calculating a χ^2-statistic corresponding to a split of the segregants based on the genotype at the marker, we can test the significance of that locus for modulation of connectivity of a protein to its target genes. We can then calculate a χ^2-statistic for each of the 2956 marker and obtain a chromosomal profile of the χ^2 values. For each of the 7 selected TFs, we either used all the genes or only the positive targets (see Methods). As explained earlier in Chapter 3, linkage disequilibrium results in relatively low resolution for identifying underlying causal loci. Thus we applied a forward selection algorithm to detect these causal loci. We found cQTL with known interaction for Gcn4p and Ste12p.

5.3.5 Detecting Modulators of Gcn4p-Target Connectivity

Figure 5.10 displays the cQTL analysis results for Gcn4p. We first calculated the χ^2 using all genes and identified significant markers by applying our forward selection step. **Figure 5.10A** shows the χ^2-statistic profile for chromosomal markers. We used 1% level with Bonferroni method for multiple testing correction (threshold p-value = 3.4×10^{-6} corresponding to χ^2-statistic = 4921). We identified a total of 5 markers as significant cQTL (black circle in the plot). The significant regions around each of the selected markers is shown in red color. Again, we only considered proteins that are involved in physical interaction with Gcn4p. Among these proteins, we only displayed those that are the product of genes located in the significant cQTL regions (green solid

circles). We identified a locus on chromosome IV that contains 72 genes including the SRB7 gene. The protein product of this gene, Srb7p, is a subunit of the RNA polymerase II mediator complex (Hengartner *et al.*, 1995). Gcn4p and Srb7p interaction has been experimentally validated (Natarajan *et al.*, 1998; Park *et al.*, 2000). However, we did not detect any coding or non-coding SNPs within the coding region of Srb7p between RM and S288c[2] strains, indicating that the Gcn4p-target connectivity could be affected by variation existing in the upstream regulation of the SRB7 gene expression between the two yeast strains. Also it could be that this cQTL is linked to other genes located within this region whose protein interaction with Gcn4p has not been identified. We also identified a marker on chromosome XVI whose associated region contains 59 genes including the ARP7 gene. This gene encodes a protein that is a component of both the SWI/SNF and RSC chromatin remodeling complexes (Szerlong *et al.*, 2003). SWI/SNF (SWitch/Sucrose NonFermentable) complex is one of the major ATP-dependent chromatin remodeling composed of 12 subunits (Smith *et al.*, 2003). Prior to transcription initiation of a gene, this complex alters the position of nucleoseomes occupying the *cis*-regulatory site of that gene by forming a DNA loop on the nucleosome surface (Zofall *et al.*, 2006), and Arp7 function with DNA bending proteins to enhance proper chromatin architecture (Szerlong *et al.*, 2003). It has been demonstrated that Arp7p and Gcn4p interact and recruit SWI/SNF to its targets promoter region (Natarajan *et al.*, 1999; Neely *et al.*, 1999). Alignment of ARP7 between the two strains revealed two non-coding SNPs. These SNPs could affect the steady-state abundance of Arp7p by post-transcriptional and translational mechanisms.

We also calculated a χ^2-statistic profile using only positive targets of Gcn4p at 5% FDR level(see Methods for the details). These are the target genes that are induced upon increased activity of Gcn4p. The results is presented in **Figure 5.10B**. By applying forward selection, we identified 3 significant markers (threshold p-value = 3.4×10^{-6} corresponding to χ^2-statistic = 332.4). The most significant selected marker

[2]S288c is a laboratory yeast strain that is isogenic to the BY strain with only about 39 SNPs occuring between the their genomes.

is located in a region close to TAF13 gene on chromosome XIII, which indicated with a black arrow in the plot. By aligning the amino acid sequences of Taf13p from the RM and S288c strains, we identified 3 coding SNPs (see **Figure 5.11**). This region encompasses a total of 59 genes. Considering that 55 proteins have been identified that are involved in direct physical interaction with Gcn4p, the p-value for detecting one of these proteins in this region by chance is equal to 9% based on hypergeometric probability distribution. Therefore, it is highly likely that the polymorphisms within the TAF13 gene are responsible for the observed linkage to this region on chromosome XIII.

Taf13p is one of the proteins comprising TFIID which itself is a subunit of the RNA polymerase II holoenzyme. It has been demontrated that Gcn4p and TFIID complex physically interact (Lim *et al.*, 2007). Next section provides more details on the interaction between Taf13p and Gcn4p and its role in RNA pol II transcriptional initiation. Considering the significance of the χ^2-statistic at this locus, it is plausible to think that mutations in the subunits TFIID complex can effect the efficiency of transcription initiation by Gcn4p.

Finally, we applied this step to the Gcn4p negative targets, genes that are repressed by Gcn4p activity (results not shown). However we could not get any significant loci involving known protein interaction with Gcn4p.

5.3.6 Interaction between Taf13p and Gcn4p

TFIID is a transcription factor complex that is required for RNA polymerase II-mediated transcription of protein-coding genes. Recognition of promoter of the DNA by the TFIID complex is required for the recruitment of RNA pol II to the transcription start site. The results of genome-wide studies indicate that TFIID functions primarily at the TATA-less promoters (Burker and Kadonaga, 1996). This is facilitated by the interaction between the transcription activating factor, in our case Gcn4p, and the TFIID complex. It has been demonstrated experimentally that a sub-

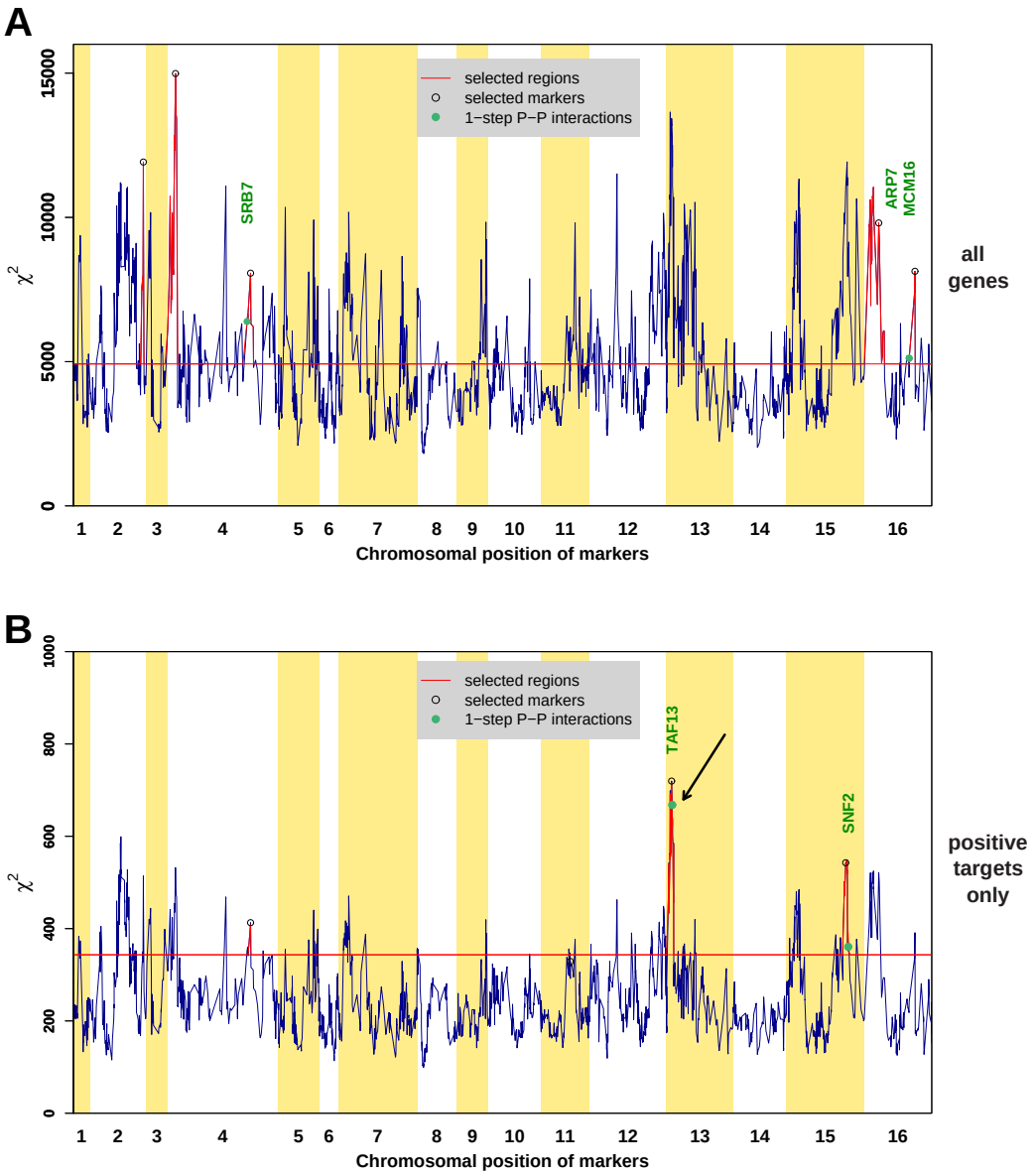

Figure 5.10: Detection of Taf13p as cQTL Modulator for Gcn4p Positive Targets. (A) Using all genes (4482), (B) Using only positive targets at 5% FDR level with Benjamini-Hochberg multiple testing correction (224 genes). We performed forward selection to detect significant peaks at 1% level with Bonferroni correction. Only when we considered positive targets of Gcn4p we were able to detect the marker close to TAF13. By aligning TAF13 protein sequences with BLAST software between RM and S288c strains, we identified 3 coding SNPs for Taf13p between the two strains. In both plots the detected markers by forward selection algorithm are marked with black circles, the significant regions around each selected markers with red color and the direct physical interaction of Gcn4p and genes product located in the significant regions with solid green circles. The horizontal red line represents the χ^2-statistic significant threshold at 1% level with Bonferroni correction.

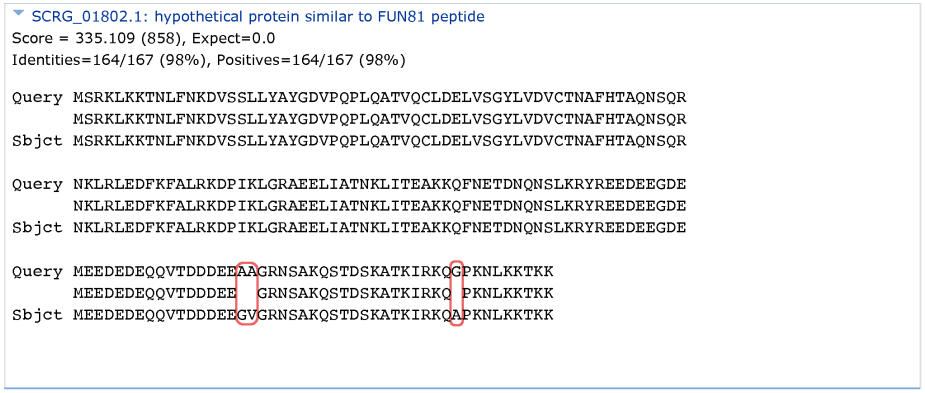

Figure 5.11: Taf13p Sequence Alignment Between the S288c and RM Strains. Here Query is for the S288c strain and Sbjct is for the RM strain. We identified 3 coding polymorphisms between the two sequences. The three polymorphisms are pointed out with red ovals
. Figure generated by BLAST (Altschul *et al.*, 1997).

unit of TFIID, Taf13p, interacts with Gcn4p *in vivo* (Lim *et al.*, 2007). The TFIID acts like the bridge between the TF and the RNA pol II, any mutation TAF proteins could effect the efficiency of the transcription rate of Gcn4p target genes.

5.3.7 Detecting Putative Modulators of Ste12p-Target Connectivity

Figure 5.13 displays the cQTL analysis results for Ste12p. We first calculated the χ^2 using all genes and identified significant markers by applying our forward selection step. **Figure 5.13A** shows the χ^2-statistic profile for the chromosomal markers. We used 1% level with Bonferroni method for multiple testing correction resulting in a threshold p-value $= 3.4 \times 10^{-6}$ corresponding to a χ^2-statistic $= 4921$. We identified a total of 6 markers as significant cQTL (black circle in the plot). The significant regions for each of the selected markers are shown in red. Since we are interested in cofactors whose allelic variation modulates connectivity between the TF and its targets, we only considered proteins that are involved in physical interaction with Ste12p. Among these proteins, we only displayed those that are the product of genes located in the significant cQTL regions (green solid circles). We identified a locus

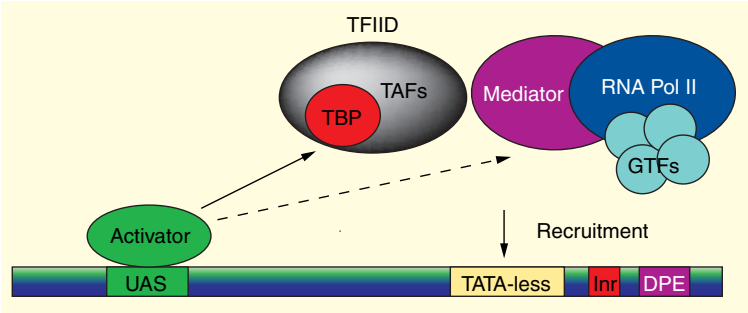

Figure 5.12: Model for RNA Polymerase II-Mediated Transcriptional Activation Involving the TFIID Complex. The RNA transcriptional machinery consists of many subunits: RNA pol II, Mediator and general transcription complexes including TFIID. TFIID subunit itself is comprised of TATA-binding protein (TBP) and about a dozen TBP-associated factors (TAFs). This machinery gets recruited to the transcription start site with the help of transcription activating factor, in our case Gcn4p, that binds to the upstream activation sequence (UAS) and, also the nucleosome displacement complex SWI/SNF. TFIID-dependent pol II recruitment typically is for the promoters lacking TATA element. The TFIID subunit recognizes either the initiator or down stream promoter element (DPE), both of which are present in TATA-less promoters. Figure from Chen and Hampsey (2002).

on chromosome IV that contains 125 genes including the DIG2 gene. This marker is closest to DIG2 on this chromosome (Green solid circle and black circle overlap; shown with black arrow). The protein product of this gene, Dig2p, is a known inhibitor of Ste12p activity (Cook *et al.*, 1996; Pi *et al.*, 1997; Tedford *et al.*, 1997).

We then used BLAST (Altschul *et al.*, 1997) to align the protein sequences of Dig2p in the RM strain and the S288c strain, strain isogenic to the BY strain. We identified a single amino acid polymorphism shown in **Figure 5.14**. At position 83 the isoleucine amino acid (I) in RM is mutated to threonine (T) in S288c (see **Figure 5.14**). Three other genes whose protein product have shown to interact with Ste12p are located near the DIG2 locus: ADA2, SAM2 and SMT3. Among these three, Smt3p has the most closely related function in pheromone response pathway. Smt3p is a small ubiquitin-related modifier protein (SUMO) where it is shown to be responsible for switching from filamentous growth to the mating differentiation program in the presence of pheromone (Wang and Dohlman, 2006). Through alignment of the Smt3p protein sequence between RM and S288c strains, we identified no coding SNPs. Experimental

validation is needed to further investigate the role of either of the DIG2 and SMT3 SNPs as causal cQTL of Ste12p activity at this locus. However, since the χ^2-statistic at DIG2 location is significantly larger than at the SMT3 locus, we expect the former to be the main modulator. We also identified two more loci containing known protein interactions with Ste12p: Kss1p on chromosome VII and Mss11p on chromosome XIII, with zero and about 30 coding SNPs between RM and S288c strains respectively. Both of these protein are mainly involved in invasive filamentous growth in nutrient poor conditions (Cook *et al.*, 1996; Gagiano *et al.*, 2003). The kinase activity of Kss1p, a member of the mitogen[3]-activated protein kinases (MAPKs), induces filamentation. However, in both cases the χ^2-statistic are only marginally significant.

We also calculated χ^2-statistic profile using only 139 positive targets of Ste12p at 5% FDR level(see Methods for the details). These are the target genes that are induced upon increased activity of Ste12p. The results are presented in **Figure 5.13B**. By applying forward selection, we identified 3 significant markers on chromosomes IV, IX and XV (threshold p-value = 3.4×10^{-6} corresponding to χ^2-statistic = 227.2). The selected marker on chromosome IV is again the DIG2 locus.

We used the hypergeometric distribution to estimate the probability that this region contains one of the known protein interactions with Ste12p. There are total of 40 identified proteins to have physical interaction with Ste12p. Considering the total number of yeast genes about 6000 where 22 are located within this region, the probability for observing one of the cofactors such as Dig2p at this locus equals 0.9%. Considering this low probability and the results from χ^2-statistic profile and forward selection, we think that DIG2 locus has the most potential to act as the modulator of Ste12p connectivity among the other selected loci for the used experimental datasets. To validate this hypothesis, we are performing a DIG2 allele replacement experiment between BY and RM[4] (see **Figure 5.15**). The allele swap method is based on the *delito perfetto* approach, explained in details in **Section 1.5.5** (Storici and Resnick, 2006). The idea is to compare the activity levels of Ste12p between the new strains

[3]Mitogen is the chemical that triggering mitosis and cell division.

[4]The experiment is designed by Harmen Bussemaker and Ivor Muroff.

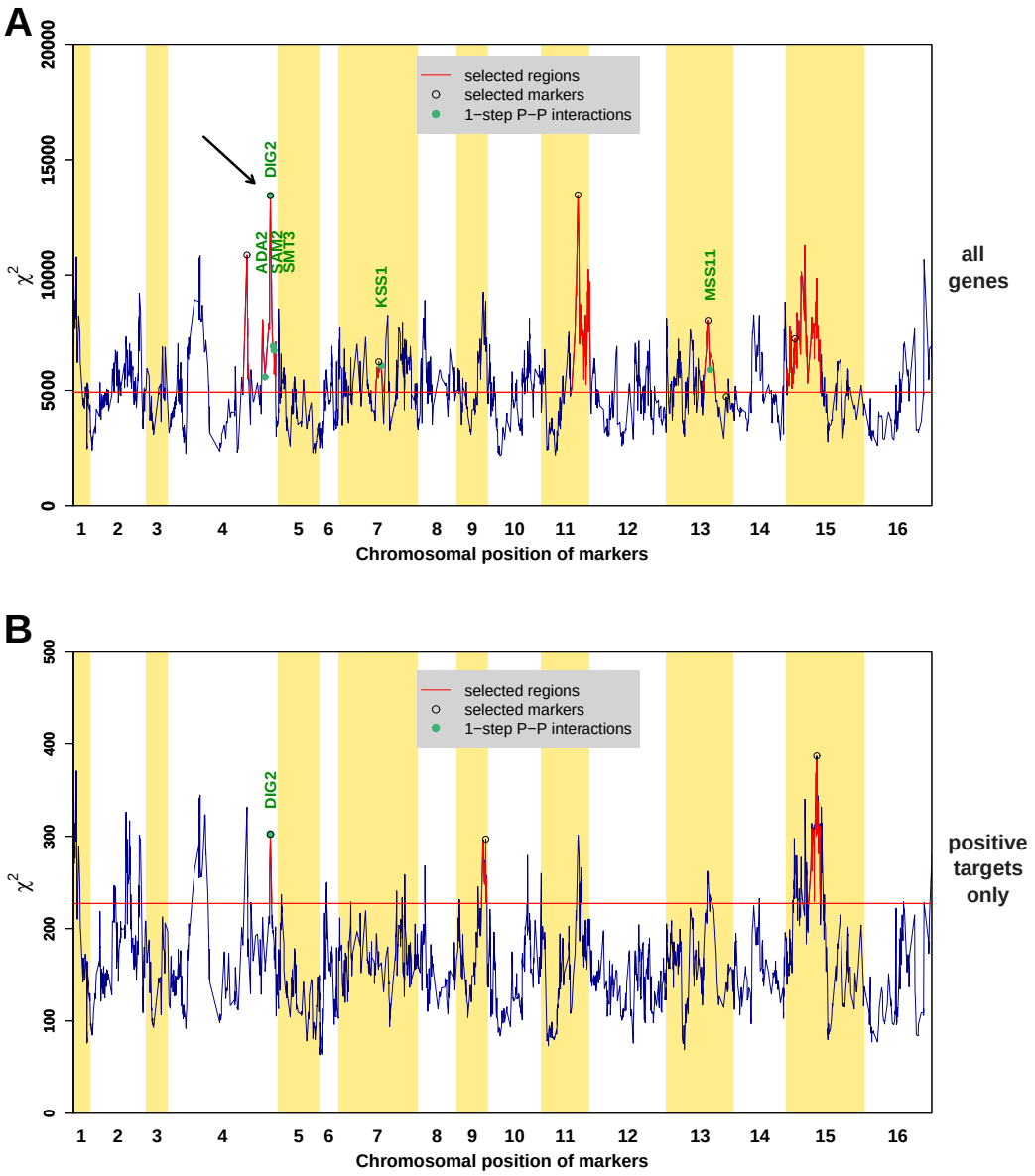

Figure 5.13: Detection of Dig2p as a Putative cQTL Modulator for Ste12p, (A) using all genes (4482), (B) using only positive targets at 5% FDR level with Benjamini-Hochberg multiple testing correction (139 genes). We performed forward selection to detect significant peaks at 1% level with Bonferroni correction. In both cases the marker closest to DIG2 location was selected. Aligning Dig2 protein sequences with BLAST software between the RM and S288c strains, we found a SNP at amino acid number 83 switched from isoleucine in RM to threonine in S288c. In both plots the detected markers by forward selection algorithm are markered with black circles, the significant regions around each selected markers with red color and the direct physical interaction of Ste12p and genes product located in the significant regions with solid green circles. The horizontal red line represents the χ^2-statistic significant threshold at 1% level with Bonferroni correction.

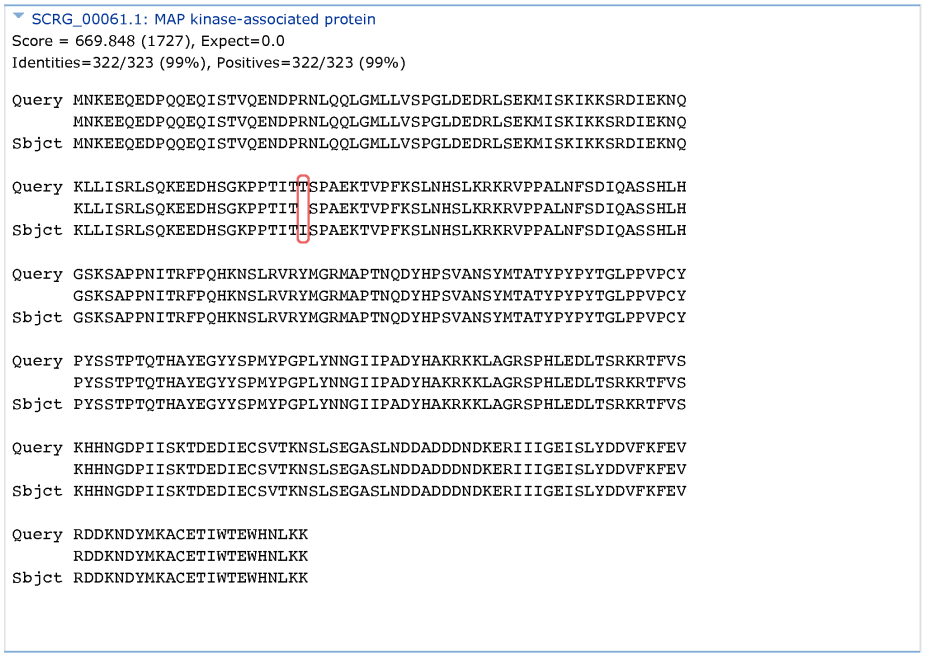

Figure 5.14: Dig2p Sequence Alignment Between S288c and RM Strains. Here Query is for the S288c strain and Sbjct is for the RM strain. We identified a single polymorphism between the two sequences at amino acid 83 where an isoleucine (I) in RM is replaced by threonine (T) in S288c (indicated with red oval). This amino acid is located on moderate phosphorylation site on Dig2p which can effect the binding strength of other cofactors or the phosphorylation complex involving Fus3p. MAPK Fus3p phosphorylates the serine or threonine residues of its target protein immediately followed by proline amino acid. The mutation on position 83 is right upstream of serine-proline, which makes it a likely site for Fus3p-dependent phosphorylation (see **Section 5.3.8**). Figure generated by BLAST (Altschul *et al.*, 1997).

under different levels of pheromone exposures.

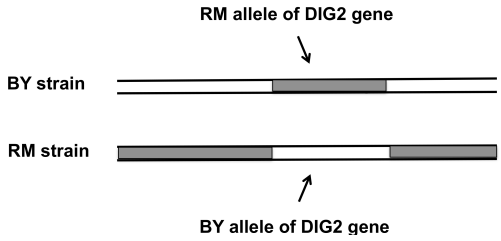

Figure 5.15: DIG2 gene allele replacement between the BY and RM strains.

Finally, we calculated the χ^2-statistic profile using the negative targets of Ste12p, genes that are repressed by Ste12p activity (results not shown). However we could not identify any significant loci involving known protein interaction with Ste12p.

5.3.8 Dig2p, an Inhibitor of Ste12p Activity

In this section we expand more on the *Saccharomyces cerevisiae* (baker's yeast) pheromone response pathway, focusing mainly on the interaction between Ste12p and Dig2p based on the current literature (Bardwell, 2005; Cook *et al.*, 1996; Garrington and Johnson, 1999; Houser *et al.*, 2012; Tedford *et al.*, 1997). The intracellular signal transduction pathway by which the yeast responds to the presence of mating pheromones is known as the yeast mating pheromone response pathway. Yeast have two mating types, **a** and $\boldsymbol{\alpha}$. The mating type identity of a cell is determined based on having the genotype MAT**a** or MAT$\boldsymbol{\alpha}$. MAT**a** cells secrete **a**-factor pheromone and respond to $\boldsymbol{\alpha}$-factor pheromone. Conversely, MAT$\boldsymbol{\alpha}$ cells secrete $\boldsymbol{\alpha}$-factor pheromone and respond to **a**-factor pheromone. As a result of mating and fusion of two haploid cells MAT**a** and MAT$\boldsymbol{\alpha}$, a MAT**a**/MAT$\boldsymbol{\alpha}$ diploid cell is formed. When a yeast cell detects the opposite mating type pheromone in its surroundings, it initiates a series of physiological changes to prepare for mating. These include significant changes in the expression levels of mating related genes, arrest of the cell cycle processes such as DNA and RNA replication for cell growth.

Mating is initiated by the binding of the pheromone to the transmembrane receptor on

the cell surface. The receptor releases a compound which initiates a cascade of protein kinase. We fast forward to the last step of the cascade, the mitogen activated protein kinase (MAPK) conducted by two MAPKs Kss1p and Fus3p. They phosphorylate their target proteins on serine or threonine residues that are immediately upstream of a proline residue on the protein amino acids chain (Payne *et al.*, 1991). The main phosphorylation substrates of Fus3p and Kss1p are Ste12p/Dig2p/Dig1p transcription factor complex. **Figure 5.16** illustrates the interaction network between these proteins. Tec1p/Dig1p and mostly Kss1p are involved in activation and regulation of yeast filamentous growth under poor nutrient condition and so we neglect this part of the network. Ste12p has a DNA binding domain that recognizes the TGAAACA motif on the promoters of its targets validated both experimentally and computationally (Dolan *et al.*, 1989; MacIsaac *et al.*, 2006). This motif is known as the pheromone response element (PRE). However, Dig1p and Dig2p bind and inhibit Ste12p binding to PREs in the absence of pheromone. In strains lacking DIG1 and DIG2, the genes induced by pheromone are upregulated. Fus3p is shown to constantly shuttle between the cytoplasm and nucleus, whereas Dig2p is a nuclear resident protein (van Drogen *et al.*, 2001). Upon pheromone stimulation, the Fus3p level rises about four-fold (Bardwell *et al.*, 1996). Activated Fus3p and Kss1p directly phosphorylates Dig1p and Dig2p. This results in dissociation of Ste12p. The active Ste12p then binds the PRE in the promoter of its targets such as FUS1 gene. However, two independent studies have demonstrated a decrease in pheromone induced FUS1 mRNA in strains lacking the DIG2 gene (Chou *et al.*, 2008; Houser *et al.*, 2012). This means that Dig2p has a positive role in regulation of mating response. Houser *et al.* suggested a model where bound Ste12p-Dig2p complex protects Ste12p from degradation. Hence upon pheromone stimulation, the steady-state level of active Ste12p is high. Ste12p also can bind to its promoter and upregulate its own expression (Dolan and Fields, 1990). Active Fus3p also phosphorylates Ste12p enhancing its degradation where it contributes to the attenuation of the mating response (Esch *et al.*, 2006). As soon as the two mating cells fuse and form a diploid cell, the pheromone response pathway needs to be turned off. In the zygote the interaction between **a**-factor receptor and

$\boldsymbol{\alpha}$-factor receptor is thought to play a role in transition of the zygote to vegetative growth regime (Roth *et al.*, 2000).

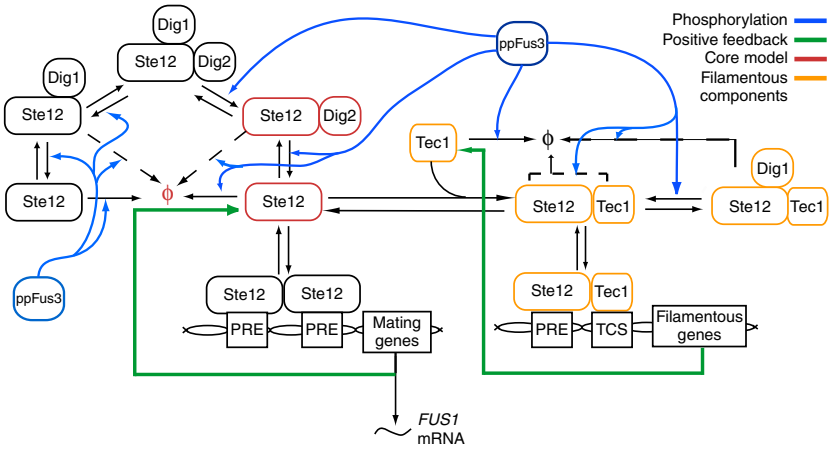

Figure 5.16: A Model for Transcriptional Regulation by the Pheromone Response Pathway Involving Ste12p and Dig2p (indicated in red). Blue arrows represent ppFus3-dependent phosphorylation. Green arrows are for positive fedback loops. For example, increased transcription of FUS1 mRNA, will result in more Ste12p binding to mating genes. The symbol ϕ indicates component degradation. In this model Dig2p binds to Ste12p inhibiting it form binding to mating genes. While Ste12p is bound by Dig2p, it is preserved from degradation. In the presence of pheromones, Fus3p phosphorylates Dig2p resulting in free Ste12p. Ste12p then can bind and initiate mating genes upregulation. Figure from Houser *et al.* (2012).

Considering the critical role of Dig2p in pheromone response activation by Ste12p, we believe the DIG2 locus is a putative modulator of Ste12p-target connection based on our cQTL analysis. Furthermore, the detection of a coding SNP causing a polymorphism in therionine near a high Fus3p-dependant phorphorylation site on Dig2p in S288c strain (highly likely occurring in BY strain as well) strengthen our fining. The validation of this result is possible with DIG2 allele replacement experiment between BY and RM strains.

5.4 Conclusion

In this chapter, we have presented a method for identifying loci that modulate the connectivity of transcription factors (TFs) to their target genes. It takes the expression data, genotype data and TF binding preferences as inputs. Our approach builds on the concept from Chapter 4 of inferring the TF activity levels by exploiting the natural sequence variation existing between the segregants. In our analysis, we used the segregant-specific activity levels of the TFs that are modulated by *trans*-acting loci to infer gene-specific susceptibilities to the variation of these activities. These loci can affect TF activities through different mechanisms such as variation in phosphorylation levels or availability of cofactors. For each gene and chromosomal marker, the difference of the susceptibilities to a particular TF between two subgroups of segregants split based on the inherited parental alleles at that marker, are calculated. The χ^2-statistic obtained by summing the squares of these differences of all genes for each marker is a measure of each locus as a possible cQTL.

We applied our method to a population of 108 yeast segregants generated from a genetic cross between the BY and RM strains (Smith and Kruglyak, 2008). We found a locus on chromosome XIII containing TAF13 whose activity could modulate the connectivity between Gcn4p and its positive targets. The most interesting finding was the detection of the putative role for DIG2, located on chromosome IV, as a modulator of Ste12p connectivity. Validation of both of these finding can be achieved with allele replacement experiments.

Chapter 6

Future Directions

We have presented in thesis several approaches that tackle different aspects of gene expression regulatory program of an eukaryotic cell. We first presented our motif discovery approach that aims to identify stability associated motifs in mRNA sequences for various RNA-binding proteins. We then presented novel linkage approaches to detect genes (i.e. chromosomal loci) that regulate different layers of cellular gene expression regulation by exploiting natural sequence variation existing among related family members of a population. There are many layers involved in gene regulation including signal transduction, chromatin remodeling, transcription initiation, elongation, RNA splicing and post-transcriptional processing, mRNA transport, localization, mRNA stability and translation. The central goal of our first linkage approach is to detect genetic loci that regulate the expression of large number of genes with no reference to any specific regulatory stage. Our second linkage approach focuses on loci that modulate post-transcriptional activities of RNA-binding proteins (RBPs). Our final approach moves further downstream of regulatory path of the interaction between proteins and their targets. We attempted to identify genetic loci that modulate the connectivity of the transcription factors (TFs) and their targets. We will now address some possible future directions and potential applications of our approaches.

First of all, every project presented in this thesis led to some hypotheses that re-

quire further experimental validation. We identified novel motif for three RNA-bing proteins, Scp160p, Sik1p and Tdh3p. We can test these motifs by experiments that engineer the nucleotide composition within the predicted binding site on the target mRNAs or completely deleting the region containing these motifs. By performing a gel shift assay to measure the dissociation constants (K_d) for mutated samples and the RBP under study or performing mRNA expression microarray to study the downstream phenotypic effects of these modification, we can confirm these findings. With our linkage approaches, we found several loci that regulate gene expression levels, RBPs activity levels and connectivity of TFs to their targets that were provided with predictions for putative underlying regulators. For example, we identified that the DIG2 gene is a possible modulator of Ste12p connectivity. We can test the validity of this finding a the DIG2 allele replacement experiment.

Among the improvements that could be made to our motif discovery for RBPs is to integrate the RNA secondary structure in to the model. Our previous attempt was to treat this effect as a multiplicative weighing factor for each nucleotide based on the Boltzmann probability that the nucleotide is unpaired. However, it did not improve the statistical power of our approach. There has been some studies that have successfully contained RNA secondary structure information into their model for few specific factors (Foat and Stormo, 2009; Ray *et al.*, 2013; Riordan *et al.*, 2011).

One improvement to the linkage approaches presented in Chapter 3 and 5, detection of the *trans*-acting loci of gene expression and cQTL, would be to include more detailed information from the linkage disequilibrium. This could increase the resolution of our forward selection procedure.

Our innovative and unique methods based on the χ^2-statistic has greatly increased statistical power for detecting loci that even marginally regulate the expression of large number of genes. This is because our method does not enforce a threshold on the linkage of individual genes, but rather considers the commutative effect on gene expression. One very interesting application would be to apply this approach to human data, specially to cancer data and human disorders caused by uncommon

genotype variants.

Glossary

Allele - specific version of the nucleotide sequence of a gene

Amino Acid - structural unit of a protein.

aQTL - activity quantitative trait loci; chromosomal loci (i.e. genes) that modulate the activity levels of a protein (i.e. trait)

Base Pair - pairing and bonding of A nucleotide to T, and C to G by hydrogen bonding

Caenorhabditis elegans (C. eleganse) - specific worm species (roundworm) as one of the model organisms

Chip-chip - chromatin immunoprecipitation on chip, an experimental method for measuring the protein-DNA interaction

Chromatin - combination of DNA and proteins that make up genetic code of an organism and is stored in the nucleus of a cell

Chromatin Remodeling - mechanisms for decompacting chromatin required for transcription

CIM - Composite interval mapping method which corrects for linkage disequilibrium to some extent by reducing the linkage between neighboring chromosomal markers

***cis*-regulatory Element** - region of DNA that regulates the expression of the nearby genes

Codon - group of three nucleotide that are translated into amino acids during protein synthesis from an mRNA template

cQTL - connectivity quantitative trait loci; chromosomal loci (i.e. genes) that modulate the connectivity (or responsiveness) between a protein and its targets

Deoxyribonucleic Acid (DNA) - nucleotide chains inside the nucleus of a cell containing an organism genetic code

Dissociation Constant (K_d) - equilibrium constant that measures the fraction of unbound components to the bound components for a chemical reaction

eQTL - expression quantitative trait loci; genetic loci (i.e. genes) that regulate the expression of genes (i.e. trait)

FDR - false discovery rate

Gene - regions of DNA sequences that are translated into proteins

Gene Expression - the level of mRNA molecules transcribed from a gene

Gene Expression Regulation - mechanism that control the mRNA level of a gene by regulating the transcriptional rate and/or mRNA decay rate

Gene Ontology (GO) - database containing the annotation of genes of several organisms, which sorts the genes into three main group: cellular component, molecular function and biological process

Genotype - the genetic sequence of a specific individual

in vitro - experiments or measurements that are performed outside a living cell but in a control environment

in vivo - experiments or measurements that are carried out inside a living cell/organism

Kinase - class of proteins that can chemically modulate other proteins

Linkage Disequilibrium - the inheritance of specific parental alleles of different genomic loci more or less than by chance for the offspring/segregants

messenger RNA - mature RNA that is transported to the cytoplasm for protein synthesis by ribosomes

Microarray - chip for measuring gene expression (mRNA levels)

Motif - sequence pattern for a TF or RBP binding site

Nucleosome - structures made of segments of DNA wound around protein cores, being the fundamental units of chromatin

Nucleotide - building blocks of DNA and RNA including A, C, G and T/U

Open Reading Frame (ORF) - region of RNA molecule containing the information for protein synthesis

Phenotype - morphological, biochemical and physiological characteristics of an individual/cell

Post-Transcriptional Processing - processes carried out by RNA-binding protein including: splicing, transportation, localization and stability of RNAs

Promoter - non-coding DNA region for transcription factor binding

PSAM - position-specific affinity matrix containing the binding preferences of a protein

Quantitative Trait Loci (QTL) - regions of DNA containing genes linked a quantitative trait (i.e. characteristic)

Ribonucleic Acid (RNA) - transportable version of genes, used for protein synthesis

RNA-Binding Protein (RBP) - protein binding to RNA transcript and carry out diverse post-transcriptional processing

RNA polymerase II - main component of gene transcription machinery that produces RNA

rRNA - ribosome RNA, type of RNA molecules with role in biogenesis of ribosomes

Ribosome - cell organelles that translate mRNAs to amino acid chains during protein synthesis

Ridge Regression - type of biased regression minimizing the sum of the least squares error function and a penalty term

r-squared - statistical measure representing the percentage of the variance in dependent variable explainable by the independent variables

Saccharomayces cerevisiae. - specific yeast organism (baker's yeast)

Single Nucleotide Polymorphism (SNP) - single nucleotide variation of a DNA sequence compared to a reference sequence

Splicing - process of removal of segments of RNA transcript before protein synthesis

Susceptibility - measure of responsiveness of target genes to the variation in activity level of a protein

Target Genes - the set of genes that their expression is regulated by a protein

Transcription - the process of copying a particular segment of DNA (i.e. gene) into RNA

Transcription Factor (TF) - protein binding to DNA regulating expression of target genes

Translation - the process of translating the genetic information contained in mRNA to the amino acid chain during protein synthesis

UTR - untranslated region are regions of mRNA transcript at either end required for stability regulation

WMW test - Wilcoxon-Mann-Whitney test, a non-parametric test used to test whether two independent samples are drawn from the same population

YPD - medium containing required nutrients for yeast cell growth

Bibliography

Abovich, N. and Rosbash, M., *Cross-Intron Bridging Interactions in the Yeast Commitment Complex are Conserved in Mammals*, Cell, **89** (1997), 403.

Ainger, K., *et al.*, *Transport and Localization Elements in Myelin Basic Protein mRNA*, J. Cell. Biol., **138** (1997), 1077.

Altschul, S. F., *et al.*, *Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs*, Nucleic Acids Research, **25** (1997).

Ambler, R. P. and Rees, M. W., *ε-N-methyl-lysine in Bacterial Flagellar Protein*, Nature, **184** (1959), 56.

Anderson, J. T., Paddy, M. R., and Swanson, M. S., *Pub1 is a Major Nuclear and Cytoplasmic Polyadenylated RNA-Binding Protein in Saccharomyces cerevisiae*, Mol. Cell. Biol., **13** (1993), 6102.

Anfinsen, C. B., *Principles That Govern the Folding of Protein Chains*, Science, **181** (1973), 223.

Arndt, K. T., Styles, C., and Fink, G. R., *Multiple Global Regulators Control HIS4 Transcription in Yeast*, Science, **237** (1987), 874.

Ashburner, M., *et al.*, *Gene Ontology:Toll for the Unification of Biology*, Nature Genetics, **25** (2000), 25.

Aslam, M. L., *et al.*, *Whole genome QTL Mapping for Growth, Meat Quality and Breast Meat Yield Traits in Turkey*, BMC Genet., **12** (2011).

Axelrod, J. D., Reagan, M. S., and Majors, J., *GAL4 Disrupts a Repressing Nucleosome During Activation of GAL1 Transcription in vivo*, Genes. Dev., **7** (1993), 857.

Badner, J. A. *et al.*, *Genome-Wide Linkage Analysis of 972 Bipolar Pedigrees Using Single-Nucleotide Polymorphisms*, Mol. Psychiatry, **17** (2012), 818.

Bardwell, L., *A Walk-Through of the Yeast Mating Pheromone Response Pathway*, Peptides, **26** (2005), 339.

Bardwell, L., *et al.*, *Signaling in the Yeast Pheromone Response Pathway: Specific and High-aAfinity Interaction of the Mitogen-Activated Protein (MAP) Kinases Kss1 and Fus3 with the Upstream MAP Kinase Kinase Ste7.*, Mol. Cell. Biol., **16** (1996), 3637.

Bashirullah, A., Cooperstock, R. L., and Lipshitz, H. D., *Spatial and Temporal Control of RNA Stability*, PNAS, **98** (2001), 7025.

Benayoun, B. A. and Veitia, R. A., *A Post-Translational Modification Code for Transcription Factors: Sorting Through a Sea of Signals*, Trends Cell Biol., **19** (2009), 189.

Benjamini, Y. and Hochberg, Y., *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing*, J. of the Royal Statistical Society, (1995), 289.

Benos, P. V., Bulyk, M. L., and Stormo, G. D., *Additivity in Protein-DNA Interactions: How Good an Approximation is it?*, Nucleic Acids Res., **30** (2002), 4442.

Bentley, D. L., *Rules of Engagement: Cotranscriptional Recruitment of Pre-mRNA Processing Factors*, Cur. Opin. Cell Biol., **17** (2005), 251.

Berg, J. M., Tymoczko, J. L., and Stryer, L., *Biochemistry*, W. H. Freeman, fifth ed., 2002.

Berget, S. M., Moore, C., and Sharp, P., *Spliced Segments at the* $5'$ *Terminus of Adenovirus 2 Late mRNA*, PNAS, **74** (1977), 3171.

Beyer, A. L. and Ocheim, Y. N., *Visualization of RNA Transcription and Processing*, Semin. Cell Biol., **2** (1991), 131.

Black, D. L., *Mechanisms of Alternative Pre-Messenger RNA Splicing*, Annu. Rev. Biochem., **72** (2003), 291.

Blaiseau, P. and Thomas, D., *Multiple Transcriptional Activation Complexes Tether the Yeast Activator Met4 to DNA*, The EMBO J., **17** (1998), 6327.

Boorsma, A., *et al.*, *T-profiler: Scoring the Activity of Predefined Groups of Genes Using Gene Expression Data*, Nucleic Acids Research, **33** (2005), W592.

Boorsma, A., *et al.*, *Inferring Condition-Specific Modulation of Transcription Factor Activity in Yeast through Regulon-Based Analysis of Genomewide Expression*, PLoS ONE, **3** (2008), e3112.

Botstein, D., Chervitz, S. A., and Cherry, J. M., *Yeast as a Model Organism*, Science, **277** (1997), 1259.

Breitkreutz, A., *et al.*, *A Global Protein Kinase and Phosphatase Interaction Network in Yeast*, Science, **328** (2010), 1043.

Brem, R. B., *et al.*, *Genetic Dissection of Transcriptional Regulation in Budding Yeast*, Science, **296** (2002), 752.

Broman, K. W., *et al.*, *R/qtl: QTL Mapping in Experimental Crosses*, Bioinformatics, **19** (2003), 889.

Bugl, H., *et al.*, *RNA Methylation under Heat Shock Control*, Mol. Cell, **6** (2000), 349.

Buratowski, S., *Connections Between mRNA 3′ End Processing and Transcription Termination*, Curr. Opin. Cell Biol., **17** (2005), 257.

Burker, T. W. and Kadonaga, J. T., *Drosophila TFIID Binds to a Conserved Downstream Basal Promoter Element that is Present in Many TATA-Box-Deficient Promoters*, Genes Dev., **10** (1996), 711.

Burnett, G. and Kennedy, E. P., *The Enzymatic Phosphorylation of Proteins*, J. Biol. Chem., **211** (1954), 969.

Bussemaker, H. J., Foat, B. C., and Ward, L. D., *Predictive Modeling of Genome-Wide mRNA Expression: From Modules to Modules*, Annu. Rev. Biophys. Biomol. Struct., **36** (2007), 329.

Carroll, K. L., *et al.*, *Identification of cis Elements Directing Termination of Yeast Nonpolyadenylated snoRNA Transcripts*, Mol. Cell Biol., **24** (2004), 6241.

Causton, H. C., *et al.*, *Remodeling of Yeast Genome Expression in Response to Environmental Changes*, Mol. Biol. Cell, **12** (2001), 323.

Celedon, J. C. *et al.*, *Significant Linkage to Airway Responsiveness on Chromosome 12q24 in Families of Children with Asthma in Costa Rica*, Hum. Genet., **120** (2007), 691.

Chamberlain, J. R., *et al.*, *Purification and Characterization of the Nuclear RNase P Holoenzyme Complex Reveals Extensive Subunit Overlap with RNaseMRP*, Genes Dev., **12** (1998), 1678.

Chelm, B. K. and Geiduschek, E. P., *Gel Electrophoretic Separation of Transcription Complexes: an Assay for RNA Polymerase Selectivity and a Method for Promoter Mapping*, Nucleic Acids Res., **7** (1979), 1851.

Chen, B. S. and Hampsey, M., *Transcription Activation: Unveiling the Essential Nature of TFIID*, Curr. Biol., **12** (2002), R620.

Cheong, C. and Hall, M. T., *Engineering RNA Sequence Specificity of Pumilio Repeats*, PNAS, **103** (2006), 13635.

Chotai, J., *On the LOD Score Method in Linkage Analysis*, Ann. Hum. Genet., **48** (1984), 359.

Chou, S., *et al.*, *Fus3-triggered Tec1 Degradation Modulates Mating Transcriptional Output During the Pheromone Response*, Mol. Sys. Biol., **4** (2008).

Cloonan, N. *et al.*, *Stem Cell Transcriptome Profiling via Massive-Scale mRNA Sequencing*, Na. Methods, **5** (2008), 585.

Conne, B., Stutz, A., and Vassalli, J. D., *The 3′ Untranslated Region of Messenger RNA: A Molecular 'Hotspot' for Pathology?*, Nature Med., **6** (2000), 637.

Cook, J. G., *et al.*, *Two Novel Targets of the MAP Kinase Kss1 are Negative Regulators of Invasive Growth in the Yeast Saccharomyces cerevisiae*, Genes Dev., **10** (1996), 2831.

Cooper, T. A., Wan, L., and Dreyfuss, G., *RNA and Disease*, Cell, **136** (2009), 777.

Darvasi, A. and Soller, M., *Advanced Intercross Lines, an Experimental Population for Fine Genetic Mapping*, Genetics, **141** (1995), 1199.

Delgado, M. L., *et al.*, *The Glyceraldehyde-3-Phosphate Dehydrogenase Polypeptides Encoded by the Saccharomyces cerevisiae TDH1, TDH2 and TDH3 Genes are Also Cell Wall Proteins*, Microbiology, **147** (2001), 411.

Dervan, P. B. and Burli, R. W., *Sequence-Specific DNA Recognition by Polyamides*, Curr. Opin. Chem. Biol., **3** (1999), 688.

Dietzel, C. and Kurjan, J., *The Yeast SCG1 Gene: a G alpha-like Protein Implicated in the a- and alpha-Factor Response Pathway*, Cell, **50** (1987), 1001.

Dolan, J. W. and Fields, S., *Overproduction of the Yeast STE12 Protein Leads to Constitutive Transcriptional Induction*, Genes Dev., **4** (1990), 492.

Dolan, J. W., Kirkman, C., and Fields, S., *The Yeast STE12 Protein Binds to the DNA Sequence Mediating Pheromone Induction*, PNAS, **86** (1989), 5703.

Domon, B. and Aebersold, R., *Mass Spectrometry and Protein Analysis*, Science, **312** (2006), 212.

Drinnenberg, I. A., *et al.*, *RNAi in Budding Yeast*, Science, **326** (2009), 544.

Duncan, R. F., Peterson, H., and Sevanian, A., *Signal Transduction Pathways Leading to Increased eIF4E Phosphorylation Caused by Oxidative Stress*, Free Radic. Biol. Med., **38** (2005), 631.

Eastmond, D. L. and Nelson, H. C., *Genome-Wide Analysis Reveals New Roles for the Activation Domains of the Saccharomayces cerevisiae Heat Shock Transcription Factor (Hsf1) During Transient Heat Shock Response*, J. Biol. Chem., **281** (2006), 32909.

Elion, E. A., Satterberg, B., and Kranz, J. E., *FUS3 Phosphorylates Multiple Components of the Mating Signal Transduction Cascade: Evidence for STE12 and FAR1*, Mol. Biol. Cell, **4** (1993), 495.

Esch, R. K., Wang, Y., and Errede, B., *Pheromone-Induced Degradation of Ste12 Contributes to Signal Attenuation and the Specificity of Developmental Fate*, Eukaryot Cell, **5** (2006), 2147.

Fechter, P. and Brownlee, G. G., *Recognition of mRNA Cap Structures by Viral and Cellular Proteins*, J. Gen. Virol., **86** (2005), 1239.

Fiedler, D. *et al.*, *Functional Organization of the S. cerevisiae Phosphorylation Network*, Cell volume =, ().

Fields, S. and Song, O., *A Novel Genetic System to Detect Protein-Protein Interactions*, Nature, **340** (1989), 245.

Filipowicz, S. N., W. Bhattacharyya and Sonenberg, N., *Mechanisms of Post-Transcriptional Regulation by MicroRNAs: Are the Answers in Sight?*, Nature Rev. Genet., **9** (2008), 102.

Fleischer, T. C., *et al.*, *Systematic Identification and Functional Screens of Uncharacterized Proteins Associated with Eukaryotic Ribosomal Complexes*, Genes Dev., **20** (2006), 1294.

Foat, B. C., Morozov, A. V., and Bussemaker, H. J., *Statistical Mechanical Modeling of Genome-wide Transcription Factor Occupancy Data by MatrixREDUCE*, Bioinformatics, **22** (2006), 7068.

Foat, B. C. and Stormo, G. D., *Discovering Structural cis-regulatory Elements by Modeling the Behaviors of mRNAs*, Mol. Syst. Biol., **5** (2009).

Foat, B. C., *et al.*, *Profiling Condition-Specific, Genome-Wide Regulation of mRNA Stability in Yeast*, PNAS, **102** (2005), 17675.

Fodor, S. P. A., *et al.*, *Light-Directed, Spatially Addressable Parallel Chemical Synthesis*, Science, **251** (1991), 767.

Forsburg, S. L. and Guarente, L., *Identification and Characterization of HAP4: a Third Component of the CCAAT-Bound HAP2/HAP3 Heteromer*, Genes Dev., **3** (1989), 1166.

Friden, P. and Schimmel, P., *LEU3 of Saccharomyces cerevisiae Activates Multiple Genes for Branched-Chain Amino Acid Biosynthesis by Binding to a Common Decanucleotide Core Sequence*, Mol. Cell. Biol., **8** (1988), 2690.

Gagiano, M., *et al.*, *Mss11p is a Transcription Factor Regulating Pseudohyphal Differentiation, Invasive Growth and Starch Metabolism in Saccharomyces cerevisiae in Response to Nutrient Availability*, Mol. Microbiol., **47** (2003), 119.

Gaisne, M., *et al.*, *A "Natural" Mutation in Saccharomyces cerevisiae Strains Derived from S288c Affects the Complex Regulatory Gene HAP1 (CYP1)*, Curr. Genet., **36** (1999), 195.

Gao, F., Foat, B. C., and Bussemaker, H. J., *Defining Transcriptional Networks Through Integrative Modeling of mRNA Expression and Transcription Factor Binding Data*, BMC Bioinformatics, **5** (2004).

Gao, X. L., Mirau, P., and Patel, D. J., *Structure Refinement of the Chromomycin Dimer-DNA Oligomer Complex in Solution*, J. Mol. Biol., **223** (1992), 259.

Garner, M. M. and Revzin, A., *A Gel Electrophoresis Method for Quantifying the Binding of Proteins to Specific DNA Regions: Application to Components of the Escherichia coli Lactose Operon Regulatory System*, Nucleic Acids Res., **9** (1981), 3047.

Garrey, S. M., Voelker, R., and Berglund, J. A., *An Extended RNA Binding Site for the Yeast Branch Point-Binding Protein and the Role of its Zinc Knuckle Domains in RNA Binding*, J. Biol. Chem., **281** (2006), 27433.

Garrington, T. P. and Johnson, G. L., *Organization and Regulation of Mitogen-Activated Protein Kinase Signaling Pathways*, Curr. Opin. Cell Biol., **11** (1999), 211.

Gasch, A. P., *et al.*, *Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes*, Mol. Biol. Cell, **11** (2000), 4241.

Gavin, A. C. *et al.*, *Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes*, Nature, **415** (2002), 141.

Gerber, A. P., Herschlag, D., and Brown, P. O., *Extensive Association of Functionally and Cytotopically Related mRNAs with Puf Family RNA-Binding Proteins in Yeast*, PLoS Biol., **2** (2004), e79.

Ghaemmaghami, S., *et al.*, *Global Analysis of Protein Expression in Yeast*, Nature, **425** (2003), 737.

Gill, G. and Ptashne, M., *Mutants of GAL4 protein Altered in an Activation Function*, Cell, **51** (1987), 121.

Gimeno, C. J., *et al.*, *Unipolar Cell Divisions in the Yeast S. cerevisiae lead to Filamentous Growth: Regulation by Starvation and RAS*, Cell, **68** (1992), 1077.

Giniger, E., Varnum, S. M., and Ptashne, M., *Specific DNA Binding of GAL4, a Positive Regulatory Protein of Yeast*, Cell, **40** (1985), 767.

Glanzer, J., *et al.*, *RNA Splicing Capability of Live Neuronal Dendrites*, PNAS, **102** (2005), 16859.

Goffeau, A. *et al.*, *Life with 6000 Genes*, Science, **274** (1996), 546.

Goldstein, A. L. and McCusker, J. H., *Three New Dominant Drug Resistance Cassettes for Gene Disruption in Saccharomayces cerevisiae*, Yeast, **15** (1999), 1541.

Gregory, T. R., *Synergy Between Sequence and Size in Large-Scale Genomics*, Nat. Rev. Genet., **6** (2005), 699.

Griggs, D. W. and Johnston, M., *Regulated Expression of the GAL4 Activator Gene in Yeast Provides a Sensitive Genetic Switch for Glucose Repression*, PNAS, **88** (1991), 8597.

Grigull, J., *et al.*, *Genomewide Analysis of mRNA Stability Using Transcription Inhibitors and Microarrays Reveals Posttranscriptional Control of Ribosome Biogenesis Factors*, Mol. Cell. Biol., **24** (2004), 5534.

GSA, *Genetics Society of America*, , `http://www.genetics.org/site/misc/images/CCThumbnails/001818.tiff`.

Guo, B., *et al.*, *A Saccharomyces Gene Family Involved in Invasive Growth, Cell-Cell Adhesion, and Mating*, PNAS, **97** (2000), 12158.

Guo, M., *et al.*, *The Yeast G Protein Alpha Subunit Gpa1 Transmits a Signal Through an RNA Binding Effector Protein Scp160*, Mol. Cell, **12** (2003), 517.

Guo, Y., *et al.*, *A Genome-Wide Linkage and Association Scan Reveals Novel Loci for Hypertension and Blood Pressure Traits*, PLoS One, **7** (2012), e31489.

Gygi, S. P., *et al.*, *Correlation between protein and mRNA Abundance in Yeast*, Mol. and Cell. Biol., **19** (1999), 1720.

Hahn, S. and Young, E. T., *Transcriptional Regulation in Saccharomayces cerevisiae: Transcription Factor Regulation and Function, Mechanisms of Initiation, and Roles of Activators and Coactivators*, Genetics, **189** (2011), 705.

Halme, A., *et al.*, *Genetic and Epigenetic Regulation of the FLO Gene Family Generates Cell-Surface Variation in Yeast*, Cell, **116** (2004), 405.

Han, J., *et al.*, *Pre-mRNA Splicing: Where and When in the Nucleus*, Trends in Cell Biol., **21** (2011), 336.

Harbison, C. T., *et al.*, *Transcriptional Regulatory Code of a Eukaryotic Genome*, Nature, **431** (2004), 99.

Hartl, F. U., Bracher, A., and Hayer-Hartl, M., *Molecular Chaperones in Protein Folding and Protestasis*, Nature, **475** (2011), 324.

Hasegawa, Y., Irie, K., and Gerber, A. P., *Distinct Roles for Khd1p in the Localization and Expression of Bud-Localized mRNAs in Yeast*, RNA, **14** (2008), 2333.

Heaton, D., *et al.*, *Mutational Analysis of the Mitochondrial Copper Metallochaperone Cox17*, J. Biol. Chem., **275** (2000), 37582.

Hector, R. E., *et al.*, *Dual Requirement for Yeast hnRNP Nab2p in mRNA Poly(A) Tail Length Control and Nuclear Export*, EMBO J., **21** (2002), 1800.

Heidari, B., *et al.*, *Mapping QTL for Grain Yield, Yield Components, and Spike Features in a Doubled Haploid Population of Bread Wheat*, Genome, **54** (2011), 517.

Heitz, E., *Das Heterochromatin der Moose*, I Jahrb Wiss Botanik, **69** (1928), 762.

Hengartner, C. J., *et al.*, *Association of an Activator with an RNA Polymerase II Holoenzyme*, Genes Dev., **9** (1995), 897.

Hentze, M. W. and Kulozik, A. E., *A Perfect Message: RNA Surveillance and Nonsense-Mediated Decay*, Cell, **96** (1999), 307.

Hentze, M. W., *et al.*, *Identification of the Iron-Responsive Element for the Translational Regulation of Human Ferritin mRNA*, Science, **238** (1987), 1570.

Hinnebusch, A. G. and Fink, G. R., *Positive Regulation in the General Amino Acid Control of Saccharomyces cerevisiae*, PNAS, **80** (1983), 5374.

Hoerl, A. E. and Kennard, R. W., *Ridge Regression:Biased Estimation for Nonorthogonal Problems*, Technometrics, **12** (1970), 55.

Hogan, D. J., *et al.*, *Diverse RNA-bing Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System*, PLoS Biol., **6** (2008), e255.

Holmberg, S. and Schjerling, P., *Cha4p of Saccharomyces cerevisiae Activates Transcription via Serine/Threonine Response Elements*, Genetics, **144** (1996), 467.

Hook, B. A., *et al.*, *Two Yeast PUF Proteins Negatively Regulate a Single mRNA*, J. of Biol. Chem., **282** (2007), 15430.

Houseley, J. and Tollervey, D., *The Many Pathways of RNA Degradation*, Cell, **136** (2009), 763.

Houser, J. R., *et al.*, *Positive Roles for Negative Regulators in the Mating Response of Yeast*, Molecular Systems Biology, **8** (2012).

Houser-Scott, F., *et al.*, *Interactions Among the Protein and RNA Subunits of Saccharomyces cerevisiae Nuclear RNase P*, PNAS, **99** (2002), 2684.

Houshmandi, S. S. and Olivas, W. M., *Yeast Puf3 Mutants Reveal the Complexity of Puf-RNA Binding and Identify a loop Required for Regulation of mRNA Decay*, RNA, **11** (2005), 1655.

Iyer, V. R., *et al.*, *Genomic Binding Sites of the Yeast Cell-Cycle Transcription Factors SBF and MBF*, Nature, **409** (2001), 533.

Jackson, J. S. J., *et al.*, *Recruitment of the Puf3 Protein to its mRNA Target for Regulation of mRNA Decay in Yeast*, RNA, **10** (2004), 1625.

Jansen, R. C., *Studying Complex Biological Systems Using Multifactorial Perturbation*, Nature Reviews Genetics, **4** (2003), 145.

Jansen, R. C. and Nap, J. P., *Genetical Genomics: The Added Value from Segregation*, TRENDS in Genetics, **17** (2001), 388.

Jiang, F., *et al.*, *Gene Activation by Dissociation of an Inhibitor from a Transcriptional Activation Domain*, Mol. Cell. Biol., **29** (2009), 5604.

Jiang, H., Guan, W., and Gu, Z., *Tinkering Evolution of Post-Transcriptional Regulons: Puf3p in Gungi as an Example*, PLoS, **6** (2010).

Johnston, S. A., Salmeron Jr, J. M., and Dincher, S. S., *Interaction of Positive and Negative Regulatory Proteins in the Galactose Regulon of Yeast*, Cell, **50** (1987), 143.

Joshi, A., Beck, Y., and Michoel, T., *Post-Transcriptional Regulatory Networks Play a Key Role in Noise Reduction that is Conserved from Micro-Organisms to Mammals*, FEBS J., **279** (2012), 3501.

Juneau, K., *et al.*, *High-Density Yeast-Tilling Array Reveals Previously Undiscovered Introns and Extensive Regulation of Meiotic Splicing*, PNAS, **104** (2006), 1522.

Kadosh, D. and Struhl, K., *Repression by Ume6 Involves Recruitment of a Complex Containing Sin3 Corepressor and Rpd3 Histone Deacetylase to Target Promoters*, Cell, **89** (1998), 365.

Kasten, M. M. and Stillman, D. J., *Identification of the Saccharomyces cerevisiae Genes STB1-STB5 Encoding Sin3p Binding Proteins*, Mol. Gen. Genet., **256** (1997), 376.

Kataoka, N., *et al.*, *Pre-mRNA Splicing Imprints mRNA in the Nucleus with a Novel RNA-Binding Protein that Persists in the Cytoplasm*, Moll. Cell, **6** (2000), 673.

Keegan, L., Gill, G., and Ptashne, M., *Separation of DNA Binding From the Transcription-Activating Function of a Eukaryotic Regulatory Protein*, Science, **231** (1986), 699.

Kellis, M., *et al.*, *Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements*, Nature, **423** (2003), 241.

Kelly, S. M., *et al.*, *Recognition of Polyadenosine RNA by the Zinc Finger Domain of Nuclear Poly(A) RNA-Binding Protein 2 (Nab2) is Required for Correct mRNA 3'-end Formation*, J. Biol. Chem., **285** (2010), 26022.

Keng, T., *HAP1 and ROX1 form a Regulatory Pathway in the Repression of HEM13 Transcription in Saccharomyces cerevisiae*, Mol. Cell. Biol., **12** (1992), 2616.

Klein, A. P., *et al.*, *Linkage Analysis of Chromosome 4 in Families with Familial Pancreatic Cancer*, Cancer Biol. Ther., **6** (2007), 320.

Kohlhaw, G. B., *Beta-Isopropylmalate Dehydrogenase From Yeast*, Methods Enzymol., **166** (1988), 429.

Kornberg, R. D. and Thomas, J. O., *Chromatin Structure; Oligomers of the Histones*, Science, **184** (1974), 865.

Kouzarides, T., *Chromatin Modifications and Their Function*, Cell, **128** (2007), 693.

Kressler, D., *et al.*, *Spb1p is a Putative Methyltransferase Required for 60S Ribosomal Subunit Biogenesis in Saccharomyces cerevisiae*, Nucleic Acids Res., **27** (1999), 4598.

Kruglyak, L. and Lander, E. S., *A Nonparametric Approach for Mapping Quantitative Trait Loci*, Genetics, **139** (1995), 1421.

Kuchin, S., Vyas, V. K., and Carlson, M., *Snf1 Protein Kinase and the Repressors Nrg1 and Nrg2 Regulate FLO11, Haploid Invasive Growth, and Diploid Pseudohyphal Differentiation*, Mol. Cell. Biol., **22** (2002), 3994.

Kurdistani, S. K. and Grunstein, M., *Histone Acetylation and Deacetylation in Yeast*, Nat. Mol. Cell. Biol., **4** (2003), 276.

Lacroute, F., *Regulation of Pyrimidine Biosynthesis in Saccharomyces cerevisiae*, J. Bacteriol., **95** (1968), 824.

Lambrechts, M. G., *et al.*, *Muc1, a Mucin-Like Protein that is Regulated by Mss10, is Critical for Pseudohyphal Differentiation in Yeast*, PNAS, **93** (1996), 8419.

Lander, E. S. and Botstein, D., *Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps*, Genetics, **121** (1989), 185.

Lang, B. D., *et al.*, *The Brefeldin A Resistance Protein Bfr1p is a Component of Polyribosome-Associated mRNP Complexes in Yeast*, Nucleic Acids Res., **29** (2001), 2567.

Lee, D. Y., *et al.*, *Role of Protein Methylation in Regulation of Transcription*, Endocr. Rev., **26** (2005), 147.

Lee, E. and Bussemaker, H. J., *Identifying the Genetic Determinants of Transcription Factor Activity*, Mol. Sys. Biol., **6** (2010), 412.

Lee, S. I., *et al.*, *Learning a Prior on Regulatory Potential from eQTL Data*, PLoS Genet., **5** (2009), e1000358.

Leipuviene, R. and Theil, E. C., *The Family of Iron Responsive RNA Structures Regulated by Changes in Cellular Iron and Oxygen*, Cell. Mol. Life Sci., **64** (2007), 2945.

Li, H., *et al.*, *Crystal Structure of the Two N-terminal RRM Domains of Pub1 and the Poly(U)-Binding Properties of Pub1*, J. Strucut. Biol., **171** (2010), 291.

Lim, M. K., *et al.*, *Gal11p Dosage-Compensates Transcriptional Activator Deletions via Taf14p*, J. Mol. Biol., **374** (2007), 9.

Liu, H., Styles, C. A., and Fink, G. R., *Elements of the Yeast Pheromone Response Pathway Required for Filamentous Growth of Diploids*, Science, **262** (1993), 1741.

Lodish, H., *et al.*, *Molecular Cell Biology*, W. H. Freeman, 6th ed., 2007.

Lukong, K. E., *et al.*, *RNA-Binding Proteins in Human Genetic Disease*, Trends in Genetics, **24** (2008), 416.

Lunde, B. M., Horner, M., and Meinhart, A., *Structural Insights into cis Element Recognition of Non-Polyadenylated RNAs by the Nab3-RRM*, Nucleic Acids Res., **39** (2011), 337.

Ma, J. and Ptashne, M., *The Carboxy-Terminal 30 Amino Acids of GAL4 are Recognized by GAL80*, Cell, **50** (1987), 137.

MacIsaac, K. D., *et al.*, *An Improved Map of Conserved Regulatory Sites for Saccharomyces cerevisiae*, BMC Bioinformatics, **7** (2006), 113.

Madsen, K., Nielsen, H. B., and Tingless, O., *Methods for Non-Linear Least Squares Problems*, 2004.

Maitra, U., Stringer, E. A., and Chaudhuri, A., *Initiation Factors in Protein Biosynthesis*, Annu. Rev. Biochem., **51** (1982), 869.

Mamoon, N. M., Song, Y., and Wellman, S. E., *Histone h1(0) and Its Carboxyl-Terminal Domain Bind in the Major Groove of DNA*, Biochemistry, **41** (2002), 9222.

Mangus, D. A., Evans, M. C., and Jacobson, A., *Poly(A)-Binding Proteins: Multifunctional Scaffolds for the Post-Transcriptional Control of Gene Expression*, Genome Biol., **4** (2003).

Mardis, E. R., *Next-Generation DNA Sequencing Methods*, Annu. Rev. Genomics Hum. Genet., **9** (2008), 387.

Marsh, L., Neiman, A. M., and Herskowitz, I., *Signal Transduction During Pheromone Response in Yeast*, Annu. Rev. Cell Biol., **7** (1991), 699.

Martinez-Pastor, M. T., *et al.*, *The Saccharomyces cerevisiae Zinc Finger Proteins Msn2p and Msn4p are Required for Transcriptional Induction Through the Stress Response Element (STRE)*, EMBO J., **15** (1996), 2227.

Matunis, M. J., Matunis, E. L., and Dreyfuss, G., *PUB1: a Major Yeast Poly(A)+ RNA-Binding Protein*, Mol. Cell Biol., **13** (1993), 6114.

McAlister, L. and Holland, M. J., *Isolation and Characterization of Yeast Strains Carrying Mutations in the Glyceraldehyde-3-Phosphate Dehydrogenase Genes*, J. Biol. Chem., **260** (1985), 15013.

McIsaac, R. S., *et al.*, *Fast-Acting and Nearly Gratuitous Induction of Gene Expression and Protein Depletion in Saccharomayces cerevisiae*, Mol. Biol. of the Cell, **22** (2011), 4447.

McIsaac, R. S., *et al.*, *Synthetic Gene Expression Perturbation Systems with Rapid, Tunable, Single-Gene Specificity in Yeast*, Nucleic Acids Res., **41** (2012).

McIsaac, R. S. *et al.*, 2013, to be published.

Meisinger, C., *et al.*, *The Preprotein Translocase of the Outer Mitochondrial Membrane: Receptors and a General Import Pore*, Cell Mol. Life Sci., **56** (1999), 817.

Melhem, N. and Devlin, B., *Shedding New Light on Genetic Dark Matter*, Genome Medicine, **2** (2010).

Mignone, F., *et al.*, *Untranslated REgions of mRNAs*, Genome Biology, **3** (2002).

Mili, S. and Steitz, J. A., *Evidence for Reassociation of RNA-Binding Proteins After Cell Lysis: Implications for the Interpretation of Immunoprecipitation Analyses*, RNA, **10** (2004), 1692.

Miller, M. T., Higgin, J. J., and Hall, T. M., *Basis of Altered RNA-Binding Specificity by PUF Proteins Revealed by Crystal Structures of Yeast Puf4p*, Nat. Struct. Mol. Biol., **15** (2008), 397.

Minehart, P. L. and Magasanik, B., *Sequence and Expression of GLN3, a Positive Nitrogen Regulatory Gene of Saccharomyces cerevisiae Encoding a Protein with a Putative Zinc Finger DNA-Binding Domain*, Mol. Cell. Biol., **11** (1991), 6216.

Model, K., *et al.*, *Multistep Assembly of the Protein Import Channel of the Mitochondrial Outer Membrane*, Nat. struct. Biol., **8** (2001), 361.

Moore, M. J., *From Birth to Death: The Complex Lives of Eukaryotic mRNAs*, Science, **309** (2005), 1514.

Mortazavi, A., *et al.*, *Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq*, Nature Methods, **5** (2008), 621.

Muhlrad, D. and Parker, R., *Premature Translational Termination Triggers mRNA Decapping*, Nature, **370** (1994), 578.

Nagalakshmi, U., Waern, K., and Snyder, M., *RNA-Seq: A Method for Comprehensive Transcriptome Analysis*, Curr. Protoc. Mol. Biol., (2010), 14.11.1.

Nagalakshmi, U., *et al.*, *The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing*, Science, **320** (2008), 1344.

Natarajan, K., *et al.*, *yTAFII61 has a General Role in RNA Polymerase II Transcription and is Required by Gcn4p to Recruit the SAGA Coactivator Complex*, Mol. Cell, **2** (1998), 683.

Natarajan, K., *et al.*, *Transcriptional Activation by Gcn4p Involves Independent Interactions with the SWI/SNF Complex and the SRB/Mediator*, Moll. Cell, **4** (1999), 657.

Natarajan, K., *et al.*, *Transcriptional Profiling Shows That Gcn4p is a Master Regulator of Gene Expression During Amino Acid Starvation in Yeast*, Mol. Cell. Biol., **21** (2001), 4347.

Neely, K. E., *et al.*, *Activation Domain-Mediated Targeting of the SWI/SNF Complex to Promoters Stimulates Transcription from Nucleosome Arrays*, Mol. Cell, **4** (1999), 649.

Nehlin, J. O., Carlberg, M., and Ronne, H., *Control of Yeast GAL Genes by MIG1 Repressor: a Transcriptional Cascade in the Glucose Response*, EMBO J, **10** (1991), 3373.

Ness, F., *et al.*, *SUT1 is a Putative Zn[II]2Cys6-Transcription Factor Whose Upregulation Enhances Both sterol Uptake and Synthesis in Aerobically Growing Saccharomyces cerevisiae Cells*, Eur. J. Biochem., **268** (2001), 1585.

Nishikawa, S. and Ono, S., *Transmission of X-rays Through Fibrous, Lamellar and Granular Substances*, Proc. Tokyo. Math. Phys. Soc., **7** (1913), 131.

Okoniewski, M. J. and Miller, C. J., *Hybridization Interactions Between Probesets in Short Oligo Microarrays Lead to Spurious Correlations*, BMC Bioinformatics, **7** (2006), 276.

Olivas, W. and Parker, R., *The Puf3 Protein is a Transcript-Specific Regulator of mRNA Degradation in Yeast*, EMBO J., **19** (2000), 6602.

Orphanides, G. and Reinberg, D., *A Unified Theory of Gene Expression*, Cell, **108** (2002), 439.

Otero, J. M., *et al.*, *Whole Genome Sequencing of Saccharomyces cerevisiae: From Genotype to Phenotype for Improved Metabolic Engineering Applications*, BMC Genomics, **11** (2010), 723.

Pandey, A. and Mann, M., *Proteomics to Study Genes and Genomes*, Nature, **405** (2000), 837.

Park, J. M., *et al.*, *In vivo Requirement of Activator-Specific Binding Targets of Mediator*, Mol. Cell. Biol., **20** (2000), 8709.

Parker, R. and Song, H., *The Enzymes and Control of Eukaryotic mRNA Turnover*, PNAS, **11** (2004), 121.

Payne, D. M., *et al.*, *Identification of the Regulatory Phosphorylation Sites in pp42/Mitogen-Activated Protein Kinase (MAP Kinase)*, EMBO J., **10** (1991), 885.

Pi, H., Chien, C. T., and Fields, S., *Transcriptional Activation Upon Pheromone Stimulation Mediated by a Small Domain of Saccharomyces cerevisiae Ste12p*, Mol. Cell. Biol., **17** (1997), 6410.

Polymenidou, M., *et al.*, *Misregulated RNA Processing in Amyotrophic Lateral Sclerosis*, Brain Research, **26** (2012), 3.

Ptashne, M., *How Eukaryotic Transcriptional Activators Work*, Nature, **335** (1988), 683.

Rao, A. R. and Pellegrini, M., *Regulation of the Yeast Metabolic Cycle by Transcription Factors with Periodic Activities*, BMC sys. Biol., **5** (2011).

Ray, D., *et al.*, *Rapid and Systematic Analysis of the RNA Recognition Specificities of RNA-Binding Proteins*, Nat Biotechnol., **7** (2013), 667.

Reeck, G. R. *et al.*, *"Homology" in Proteins and Nucleic Acids: A Terminology Muddle and a Way Out of it*, Cell, **50** (1987).

Rigaut, G., *et al.*, *A Genetic Protein Purification Method for Protein Complex Characterization and Proteome Exploration*, Nat. Biotech., **17** (1999), 1030.

Riordan, D. P., Herschlag, D., and Brown, P. O., *Identification of RNA Recognition Elements in the Saccharomyces cerevisiae Transcriptome*, Nucleic Acids Res., **39** (2011), 1501.

Roberts, R. L. and Fink, G. R., *Elements of a Single MAP Kinase Cascade in Saccharomyces cerevisiae Mediate Two Developmental Programs in the Same Cell Type: Mating and Invasive Growth*, Genes Dev., **8** (1994), 2974.

Rockman, M. V. and Kruglyak, L., *Genetics of Global Gene Expression*, Nature Reviews Genetics, **7** (2006), 862.

Rockman, M. V. and Kruglyak, L., *Breeding Designs for Recombinant Inbred Advanced Intercross Lines*, Genetics, **179** (2008), 1069.

Rockman, M. V., Skrovanek, S. S., and Kruglyak, L., *Selection at Linked Sites Shapes Heritable Phenotypic Variation in C.elegans*, Science, **330** (2010), 372.

Rosenbaum, D. M., Rasmussen, S. G., and Kobilka, B. K., *The Structure and Function of G-Protein-Coupled Receptors*, Nature, **459** (2009), 356.

Roth, A. F., *et al.*, *Asg7p-Ste3p Inhibition of Pheromone Signaling: Regulation of the Zygotic Transition to Vegetative Growth*, Mol. Cell. Biol., **20** (2000), 8815.

Ryder, S. P., *Pumilio RNA Recognition: The Consequence of Promiscuity*, Structure, **19** (2011), 277.

Saint-Georges, Y., *et al.*, *Yeast Mitochondrial Biogenesis: a Role for the PUF RNA-Binding Protein Puf3p in mRNA Localization*, PLoS One, **3** (2008), e2293.

Sanger, F., Nicklen, S., and Coulson, A. R., *DNA Sequencing with Chain-Terminating Inhibitors*, PNAS, **74** (1977), 5463.

Sapolsky, R. J., *et al.*, *High-Throughput Polymorphism Screening and Genotyping with High-Density Oligonucleotide Arrays*, Genet. Anal., ().

Schacherer, J., *et al.*, *Genome-Wide Analysis of Nucleotide-Level Variation in Commonly Used Saccharomyces cerevisiae Strains*, PLoS One, **2** (2007), e322.

Schellenberg, G. D. *et al.*, *Linkage Analysis of Familial Alzheimer Disease, Using Chromosome 21 Markers*, Am. J. Hum. Genet., **48** (1991), 563.

Schena, M., *et al.*, *Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray*, Science, **270** (1995), 467.

Schmitt, A. P. and McEntee, K., *Msn2p, a Zinc Finger DNA-Binding Protein, is the Transcriptional Activator of the Multistress Response in Saccharomyces cerevisiae*, PNAS, **93** (1996), 5777.

Schwab, S. G., *et al.*, *Evaluation of a Susceptibility Gene for Schizophrenia on Chromosome 6p by Multipoint Affected Sib-Pair Linkage Analysis*, Nat. Genet., **11** (1995).

Schwank, S., *et al.*, *Yeast Transcriptional Activator INO2 Interacts as an Ino2p/Ino4p Basic Helix-Loop-Helix Heteromeric Complex with the Inositol/Choline-Responsive Element Necessary for Expression of Phospholipid Biosynthetic Genes in Saccharomyces cerevisiae*, Nucleic Acids Res., **23** (1995), 230.

Schwartz, D. and Parker, R., *Mutations in Translation Initiation Factors Lead to Increased Rates of Deadenylation and Decapping of Yeast mRNA*, Mol. Cell. Biol., **19** (1999), 5247.

Shalgi, R., *et al.*, *A Catalog of Stability-Associated Sequence Elements in 3′ UTRs of Yeast mRNAs*, Genome Biol., **6** (2005).

Shendure, J. and Ji, H., *Next-Generation DNA sequencing*, Nature Biotechnol., **26** (2008), 1135.

Siggers, T., *et al.*, *Non-DNA-Binding Cofactors Enhance DNA-Binding Specificity of a Transcriptional Regulatory Complex*, Mol. Sys. Biol., **7** (2011).

Singh, N. N. and Lambowitz, A. M., *Interaction of a Group II Intron Ribonucleoprotein Endonuclease with Its DNA Target Site Investigated by DNA Footprinting and Modification Interference*, J. Mol. Biol., **309** (2001), 361.

Smith, C. L., *et al.*, *Structural Analysis of the Yeast SWI/SNF Chromatin Remodeling Complex*, Nat. Struct. Biol., **10** (2003), 141.

Smith, E. N. and Kruglyak, L., *Gene-Environment Interaction in Yeast Gene Expression*, PLoS. Biol., **6** (2008), e83.

Smith, P. *et al.*, *A Genome Wide Linkage Search for Breast Cancer Susceptibility Genes*, Genes Chro. Cancer, **45** (2006), 646.

Smith, R. L. and Johnson, A. D., *Turning Genes Off by Sssn6-Tup1: a Conserved System of Transcriptional Repression in Eukaryotes*, Trends Biochem. Sci., **25** (2000), 325.

Spellman, P. T., *et al.*, *Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization*, Mol. Cell. Biol., **9** (1998), 3273.

Sterne-Weller, T., *et al.*, *Loss of Exon Identity is a Common Mechanism of Human Inherited Disease*, Genome Res., **21** (2011), 1563.

Storici, F. and Resnick, M. A., *The Delitto Perfetto Approach to In Vivo Site-Directed Mutagenesis and Chromosome Rearrangement with Synthetic Oligonucleotides in Yeast*, Methods in Enzymology, **409** (2006), 329.

Struhl, K., *Histone Acetylation and Transcriptional Regulatory Mechanisms*, Genes Dev., **12** (1998), 599.

Swan, K. A., *et al.*, *High-Throughbput Gene Mapping in Caenorhabditis elegans*, Genome Research, **12** (2002), 1100.

Szerlong, H., Saha, A., and Cairns, B. R., *The Nuclear Actin-Related Proteins Arp7 and Arp9: a Dimeric Module that Cooperates with Architectural Proteins for Chromatin Remodeling*, EMBO J., **22** (2003), 3175.

Tachibana, C., *et al.*, *Combined Global Localization Analysis and Transcriptome Data Identify Genes that are Directly Coregulated by Adr1 and Cat8*, Mol. Cell. Biol., **25** (2005), 2138.

Taft, R. J., Pheasant, M., and Mattick, J. S., *The Relationship Between Non-Protein-Coding DNA and Eukaryotic Complexity*, Bioessays, **29** (2007), 288.

Takemaru, K., *et al.*, *Yeast Coactivator MBF1 Mediates GCN4-Dependent Transcriptional Activation*, Mol. and Cell. Biol., **18** (1998), 4971.

Talibi, D., Grenson, M., and Andre, B., *Cis- and Trans-Acting Elements Determining Induction of the Genes of the Gamma-Aminobutyrate (GABA) Utilization Pathway in Saccharomyces cerevisiae*, Nucleic Acids Res., **23** (1995), 550.

Tanaka, K., *et al.*, *IRA2, a Second Gene of Saccharomyces cerevisiae that Encodes a Protein with a Domain Homologous to Mammalian Ras GTPase-Activating Protein*, Mol. Cell Biol., **10** (1990), 4303.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J., *A Genomic Perspective on Protein Families*, Science, **278** (1997), 631.

Tedford, K., *et al.*, *Regulation of the Mating Pheromone and Invasive Growth Responses in Yeast by Two MAP Kinase Substrates*, Current Biology, **7** (1997), 228.

Tenenbaum, S. A., *et al.*, *Identifying mRNA Subsets in Messenger Ribonucleoprotein Complexes by Using cDNA Arrays*, PNAS, **97** (2000), 14085.

Thomas, D., Jacquemin, I., and Surdin-Kerjan, Y., *MET4, a Leucine Zipper Protein, and Centromere-Binding Factor 1 Are Both Required for Transcriptional Activation of Sulfur Metabolism in Saccharomayces cerevisiae*, Mol. and cell. Biol., **12** (1992), 1719.

Tomari, Y. and Zamore, P. D., *Perspective: Machines for RNAi*, Genes Dev., **19** (2005), 517.

Towbin, H., Staehelin, T., and Gordon, J., *Electrophoretic Transfer of Proteins from Polyacrylamide Gels to Nitrocellulose Sheets: Procedure and Some Applications*, PNAS, **76** (1979), 4350.

Tuch, B. B., *et al.*, *The Evolution of Combinatorial Gene Regulation in Fungi*, PLoS Biol., **6** (2008), e38.

van Drogen, F., *et al.*, *MAP Kinase Dynamics in Response to Pheromones in Budding Yeast*, Nature Cell Biol., **3** (2001), 1051.

van't Veer, L. J. *et al.*, *Expression Profiling Predicts Outcome in Breast Cancer*, Nature, **415** (2002), 530.

Vasiljeva, L., *et al.*, *The Nrd1-Nab3-Sen1 Termination Complex Interacts with the Ser5-Phosphorylated RNA Polymerase II C-terminal Domain*, Nat. Struct. Mol. Biol., **15** (2008), 795.

Vasudevan, S., Seli, E., and Steitz, J. A., *Metazoan Oocyte and Early Embryo Development Program: a Progression Through Translation Regulatory Cascades*, Genes Dev., **20** (2006), 138.

Wahl, M. C., Will, C. L., and Luhrmann, R., *The Splicosome: Design Principles of a Dynamic RNP Machine*, Cell, **136** (2009), 701.

Wang, D., *et al.*, *Evidence that Intermolecular Interactions are Involved in Masking the Activation Domain of Transcriptional Activator Leu3p*, J. Biol. Chem., **272** (1997), 19383.

Wang, Y. and Dohlman, H. G., *Pheromone-Regulated Sumoylation of Transcription Factors That Mediate the Invasive to Mating Developmental Swith in Yeast*, J. of Biol. Chem., **281** (2006), 1964.

Wang, Y., *et al.*, *Precision and Functional Specificity in mRNA Decay*, PNAS, **99** (2002), 5860.

Watson, J. D. and Crick, F. H., *The Structure of DNA*, Cold Spring Harb. Symp. Quant. Biol., **18** (1953), 123.

Wickens, M., *et al.*, *A PUF Family Portrait: 3′ UTR Regulation as a Way of Life*, Trends in Genet., **18** (2002), 150.

Wilhelm, B. T., *et al.*, *Dynamic Repertoire of a Eukaryotic Transcriptome Surveyed at Single-Nucleotide Resolution*, Nature, **453** (2008), 1239.

Wilmes, G. M., *et al.*, *A Genetic Interaction Map of RNA Processing Factors Reveals Links Between Sem1/Dss1-Containing Complexes and mRNA Export and Splicing*, Mol. Cell, **32** (2008), 735.

Winzeler, E. A., *et al.*, *Direct Allelic Variation Scanning of the Yeast Genome*, Science, **281** (1998), 1194.

Wolf, J. J., *et al.*, *Feed-Forward Regulation of a Cell Fate Determinant by an RNA-Binding Protein Generates Asymmetry in Yeast*, Genetics, **185** (2010), 513.

Wolffe, A. P., *Histone Deacetylase: A Regulator of Transcription*, Science, **272** (1996), 371.

Wyrick, J. J. and Young, R. A., *Deciphering Gene Expression Regulatory Networks*, Curr. Opinion in Genetics and Development, **12** (2002), 130.

Yamaguchi-Iwai, Y., Dancis, A., and Klausner, R. D., *AFT1: a Mediator of Iron Regulated Transcriptional Control in Saccharomyces cerevisiae.*, EMBO J., **14** (1995), 1231.

Yamazaki, T., *et al.*, *FUS-SMN Protein Interactions Link the Motor Neuron Diseases ALS and SMA*, Cell Rep., **2** (2012), 799.

Yosefzon, Y., *et al.*, *Divergent RNA Binding Specificity of Yeast Puf2p*, RNA, **17** (2011), 1479.

Young, E. T., Kacherovsky, N., and Van Riper, K., *Snf1 Protein Kinase Regulates Adr1 Binding to Chromatin but Not Transcription Activation*, J. Biol. Chem., **277** (2002), 38095.

Zeng, Z. B., *Precision Mapping of Quantitative Trait Loci*, Genetics, **136** (1994), 1457.

Zhao, H., *et al.*, *Regulation of Zinc Homeostasis in Yeast by Binding of the ZAP1 Transcriptional Activator to Zinc-Responsive Promoter Elements*, J. Biol. Chem., **273** (1998), 28713.

Zhou, H. and Winston, F., *NRG1 is Required for Glucose Repression of the SUC2 and GAL Genes of Saccharomyces cerevisiae*, BMC Genet., **2** (2001).

Zhu, J., *et al.*, *Integrating Large-Scale Functional Genomic Data to Dissect the Complexity of Yeast Regulatory Networks*, Nature Genetics, **40** (2008), 854.

Zofall, M., *et al.*, *Chromatin Remodeling by ISW2 and SWI/SNF Requires DNA Translocation Inside the Nucleosome*, Nat. Struct. Mol. Biol., **13** (2006), 339.