Copyright

by

Akshay Anant Bhinge

2009

The Dissertation Committee for Akshay Anant Bhinge Certifies that this is the approved version of the following dissertation:

A Functional Genomics Approach To Map Transcriptional and Post-transcriptional Gene Regulatory Networks

Committee:

Vishwanath Iyer, Supervisor

Makkuni Jayaram

Krishnendu Roy

Scott Stevens

Christopher Sullivan

A Functional Genomics Approach To Map Transcriptional and Post-transcriptional Gene Regulatory Networks

by

Akshay Anant Bhinge, M.B.B.S.; M.Tech.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2009

Dedication

I dedicate this work

to my parents who, against great odds, gave me the opportunity to pursue my goals

and

to Seema, my best friend and the love of my life.

Acknowledgements

I would like to thank my advisor Dr. Vishwanath Iyer for his incredible patience as he mentored me throughout graduate school. He always made it a priority to make himself available to his students. Every project idea, however wacky, that I have presented to him was granted serious thought and critical comments. Without his constant encouragement, I would not have made it this far. I would also like to thank my committee members for their insightful and helpful suggestions.

To my family, I am very grateful. I would not have been able to pursue graduate studies without my parent's unwavering support or my brother's motivation. I want to thank Yogish for sparking my interest in science. He has been more like a mentor than just a good neighbor. I have been friends with Ganapathy as far back as I can remember and have turned to him for advice on numerous occasions. I want to thank all my friends from Grant Medical College, IIT-Bombay and IISc. They have made my otherwise drab academic journey a whole lot interesting.

The best thing that happened to me in medical school was meeting my wife, Seema. Her love and care has kept me going through the worst. Its very comforting to know there are some aspects of your life you can always count on.

I want to thank the present and past members of the Iyer lab for enriching my graduate life, especially Jonghwan Kim from whom I gained a lot of experimental skills. My thanks to Dr. Chris Sullivan and his lab members for their help and guidance in the microRNA project. I would also like to thank Dr. Edward Marcotte for scientific discussions during our weekly lab meetings. His ability to describe complex ideas in the simplest few words never ceases to amaze me.

A Functional Genomics Approach To Map Transcriptional and Post-transcriptional Gene Regulatory Networks

Publication No._____

Akshay Anant Bhinge, Ph.D.

The University of Texas at Austin, 2009

Supervisor: Vishwanath Iyer

It has been suggested that organismal complexity correlates with the complexity of gene regulation. Transcriptional control of gene expression is mediated by binding of regulatory proteins to cis-acting sequences on the genome. Hence, it is crucial to identify the chromosomal targets of transcription factors (TFs) to delineate transcriptional regulatory networks underlying gene expression programs. The development of ChIP-chip technology has enabled high throughput mapping of TF binding sites across the genome. However, there are many limitations to the technology including the availability of whole genome arrays for complex organisms such human or mouse. To circumvent these limitations, we developed the Sequence Tag Analysis of Genomic Enrichment (STAGE) methodology that is based on extracting short DNA sequences or "tags" from ChIP-enriched DNA. With improvements in sequencing technologies, we applied the recently developed ChIP-Seq technique i.e. ChIP followed by ultra high throughput sequencing, to identify binding sites for the TF E2F4 across the human genome. We

identified previously uncharacterized E2F4 binding sites in intergenic regions and found that several microRNAs are potential E2F4 targets.

Binding of TFs to their respective chromosomal targets requires access of the TF to its regulatory element, which is strongly influenced by nucleosomal remodeling. In order to understand nucleosome remodeling in response to transcriptional perturbation, we used ultra high throughput sequencing to map nucleosome positions in yeast that were subjected to heat shock or were grown normally. We generated nucleosome remodeling profiles across yeast promoters and found that specific remodeling patterns correlate with specific TFs active during the transcriptional reprogramming.

Another important aspect of gene regulation operates at the post-transcriptional level. MicroRNAs (miRNAs) are ~22 nucleotide non-coding RNAs that suppress translation or mark mRNAs for degradation. MiRNAs regulate TFs and in turn can be regulated by TFs. We characterized a TF-miRNA network involving the oncofactor Myc and the miRNA miR-22 that suppresses the interferon pathway as primary fibroblasts enter a stage of rapid proliferation. We found that miR-22 suppresses the interferon pathway by inhibiting nuclear translocation of the TF NF-kappaB. Our results show how the oncogenic TF Myc cross-talks with other TF regulatory pathways via a miRNA intermediary.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

EUKARYOTIC TRANSCRIPTIONAL REGULATION

How complex are we?

Even before the human genome was sequenced, it was largely known that genome size does not correlate with organismal complexity [1]. For example, the human genome is 200 times larger than that of the yeast *S.cerevisiae* but 200 times smaller than that of *Amoeba Dubia*. This was attributed to the fact that only a small percentage of the genome is dedicated to protein-coding genes. It was naturally expected that complexity will correlate with the number of genes an organism has and earlier estimates of the total number of human genes hovered in the range 60,000-150,000 [2]. Hence, after the human genome was sequenced, it was a surprise to note that the total number of human genes predicted was somewhere around 20,000 [3]. On the contrary, the genome sequence of the nematode *C. elegans* with ~1000 cells was predicted to have about 19,000 genes [2]. One significant difference between the human and worm genomes was the higher number of alternatively spliced transcripts found in the human genome that could potentially give rise to 100,000 different proteins [2]. In addition to alternative splicing, it has been suggested that organismal complexity arises from elaborate regulation of gene expression [4].

Eukaryotic transcription factors

Almost 5-10% of the total coding capacity of metazoans is dedicated to proteins regulating transcription [4]. Regulatory proteins that bind DNA in a sequence-specific manner and mediate gene-selective transcriptional activation or repression are called transcription factors (TFs). The yeast genome encodes a total of ~200 transcription

factors, the worm and fruit-fly genome encode roughly 700-800 transcription factors while the human genome encodes almost 2000 transcription factors [5]. Thus, organismal complexity apparently correlates with the both the absolute numbers and the relative proportion of TFs per genome. On average, yeast contains one TF per 25 genes while humans possess one TF per 12 genes. Due to the combinatorial nature of transcriptional regulation, this two-fold increase in the TF proportion might translate into several-folds of increased complexity at the network level [4].

Regulatory DNA sequences

Central to the process of differential gene regulation are cis-acting sequences that comprise eukaryotic promoters. Eukaryotic promoters are generally characterized by three basic features: the transcription start site (TSS), the TATA box and sequences bound by transcriptional regulators [6]. The TSS and the TATA box comprise what is known as the core promoter, which is sufficient for transcription initiation by the basal transcriptional machinery but has limited regulatory potential [6]. Regulatory sequences that are bound by transcription factors include activators, repressors, enhancers and silencers [7]. Some of these sequence elements such as enhancers can act in an orientation independent manner and can influence gene expression at large distances [8].

Transcriptional regulatory networks

The complete set of interactions between all transcription regulatory proteins and their corresponding cis-acting regulatory elements active in any given state of a cell is termed as a transcriptional regulatory network [9]. Cells commonly respond to external physiological stimuli by altering their transcriptional networks and errors in the network connectivity often lead to adverse phenotypes including disease [9]. Hence it is crucial to

understand how transcriptional networks are maintained to regulate gene expression under varying conditions.

Recent advances in high-throughput techniques have enabled researchers to gain a systems level understanding of how cellular networks operate. The advent of microarrays has made it possible to measure RNA expression patterns on a genome-wide scale [10]. By combining chromatin immunoprecipitation (ChIP) with microarray hybridization (ChIP-chip) it is now possible to map binding locations for regulatory proteins across the entire genome in a high-throughput manner [11, 12]. Another powerful approach involves "perturbing" the system by environmental means or by selected knockdown of genes by RNAi or genetic manipulation and assaying the effect of this perturbation on gene expression by microarrays [13]. Analysis of data generated by these high-throughput techniques has provided preliminary maps of regulatory networks. Combination of microarray data with sequence analysis has shown that transcriptional networks have a modular architecture [14, 15], while network reconstructions based on ChIP-chip data [16] or the perturbation approach have shown that environmental stimuli play a significant role in selecting transcriptional targets [13]. However, such genome-wide data is available only for lower eukaryotes like *S.cerevisiae* but sorely lacking in higher animals due to the cost and labor involved in conducting such experiments. Additionally, metazoan gene regulation has additional levels of complexity that include chromatin remodeling and post-transcription regulation.

ROLE OF CHROMATIN

The strongly negatively charged DNA phosphate backbone is neutralized by binding to basic histone proteins, enabling the compaction of the genome into chromosomes by a factor of almost 10000 [17]. However, this packaging of DNA into the nucleus creates a problem for the regulatory machinery to access cis-acting sequences.

The evolutionary solution to the problem of packaging the genome into a compact structure that is still accessible to regulatory proteins is the nucleosome [17].

The nucleosome

The nucleosome is elementary repeating unit of eukaryotic chromatin, composed of a histone protein core around which 147 bp of DNA is wrapped in 1.65 helical turns [18]. The histone core is made up of two copies each of the proteins H2A, H2B, H3 and H4 [19]. The genomic region in between consecutive nucleosomes is termed the linker region to which an additional histone H1 can bind and further compact the chromatin into 30 nm fibres [17]. Nucleosomes are arranged as a linear array along the genomic DNA giving rise to the "beads on a string" appearance [19].

Histone modifications and variants

Emanating from the histone core are amino-terminal "tails", which are subject to several covalent modifications that includes acetylation, methylation and phosphorylation [20]. Acetylation of histones is typically accompanied by nucleosome eviction while de-acetylation of histones is the preceding event required to generate a repressive chromatin structure [21]. Acetylation and phosphorylation have been implicated in activation of transcription while deactylation is associated with repression [20]. Methylation is associated with both activation and repression depending upon the context [20].

Further modulation of the nucleosomal unit is brought about by replacing the core histones by histone variants. At promoters that are active or poised for activation, the histones H3 and H4 are replaced by the histone variants H2AZ or H3.3 [22]. The H2A variant H2AX is frequently found at double-stranded DNA breaks and phosphorylation of H2AX is essential to form competent repair foci [22].

Chromatin dynamics

Nucleosome assembly, disassembly and mobilization can have a profound influence of gene regulation [23]. Hence, it is important to understand the underlying determinants of nucleosome positioning and remodeling.

Sequence determinants of nucleosome positioning

Alignment of the 147 bp of DNA sequence underlying thousands of well-positioned nucleosomes has revealed particular base combinations that are statistically enriched. The dinucleotides AA, TT, TA and GC occur at a 10 bp periodicity along the nucleosomal DNA sequence [24, 25]. The GC dinucleotide periodicity is offset by 5 bp as compared to the AA, TT and TA patterns [24, 25]. Additional nucleotide and structural patterns also seem to correlate with nucleosome positioning, however these are not considered to be universal determinants of nucleosome positioning [26]. It might be advantageous to the genome to employ favorable as well as unfavorable sequences, which results in sub-optimally positioned nucleosomes [17]. Such nucleosomes can restrict access to transcription factors but can be evicted easily enough to grant access to the regulatory machinery when needed [17].

Despite the statistical enrichment of the AA, TT, TA and GC patterns, the presence of these individual patterns in DNA sequences underlying *in vivo* positioned nucleosomes is only modestly abundant as compared to random sequences [25]. A recent prediction algorithm based solely on sequence determinants was able predict nucleosomal positions only marginally better than random guessing [25]. Thus the underlying sequence, though important in positioning, may not be the sole player.

Chromatin remodelers

Chromatin remodeling to allow access to sites buried underneath a nucleosome may require ATP-dependent chromatin remodeling complexes [27]. These ATP-dependent enzymes regulate nucleosome dynamics in multiple ways including transient exposure of DNA loops, translational repositioning, nucleosomal eviction and reassembly and replacement of histone subunits by variants [27, 28].

Nucleosome dynamics regulate DNA accessibility thereby influencing gene regulation and transcriptional fidelity. Hence, better understanding of chromatin remodeling will provide deeper insights into transcriptional regulation.

POST-TRANSCRIPTIONAL GENE REGULATION

An efficient and rapid way to modulate gene expression is to regulate mRNAs that have already been transcribed. Alterations of mRNA stability and/or translational efficiency are important components of metazoan gene regulation and have been linked to oncogenesis [29]. An additional player was introduced into the post-transcriptional regulation game with the discovery of microRNAs. MicroRNAs (miRNAs) are 20-24 nucleotide (nt) non-coding RNAs that regulate gene expression post-transcriptionally by translational repression or enhanced mRNA degradation [30]. Hundreds of miRNAs have been discovered so far in animals and plants where they are predicted to control thousands of genes [31].

Discovery

The first miRNA to be discovered, lin-4, was identified in a genetic screen for defects in temporal control of post-embryonic development in the nematode *C.elegans* [32]. Mutations in lin-4 disrupt the temporal regulation of larval development [33]. On the other hand, worms deficient in lin-14, which is a protein-coding gene, display

opposing phenotypes to those observed in lin-4 mutants [34]. It was found that lin-4 encodes a 22 nucleotide non-coding RNA that base pairs imperfectly with the 3' untranslated (3' UTR) region of lin-14 and downregulates lin-14 protein expression [32]. The suppression of lin-14 by lin-4 requires a functional lin-4 gene and an intact lin-14 3' UTR [32, 35]. The suppression was due to an inhibition of protein synthesis alone and did not affect lin-14 transcript levels [36]. Almost 7 years after the discovery of lin-4, another miRNA, let-7 was discovered in *C.elegans* [37]. The miRNA let-7 is a temporally regulated 21 nucleotide small RNA that inhibits translation of lin-41 and hbl-1 by binding to their 3' UTRs [38]. Both let-7 and lin-41 orthologs were observed in diverse species such as sea urchins, flies, mice and humans [39]. This extensive conservation indicated a much wider and general role for small RNAs in development across different species. MiRNAs have been implicated in diverse biological processes such as apoptosis, oncogenesis, stress adaptation, hormone signaling as well as development [38].

Biogenesis

Most miRNA genes are transcribed from their own promoter but some are found in the introns of known genes either in the sense or the anti-sense direction [40, 41]. Transcription of miRNA genes is mediated by RNA polymerase II into primary transcripts or pri-miRNAs. Pri-miRNAs can be several kilobases long and contain a hairpin structure that has a 60-100 nt stem loop with a 2 nt overhang at the 3'end [42]. These hairpin structures are recognized and further processed by the RNaseIII enzyme Drosha in complex with the protein DGCR8 or Pasha into precursor miRNAs (pre-miRNAs) [43]. Pre-miRNAs are exported out of the nucleus by exportin-5 [44] where they are further processed into mature miRNA duplexes by the cytoplasmic RNaseIII enzyme Dicer [42].

Mode of action

Mature miRNAs are incorporated into effector ribonucleoprotein complexes called miRNA containing RNA-induced silencing complexes or miRISC. Only one strand of the duplex miRNA is retained in the miRISC assembly. This strand is termed the guide strand while the anti-guide or sense strand is degraded [45]. Selection of the guide strand is directed by thermodynamic considerations. The strand that has a lower stability at the 5' end of the duplex is selected as the guide strand and is incorporated into miRISC. The guide strand now directs the miRISC assembly towards the target mRNA [42]. The exact mechanism of selection of miRNA binding sites on target mRNAs is still unclear. However, certain rules seem to apply: 1) Binding sites are usually found in 3' UTRs of target genes though functional sites in 5' UTRs and even in coding regions have been documented [46]. 2) There is a requirement for complementarity between the 5' end of the miRNA and the mRNA especially nucleotides 2-8 on the miRNA, called the "seed" region [47]. 3) Imperfect base-pairing between the miRNA and the target results in translational repression while perfect pairing results in mRNA cleavage [48]. 4) Imperfect base pairing at the seed region can be compensated for by additional base pairing at the 3' region [30]. 5) The sequence context within which the binding site resides has also been shown to affect targeting. For example, if the site is located within a hair-pin structure, it can impede access to the miRISC assembly thereby avoiding regulation [49]. 6) Multiple binding sites are often used, especially in animals, and may act cooperatively [49].

TF-miRNA interactions

An interesting paradigm that has emerged from recent studies into miRNA mediated gene regulation is that miRNAs and TFs reciprocally regulate each other. Many such examples have been documented so far in diverse biological processes. For

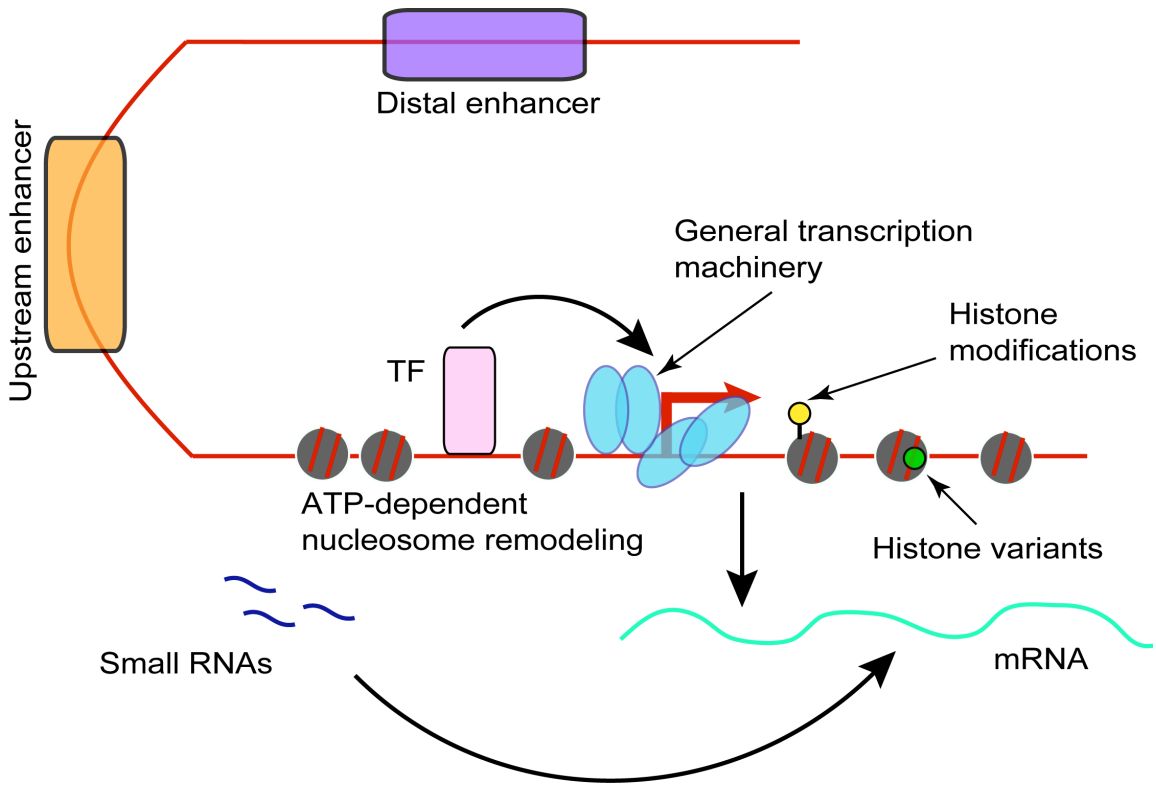


Figure 1.1 Complexity of eukaryotic gene regulation

TFs bind in a sequence-specific manner to eukaryotic promoters and activate (or repress) transcription. Gene expression can be influenced over long distances by enhancers, ATP-dependent remodelers or histone modifications/replacement. Additionally, the encoded transcript may be under post-transcriptional regulation by small RNAs such as microRNAs. It must be noted that the encoded transcript itself may code for a transcription factor or get processed to generate a microRNA.

example, the TF Myc activates the miR-17-92 cluster of miRNAs as well as another TF E2F1, while the miR-17-92 cluster suppresses E2F1 [50]. Reciprocal regulation of miRNAs in the miR-17-92 cluster and the TF AML/Runx1 by each other has also been recently reported [51]. It has been suggested that miRNAs fine-tune the broad expression programs set up by TFs [52] and that miRNAs and TFs are linked in functional

regulatory networks that modulate gene expression during normal cell growth as well as cancer [50, 53-55]. As the role of miRNAs in human disease conditions becomes increasingly prominent, it is critical to further our understanding of their role in gene regulation especially in the larger context of transcriptional networks.

The complexity of metazoan gene regulation is summarized in Fig. 1.1. It is becoming increasingly clear that in order to gain a complete understanding of gene regulatory mechanisms, it will be necessary to integrate data from the transcriptional, epigenetic as well as post-transcriptional aspects of regulatory networks.

Chapter 2: Mapping Genome-wide Transcription Factor Binding Sites Using STAGE

INTRODUCTION

The ENCODE (Encyclopedia of DNA Elements) project has suggested that a large fraction of the human genome is transcriptionally active [56]. Thus a significant portion of the human genome may be involved in regulating key aspects of human biology. Majority of this regulatory potential is exercised by the site-specific binding of transcription factors to open chromatin. Hence, identifying the genomic binding sites of regulatory proteins is important in deciphering gene regulatory networks [56]. Chromatin immunoprecipitation followed by microarray hybridization is a powerful technique to map transcription factor binding sites in the human genome on a global scale [11, 12]. Whole genome tiling arrays are now available enabling resolutions on a scale of a few hundred base pairs [57]. However, these arrays are expensive, require a large amount of starting material and suffer from the common limitations of microarrays such as potential for cross-hybridization and low dynamic range. With the advent of ultra high throughput sequencing technologies such as Solexa, ChIP-Sequencing i.e. ChIP followed by sequencing (described in a later section) has become the method of choice for identifying genome-wide TF binding sites. However, back in 2003, such technologies were not available to researchers. To overcome these obstacles, we developed a sequencing based technique called Sequence Tag Analysis of Genomic Enrichment (STAGE) (Fig. 2.1) to map transcription factor binding sites across the whole genome in an unbiased fashion. STAGE using conventional Sanger sequencing was used to map E2F4 binding sites primarily to establish a proof-of-principle [58]. STAGE using 454 sequencing was

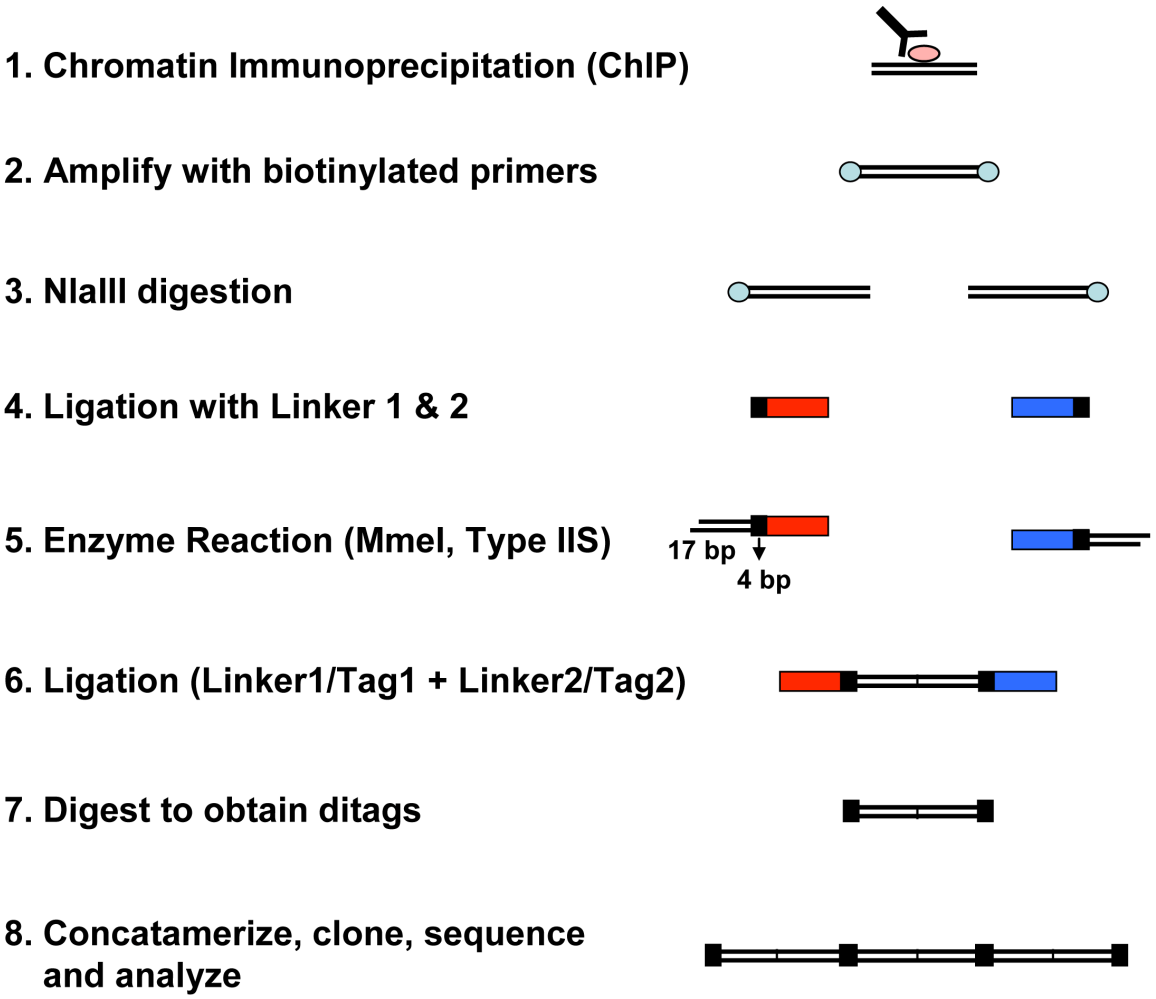


Figure 2.1 STAGE methodology

applied to map genome-wide binding sites of the transcription factors STAT1 and Myc [59]. The technology was further improved to increase throughput by 200%.

E2F4 STAGE

E2F4 belongs to the E2F family of transcription factors, which is composed of 8 members, namely E2F1-8 and primarily involved in cell cycle progression and differentiation [60]. E2F4 is a repressor and mainly localizes to the cytoplasm under conditions of cell proliferation [61]. On withdrawal of mitogenic stimuli, as cells start entering the G0 phase of the cell cycle, E2F4 heterodimerizes with the DP proteins that facilitate the nuclear translocation of E2F4 into the nucleus in CRM1 dependent manner [61]. Nuclear E2F4 complex binds to several hundred promoters and suppresses diverse classes of genes involved in cell cycle, DNA repair and apoptosis [62]. Abnormal expression of E2F4 has been associated with various tumors such as colorectal carcinomas, endometrial cancers, gastric adenocarcinomas and prostatic cancers indicating the importance of E2F4 in tumorigenesis [63].

At the time of this study a few attempts to elucidate E2F4 binding sites had been made using ChIP-chip on core promoter microarrays [62, 64]. However, an unbiased query of E2F4 binding sites across the human genome had not been attempted. Additionally, since a ChIP-grade antibody for E2F4 was available, it made sense to apply the STAGE technology to identify E2F4 target sites as a proof-of-principle.

RESULTS

The STAGE methodology is conceptually derived from the Serial Analysis of Gene Expression (SAGE) technique where 21 bp DNA sequences or “tags” are extracted from ChIP-enriched DNA and concatamerized for sequencing. Details of the STAGE method have been described in Materials and Methods. Each sequenced DNA

concatamer contained on an average 30 tags that were approximately 20-22 bp in length. Since the enzyme NlaIII was used to digest the ditags, each tag has a CATG at the 5' end. The enzyme MmeI releases 21 bp tags but sometimes may cut 20 or 22 bp from its recognition sequence. We ignored the 20 and 22 bp tags and only 21 bp tags with a CATG sequence at the 5' end were retained and constituted the sequencing pool of tags from E2F4 STAGE. We analyzed ~3500 valid tags to identify E2F4 targets in human cells by STAGE. STAGE tags now had to be mapped back to the human genome to generate coordinates. We initially tried using the BLAST software to identify the genomic coordinates of the STAGE tags but this process was very slow. So we developed a faster to map STAGE tags to the genome.

Since all STAGE derived tags had a CATG at the 5' end, we scanned the human genome (June 2003 Build 34) for the sequence CATG and extracted 21 nucleotides or "tags" from the beginning of the CATG sequence. As CATG is a palindrome, for a given CATG we extracted tags from either strand and stored the coordinate of the 5' end of each tag. We sorted all CATG tags per chromosome alphabetically, thus generating chromosome-specific sorted CATG tag libraries that had the genomic coordinates of each tag stored in the same file. To map a STAGE derived tag to the genome, we simply mapped it to our genomic tag libraries using a binary search algorithm and extracted the coordinate corresponding to that tag. Using this method, we could map 1500 tags to the entire human genome in less than 4 minutes.

The human genome has approximately 20,000 genes. We first analyzed those tag sequences generated by the STAGE procedure that were found within 2 kb upstream and 1 kb downstream of a transcription start site of each gene. Because a large number of tags cannot be mapped unambiguously to a given location on the genome, and because of the potential of observing tags that arise from non-specific genomic background, we

developed an algorithm to score tags and target genes as putative targets. The algorithm took into account, the uniqueness of a given tag, the occurrence of the tag in our STAGE tag pool, and the probability of multiple tags clustering near each other on the genome to indicate true binding events. We assigned each individual unique tag a score based on its number of hits on the human genome and number of occurrences in the total STAGE pool. In summary, each hit on the genome lowers the tag score, and each occurrence in the STAGE pool raises the score. The final score of each gene was calculated by adding the scores of the tags found within the 3 kb window. Details of the algorithm have been described below.

Identifying significant tags

We associated two numbers with each tag. The number of times the tag was found in the STAGE sequencing pool of tags was referred to as **nocc** while the number of times that tag was found in the genome was referred to as **nhit**. Higher the nocc, higher the frequency of that tag in the ChIP-enriched DNA derived sequencing pool and hence higher the likelihood that the tags represents a true binding event. However, some tags are found in repeat regions of the genome and may have as many as 10,000 hits across the entire genome. These tags would get sequenced multiple times merely by chance. Hence the nocc had to be normalized to the nhit. In order to do this, we first calculated the expected nocc of a tag in the STAGE sequencing pool.

(E_nocc) Expected nocc = p*nhit*(total number of sequenced tags) where p = the probability of selecting a single tag from the entire genomic CATG tag library i.e. 27 x 10^6 tags.

$p = 1/(27x10^6)$

Tag value = (observed nocc – E_nocc)/E_nocc

These tag values represent the likelihood of the tag being derived from a true binding event. The above calculated tag values were overtly sensitive to nhit and had values ranging from 0 to 50,000. Hence these tag values were scaled using a hyperbolic function such that the maximum possible score was 1000. The hyperbolic function was chosen empirically to be

Tag value (scaled) = (1000 * tag value)/(200 + tag value)

The final tag score was assigned as: Tag score = Tag value (scaled) * nocc

Scoring genes as putative targets

To score genes as putative binding sites, tags were mapped to within 2000 bp upstream and 1000 bp downstream of all genes annotated by the RefSeq database. Tags assigned to a given gene were now sorted according to position. The gene score was first initialized to the value of the tag farthest from the transcription start of the gene. Subsequent tag values were now added to this score. If two consecutive tags were the same, PENALTY value was subtracted from the final score. Otherwise for each distinct tag added, REWARD value was added to the final score. For this study the values used for the REWARD and PENALTY were 10 and 5 respectively.

To determine whether a given gene is a true target, gene scores were compared to controls. Two types of controls were used.

Experimental control: STAGE tags were derived from human genomic DNA and gene were assigned scores based on the genomic STAGE tags. The experimental gene scores were divided by the control scores and all genes having a normalized ratio of 900 or more were taken as putative targets.

Computational control: Tags were randomly selected from the whole genome CATG libraries and these simulated tags were used to generate scores for each gene as described above. This simulation was performed 500 times and the average score and the

standard deviation were calculated for each gene. The experimental normalized scores that were greater than 3*standard deviation of the control scores (*P-value* < 0.01) were termed significant.

We calculated the score of every gene using the E2F4 STAGE library and the genomic STAGE library separately. The score of each gene generated from the E2F4 STAGE library of tags was divided by the score of the same gene generated from the control genomic STAGE library. The values generated from this division are akin to a fold enrichment and this was considered as a measurement of promoter enrichment as a result of E2F4 binding.

DISCUSSION

We found that many of the genes with high scores but low fold enrichments were mitochondrial genes. The mitochondrial genes as a group had a tendency towards a higher number of unique tags with relatively low hit numbers. This could be possible due to the high copy numbers (~2000) of mitochondria inside cells. In later versions of the algorithms, the number of hits (nhit) for tags generated from the mitochondrial genome was multiplied by 2000 and this corrected the above mentioned problem.

Appendix A shows the 48 putative target genes of E2F4 detected by STAGE. Most of these genes were designated by at least one very unique tag sequence that had one hit on the entire human genome. The list includes three previously known targets of E2F4, such as RAD54L, SLC3A2 and MAP3K7 that were identified through the use of a human core promoter microarray.

Using E2F4 STAGE, we detected several known targets of E2F4 in addition to new putative targets although we sequenced only a small number of tags. Even though STAGE may have been biased towards the most abundant or strongly enriched targets and needed a greater depth of sequencing to elucidate all TF binding sites, the STAGE

methodology had several significant advantages over then existing methods. First, it provided a new way of studying genome-wide binding distributions without constructing intergenic microarrays. Second, it could be used to study a variety of organisms without previous information about the targets of DNA-binding proteins, as long as the genome sequence of that organism is available. Third, STAGE could be utilized to identify binding loci for any sequence specific transcription factor, chromatin components and modifying factors, DNA replication proteins, repair proteins, and localized mistargeted oncogenic proteins in cancers. Fourth, the STAGE method potentially allowed us to find novel binding sites of proteins in numerous organisms in which genome sequences were not fully annotated. Fifth, STAGE could be used in any conventional laboratory since it did not require any special machines. Finally, small amounts of starting material, for example crosslinked extract from 10^7 cells, was good enough for generating tens of thousands of STAGE tags.

Though the STAGE method provided several advantages to contemporary existing methods, it suffered from a few limitations. To sequence tags in a high-throughput manner by Sanger sequencing, ditags had to be concatamerized and cloned into sequencing vectors. This step was followed by large-scale screening to select clones that had the appropriate size of insert. Needless to say this was a laborious and time-consuming process. With the advent of better sequencing technologies like the 454 pyrosequencing method, ditags could be directly sequenced and this provided a tremendous boost to the STAGE method. STAGE performed with the 454 sequencing technology has been described in the next section.

MATERIAL AND METHODS

ChIP for E2F4

E2F4 ChIP was performed by a former graduate student in the Iyer lab, Jonghwan Kim. A human fibroblast cell line derived from foreskin (ATCC CRL 2091) was grown to about 60% confluence in 15 cm plates in DMEM containing glucose (1 g/liter), antibiotics and 10% FBS (Hyclone). Cells were washed twice with the same medium lacking FBS and low-serum medium (0.1% FBS) was added to the plates. After a 72-hour incubation, formaldehyde was added (final 1%) to crosslink the cells. Anti-E2F4 antibody (sc-1082x, Santa Cruz) at a 1:100 dilution was used for ChIP. ChIP was performed as described in [58].

STAGE using Sanger sequencing

E2F4 STAGE was performed by a former graduate student in the Iyer lab, Jonghwan Kim. A schematic representation of the STAGE technique is shown in Fig. 2.1. STAGE is conceptually derived from SAGE, however, the input to STAGE is ChIP-enriched genomic DNA. Chromatin immuno-precipitation is used to enrich genomic DNA bound by a transcription factor of interest. Enriched DNA was amplified by 5'-biotinylated primers using ligation-mediated PCR. Amplified DNA fragments (1~2ug) were cleaved with NlaIII, bound to streptavidin-coated magnetic beads (Dynal) and separated in two tubes. After ligation to linker 1 or 2, which contain recognition sites for MmeI type IIS restriction endonuclease, the DNA fragments were released from the beads by digestion with MmeI. The released tags were ligated to one another without polishing the ends to generate ditags. For STAGE using Sanger sequencing, ditags were gel extracted, amplified and further digested with NlaIII. Digested ditags were gel-extracted, concatamerized by ligation and cloned into the pZero 1.0 vector (Invitrogen).

After transformation, colonies were screened by PCR to select inserts of appropriates sizes and sequenced.

STAT1 STAGE

In order to make STAGE more competitive with genome-wide tiling arrays, we took advantage of bead-based pyrosequencing 454 technology [65]. 454 sequencing resulted in a dramatic increase in throughput and cost-effectiveness as well as significantly reducing the time and labor needed to perform STAGE. STAGE using 454 technology was used to map the genome-wide binding sites of the transcription factor STAT1. STAT (Signal Transducer and Activator of Transcription) proteins are transcription factors that mediate cellular responses to interferon, cytokine and growth factor signaling. Interferons (IFN) regulate cell proliferation, apoptosis and immune surveillance through direct stimulation of the JAK-STAT pathway [66]. IFN-gamma (IFNG) specifically activates STAT1 which homodimerizes, translocates to the nucleus and activates IFNG inducible genes by binding to the gamma-activating sequence (GAS) motif [67]. ChIP-chip analysis of STAT1 binding sites across chromosome 22 revealed that STAT1 regulates many genes involved in immune response, apoptosis, lipid metabolism and cell growth [68]. We used STAGE followed by 454 sequencing to map binding targets of STAT1 across the entire genome.

RESULTS

STAT1 bound DNA was enriched by ChIP and STAGE was used to generate ditags and amplified by linker-specific primers. Ditags were sequenced by 454 inc without the nebulization step to fragment the DNA. Thus, each read typically consists of two tags flanked by linkers. Linker sequences were removed and 21 bp tags extracted from the reads by custom PERL scripts. We sequenced a total of 179,954 reads from the

STAT1 STAGE pool representing about 17 Mb of total sequence, constituting a single 454 sequencing run. Reads that appeared multiple times were assumed to be PCR artifacts and hence represented just once. After removing duplicated reads, 162,577 tags were extracted and mapped back to the genome. About 19% (31353) tags could not be mapped perfectly to any location and hence were discarded. Further analysis was done with the remaining perfectly matched 131,224 tags.

Since STAGE tags are derived from ChIP enriched DNA, the distribution of tag frequency in the STAGE sequencing library should deviate significantly from a randomly generated library. Simulations were performed where the same number of tags as present in the STAT1 STAGE library (131,224) were selected at random from the entire genome multiple times. Tags that had multiple hits on the genome were discarded and the average frequency distribution of tags in the random library were compared to that of tags from the STAT1 STAGE library. For a frequency of occurrence equal to 1, the random and experimental libraries were similar. However, for a frequency of occurrence greater than 1, the experimental library showed a clear enrichment over the random library (Fig. 2.2A). This shows that the STAGE tag library generated by 454 sequencing is distinct from a randomly simulated background library.

STAGE identified STAT1 targets

Since a significant portion of the human genome is repetitive, tags in the STAGE sequencing library may be derived from repeat regions and hence map to multiple locations on the genome. Tags having multiple hits on the genome might be over represented in the sequencing pool just by chance and not necessarily due to ChIP-enrichment. Hence, to calculate the true enrichment of the tag over background, we have take into account the number of times the tag was found in the genome. This was accomplished by assigning each tag a probability of enrichment by assuming that the

selection of tags from the genome follows a binomial distribution. Specifics about the actual calculations have been given in the Material and Methods section. Since STAGE tags are derived from ChIP-enriched DNA, multiple tags should cluster together in short regions that are comparable to the length of sonicated genomic DNA that was isolated by ChIP. On the contrary, tags derived from a random library generated from input material without any enrichment should be uniformly sampled across much wider regions in genome. This rationale was used to differentiate binding targets from background. In order to find the optimal window that can detect true binding sites, we scanned windows of different lengths (from 100 bp to 2000 bp) across the genome. For each window size, the number of windows with a given number of single-hit tags derived from the STAGE library was recorded. These numbers were compared to those obtained from tags derived from a randomly simulated STAGE library. For a given number of single hit tags found within the window, the ratio of the number of windows found in the random sample to the ratio of the number of windows found in the STAGE library was defined as the false discovery rate (FDR). Based on these random simulations, a window a 500 bp gave a FDR of < 5% while the number of targets detected was 734 (Fig. 2.2B). The complete data for all windows sizes for this simulation has been given in Appendix B. A window of 500 bp was chosen for all further analysis.

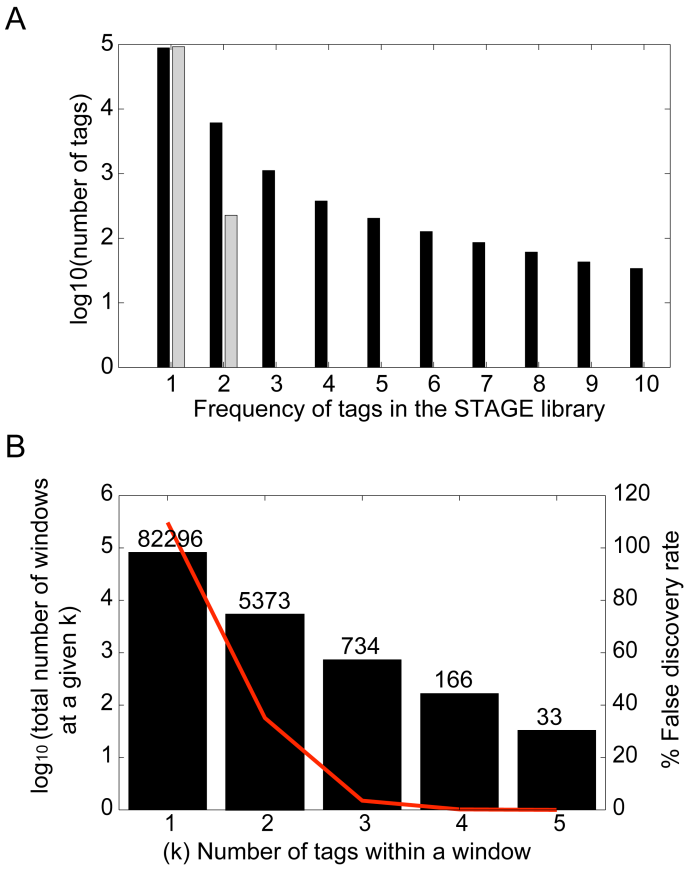


Figure 2.2 STAT1 STAGE tag library

(A) Frequency of tags in the sequencing pool derived from the STAT1 STAGE sequencing library (black bars) was compared to tags generated from a random simulation (gray bars). For a frequency of 2 or more, the STAGE library shows a higher frequency of tags as compared to the simulated library. (B) Unique tags (i.e. tags with nhit = 1 on the genome) were clustered within 500 bp windows. For a given number of unique tags within a window (termed k), we counted the number of windows generated by the STAGE library and the simulated library. The ratio of number of windows in the simulated library to the number of windows in the STAGE library was expressed as a percentage and termed the false discovery rate (FDR). The FDR is shown by the red line and represented on the right Y-axis while the number of windows discovered at each FDR is indicated by the black bars and represented on the left Y-axis on a logarithmic scale. The actual number of windows at each FDR is indicated on the top of each bar. Reproduced from [59].

To improve target detection specificity, a window that passed the above criteria was considered a target only if at least on tag assigned to that window was considered truly enriched. Thus, for each window we calculated two probabilities: a) the probability of finding a given number of tags within that window b) the probability that at least one tag assigned to that window was statistically likely to be enriched. To avoid assigning high probabilities to windows that had a single highly enriched tag, we gave greater weightage to the probability of finding a significant number of tags within the window than simply finding enriched tags. This combined probability calculation gave us a false discovery rate of < 1% at a probability threshold of 0.95 and detected 381 binding sites across the genome. It must be noted that the false discovery rate mentioned was derived from random simulations under the assumption that selection of STAGE tags follows a binomial distribution. However, it is entirely possible that experimental manipulation introduces biases that were not modeled in the simulation.

Verification by ChIP-chip and quantitative PCR

Since whole genome ChIP-chip data was not available for STAT1, we used data from STAT1 ChIP-chip performed on ENCODE regions tiling oligonucleotide arrays [56]. 7 out of the 381 sites that STAGE identified were within the ENCODE regions. 3 out of these 7 overlapped with a STAT1 ChIP-chip peak (Fig. 2.3 upper panel). We further compared STAT1 STAGE binding sites to STAT1 target promoters detected by performing ChIP-chip on a core-promoter microarray that included 9764 different promoters where a promoter was defined as approximately 1000 bp upstream and 200 bp downstream from the transcription start site (TSS) of a gene. Core-promoter ChIP-chip identified 157 promoters bound as STAT1 targets at an enrichment of more than 3-fold. 29 out of the 9764 promoters had a high-confidence STAGE binding site within 1 kb

upstream and 200 bp downstream from the TSS. 11 out of these 29 were common to the 157 targets detected by the microarray analysis (Fig. 2.3 lower panel) (*P-value* assuming a hypergeometric distribution is $< 10^{-12}$).

In order to obtain a quantitative estimate of the false positive rate of the STAGE analysis, we performed quantitative PCR on a biologically independent STAT1 ChIP sample and assayed 10 sites from the 381 binding sites with scores ranging from 0.95 to 1 for enrichment. 9 out of these 10 sites showed a quantitative enrichment over background and 8 out of these 10 had enrichment more than 2-fold (Fig. 2.4). This indicated that the true positive rate of detection was about 90% giving a false discovery rate of 0.1.

Motif analysis

If a STAT1 binding site detected by STAGE occurred within 1 kb upstream and 200 bp downstream of the TSS of a RefSeq annotated gene, that gene was termed a STAT1 target. STAGE detected 59 genes in RefSeq as STAT1 targets by the above criteria. Of these genes, 62% had a GAS motif within 1 kb upstream and 200 bp downstream of their TSS representing an enrichment greater than 2-fold over background (*P-value* $< 10^{-8}$ assuming a hypergeometric distribution). Background in this case was considered as 1 kb upstream and 200 bp downstream from the TSS of all genes in RefSeq. We applied the same analysis for genome-wide STAT1 binding sites. For each window detected as a STAT1 binding site, we searched for the GAS motif within that window extending our search to 250 bp on either side. Out of 381 binding sites detected by STAGE, 226 (~59%) had the GAS consensus sequence representing an enrichment of more than 2-fold over background level of occurrence of the GAS motif in randomly selected windows from the entire genome (*P-value* $< 10^{-43}$) (Fig. 2.5). Further, in accordance with the observation that STAT1 is known to exhibit cooperative binding with other transcription factors like AP1 [69], MYC [70]

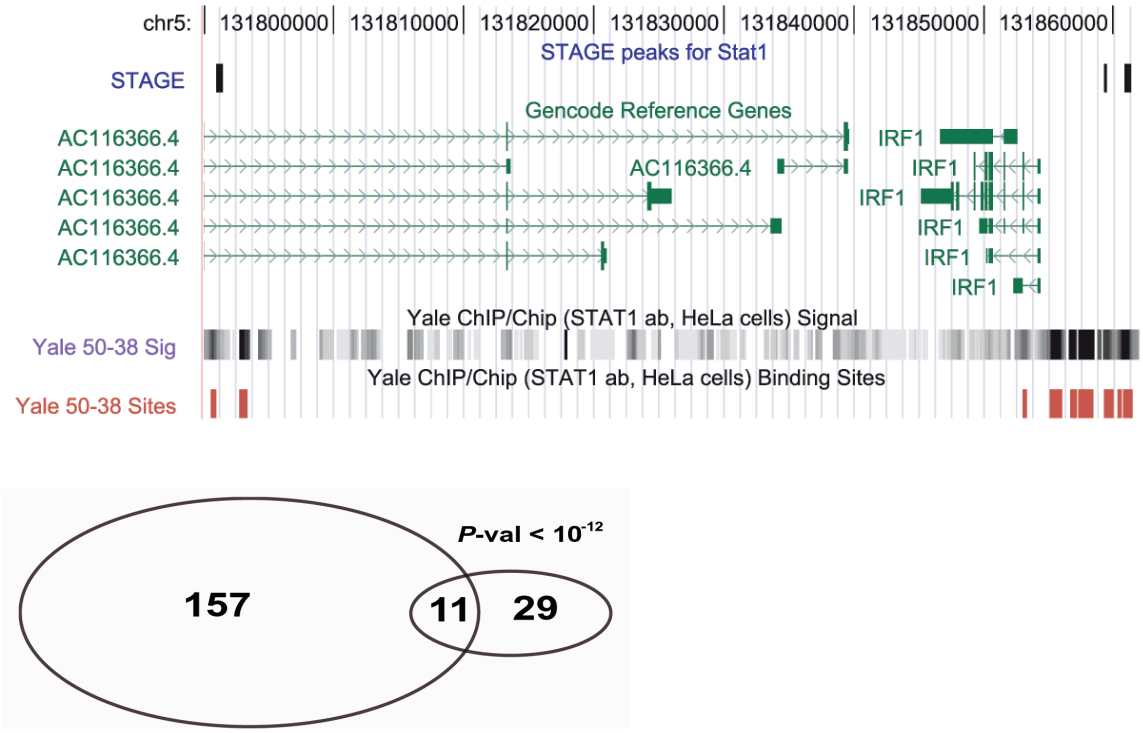


Figure 2.3 Comparing STAGE results with ChIP-chip targets

Upper panel shows a screen shot of the UCSC genome browser. STAGE detected targets are shown as black vertical lines. Raw ChIP-chip signal is shown in grayscale while peaks are shown as red vertical lines. All 3 binding sites detected by STAGE overlap with ChIP-chip detected peaks. Lower panel shows overlap between STAGE detected binding sites with core-promoter ChIP-chip data. *P-value* was calculated using a hypergeometric distribution. Reproduced from [59].

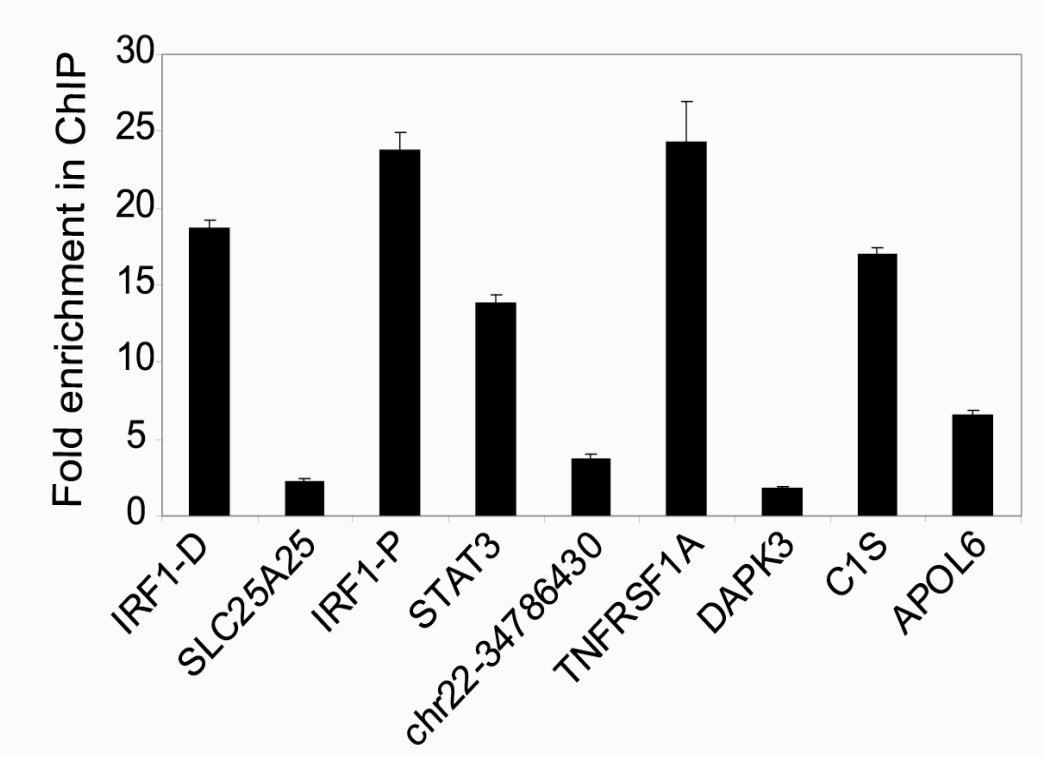


Figure 2.4 Quantitative PCR data for STAGE detected binding sites

Quantitative PCR was performed on an independent STAT1 ChIP sample for 10 randomly selected sites ranging in score from 0.95 to 1. Of the 10 sites chosen, 8 showed enrichment greater than 2 fold and 1 site showed an enrichment of more than 1.5 fold over background. Fold enrichment was calculated using the ΔΔCt method. Reproduced from [59].

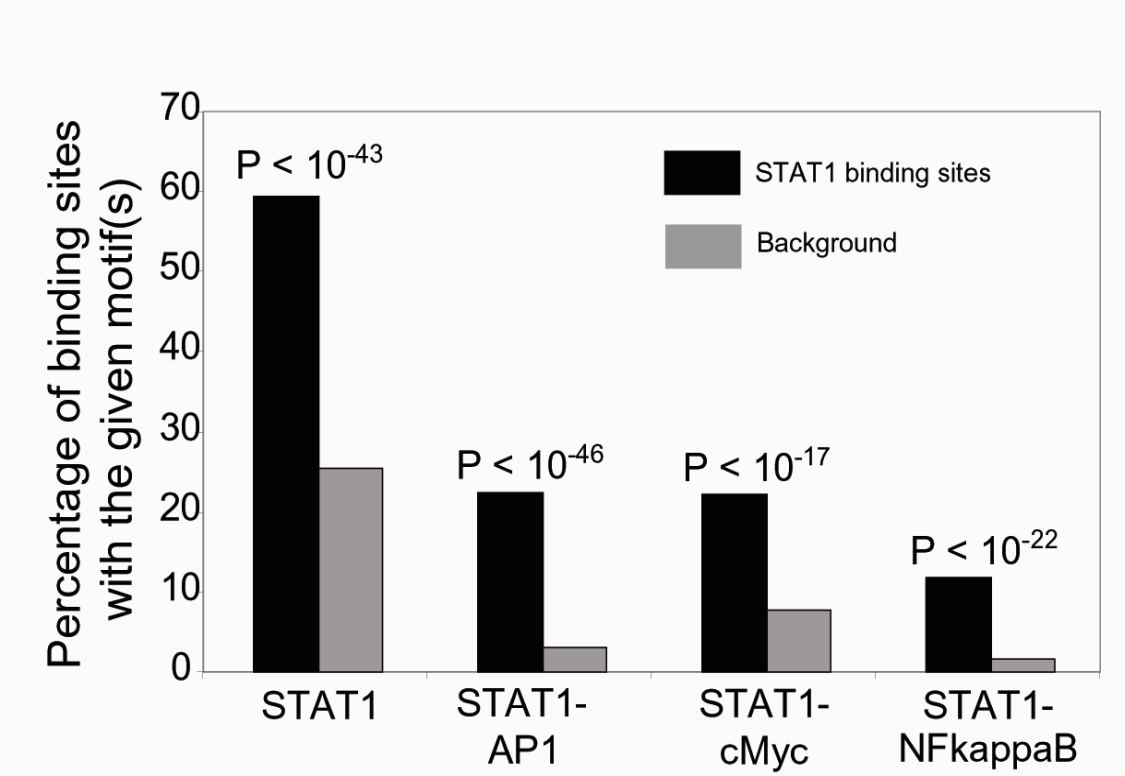


Figure 2.5 Motif analysis for STAT1 targets

Abundance of the STAT1 motif alone or co-occurrence of the STAT1 motif with motifs for other TFs such as AP1, Myc and NFkappaB have been shown as a percentage of the total number of binding sites. STAGE detected binding sites show a clear enrichment of the STAT1 motif over background. Additonally, the motifs for the TFs AP1, Myc and NFkappaB, which are known to regulate STAT1 targets, showed co-enrichment with the STAT1 motif. *P-values* (P) indicated in the chart were calculated using a binomial distribution. Reproduced from [59].

and NFKappaB [71], we found an enrichment for the STAT1 motif along with motifs for AP1, MYC and NFKappaB (Fig. 2.5).

Genes proximal to STAT1 binding sites

STAGE identified several STAT1 targets involved in interferon-gamma signaling that were unidentified before. One of the targets identified was death-associated protein kinase 3 (DAPK3), a positive regulator of programmed cell death. DAPK3 induces cell death by associating with the protein DAXX, a pro-apoptotic molecule. IFNG is known to increase DAPK3-DAXX interactions and this complex is required for the pro-apoptotic role of IFNG [72]. STAT1 regulation of DAPK3 may represent yet another way how STAT1 might mediate the apoptotic functions of IFNG signaling. Another possible mechanism of IFNG induced apoptosis could involve the protein APOL6, a protein that induces mitochondria-mediated apoptosis [73] that was identified as a STAT1 target by STAGE.

STAT3 is an anti-apoptotic protein and induces cell proliferation whereas STAT1 is pro-apoptotic and promotes growth arrest [74, 75]. In mouse embryonic fibroblasts, it has been observed that IFNG indices high levels of STAT1 expression while STAT3 levels remain low. However, in STAT1-/- cells, IFNG stimulation indices high levels of STAT3 expression [67]. Our data implicates STAT3 as a direct transcriptional target of STAT1 and suggests that STAT1 represses STAT3 during IFNG signaling further promoting its apoptotic action.

Tumor necrosis factor-α (TNF-α) is a cytokine that is involved in multiple cellular responses including cell differentiation, survival and apoptosis. TNF-α binds to it receptor TNFRS1A (Tumor Necrosis Factor Receptor Super Family 1A) or TNFR1 and causes NFKB activation, which is crucial for the expression of many pro-inflammatory

cytokines, chemokines and other regulators of apoptosis and cell differentiation. In the absence of IFNG stimulation, cytoplasmic STAT1 binds to TNFRS1A and maintains a tight control over TNF-mediated NFKB activation. However, increased sensitivity to TNF stimulation has been observed on IFNG stimulation [76]. STAGE detected a STAT1 binding site in the first intron of TNFR1 suggesting the possibility that IFNG dependent increased sensitivity to TNF-α could be a direct result of activation of TNFR1 by IFNG1-stimulated STAT1. All above mentioned target sites were verified by quantitative PCR performed on an independent ChIP sample (Fig. 2.4). We also identified other previously known STAT1 targets such as IRF1, HLA-E, ICAM1, as well as STAT1 itself, whose expression is known to be induced by IFNG. The complete list of STAT1 targets identified by STAGE is provided in Appendix C.

Identification of c-Myc targets by STAGE

We used STAGE to identify the targets of c-Myc, a known oncogenic transcription factor. Myc STAGE was carried out in HeLa cells and approximately 4500 clones from the c-Myc-STAGE library were sequenced by standard sequencing methodology. On average, each clone contained about 20 to 30 tags generating a total of 127, 351 tags out of which 19,867 were orphans that could not be mapped to the human genome. We used the remaining 107,484 tags for further analysis. Based on extrapolations from our ChIP-chip data (below) and previous observations, c-Myc is expected to have between 17,000 and 25,000 binding sites on the genome. Because our depth of sequencing of STAGE tags for c-Myc was slightly lower than for STAT1, and the possibility that c-Myc may have a larger number of binding sites on the genome, high specificity algorithm we developed for identifying STAT1 targets yielded a low number of binding targets for c-Myc. We therefore used a more relaxed algorithm as described in Methods to identify 2128 binding sites for c-Myc across the entire genome at a

probability threshold of 0.8. We estimated the false discovery rate based simulations at this threshold to be 5%. 26 of the c-Myc binding sites identified by STAGE occurred within the ENCODE region. We also identified c-Myc binding sites within the ENCODE regions by ChIP-chip using NimbleGen oligonucleotide tiling arrays [56].The ChIP-chip analysis included 3 biological replicates, and we defined c-Myc binding peaks in each replicate using the NimbleGen SignalMap software. 14 out of the 26 c-Myc binding sites within the ENCODE regions that were identified by STAGE were within 500 bp of a ChIP-chip peak in at least one of the three biological replicate experiments.

DISCUSSION

454 technology had several advantages for STAGE over standard sequencing approaches (Margulies et al. 2005). First, there was no requirement for cloning and isolation of independent recombinant clones. Rather, tags generated by the STAGE procedure could be directly sequenced. Second, the water-in-oil emulsion that was generated in making the library could be stored, and only a portion of this sample was used to generate on the order of 200,000 sequence reads in a single run of the instrument. Thus from a single chromatin immunoprecipitation reaction performed from a normally grown culture of mammalian cells, it was possible to sequence many samples, and together generate more than one million sequence reads amounting to more than 100 Mb of sequence using STAGE, greatly improving the depth of sequencing and coverage of targets enriched in the ChIP sample. Third, 454 technology was more cost-effective. In our experience, the price of sequencing a STAGE tag using 454 technology was about one-fifth that of standard clone based sequencing. We further improved STAGE by incorporating the tetratag technique described in a later section. However, even with these improvements, in depth coverage of TF binding sites across the human genome

would have been a lot more expensive with STAGE as compared to ChIP-sequencing using ultra high throughput technologies such as Solexa.

MATERIALS AND METHODS

STAT1 ChIP

STAT1 chip was performed in HeLa S3 cells by the Snyder lab at Yale as described in [68].

STAGE using 454 sequencing

STAGE procedure was carried out to generate ditags from STAT1 ChIP DNA as described above. Ditags were amplified using linker specific primers, gel extracted and directly sequenced by 454 inc. Duplicated ditags were removed and 21 bp tags were extracted by custom PERL scripts. Tags were mapped back to the May 2004 Build 35 human genome assembly as described in E2F4 STAGE.

Assigning probabilities for tag enrichment

As described earlier, the number of distinct positions in the genome to which a given tag could be mapped was termed **nhit**. For example, if a given tag could be mapped to 2 locations in the genome, the nhit for that tag would be 2. As described previously, we defined the number of times a given tag appeared in the STAGE sequencing pool as **nocc**. The selection of tags at random from the entire genome can be modeled as a binomial distribution where the success of an event is defined as selecting a tag with a given nhit, the probability of which is calculated as $p = nhit/T$, T being all possible 21 mers in the genome that start with CATG. For an observed tag with a given nhit and nocc = f, we can calculate the probability of selecting a tag with the observed nhit and nocc >= f under the null model. This probability was calculated as 1 – P(f-1) where P(f-1) is the

cumulative binomial probability of selecting that particular tag with a frequency <= f-1. P(f-1) was calculated as

$$1-\sum_{0}^{f-1}\binom{N}{x}p^{x}(1-p)^{N-x}$$

where, p is the background probability of selection of the tag and x iterates from 0 to f-1.

Multiplying this probability by the total number of tags with the given nhit in the genome yields the expected frequency of selecting tags with a given *nhit* and *nocc* >= f.

Thus under a null model, the expected frequency of a tag with a given *nhit* and *nocc* = *f* when N tags are selected at random from the genome was calculated as

$$\left(1-\sum_{0}^{f-1}\binom{N}{x}p^{x}(1-p)^{N-x}\right)M$$

where, p = *nhit*/T as described above, T = 27,429,149, and M = number of tags with a given *nhit* in the entire genome.

Probability that a given tag is enriched: p(tag) = $\left(1-\frac{Exp}{Obs}\right)$ where, *Exp* = expected frequency of the tag and *Obs* = observed frequency of the tag.

If the calculated expected frequency of the tag was greater than the observed frequency, the tag was assigned a low enrichment probability of 0.001.

STAGE target calls for STAT1

A window size of 500 bp was selected as described above to scan the genome and tags lying within each window were assigned to that window. For a given window, we defined k as the number of tags with *nhit* = 1 that were assigned to that window.
Probability that the given window is a target was calculated as:

$P = wt\_nhit*P_{hit} + wt\_nocc*P_{nocc}$ where,

P_{hit} = 1 - E(k)/O(k) where E(k) was expected frequency of windows with given k, O(k) was observed frequency of windows with given k.

E(k) was estimated by performing random simulations where N tags equal to the STAGE sequencing pool were randomly chosen from the genome 20 times and the distribution of E(k) calculated each time and finally averaged for each k.

P_{nocc} was calculated as the probability that at least one tag assigned to the window was not random:

$$P_{nocc} = 1 - \prod_i \left(1 - \frac{p(tag_i)}{nhit_i}\right)$$

where $p(tag_i)$ is the probability that tag_i was enriched and was calculated as described above. *wt_nhit* and *wt_nocc* were chose empirically and were set to 0.9 and 0.1, respectively.

Quantitative PCR (q-PCR) for confirming STAT1 binding sites

Quantitative PCR (q-PCR) was performed on an independent IFNG stimulated STAT1 ChIP sample. 10 predicted binding sites were selected at random spanning a range of STAGE probability scores. Each site was extended by 100 bp on either side and primers were designed to amplify a 60-100 bp region within the extended windows. Q-PCR was performed in triplicate in a 96-well optical plate (ABI PRISM) using SYBR green PCR Master Mix (Applied Biosystems) on an ABI 7900 instrument. For each locus, -ΔΔCt values were calculated for ChIP enriched DNA with respect to input DNA using *gapdh* promoter as a reference locus. Primer sequences have been provided in Appendix D.

IMPROVING STAGE

Since 454 technology can generate reads greater than 100 bp, sequencing ditags is inefficient as each read can now generate a maximum of two tags when it can

accommodate at least 4 tags of 21 bp each. Replacing the adapters at either ends of the ditags to generate "tetratags" would increase throughput by 200%. We used the strategy outlined in Fig. 2.6 to generate such tetratags. Ditags were digested with NlaIII and the 42 bp core ditags were gel extracted and self-ligated to generate a ladder of ditags i.e. multiples of 42 bp. The band corresponding to 84 bp i.e. tetratag band, was gel extracted and amplified by a modified ligation-mediated PCR (LM-PCR) procedure. Adapters with CATG overhangs and 3' EcoRI sites adjacent to the CATG overhangs were ligated to the gel extracted tetratags which were further amplified by a standard LM-PCR reaction using primers specific to the adapters. The adapters were then removed by digesting the amplified product by EcoRI and re-purifying the tetratags by gel extraction. However, it must be noted that the procedure required two gel extraction steps that reduced overall efficiency and hence required relatively larger amounts of starting material (at least 2ug). Additionally, the efficiency of the self-ligation step was dependent on the ratio of the amount of DNA to ligase used and it was moderately difficult to get the optimal ratio.

The tetratag version of STAGE was applied as a proof-of-principle to derive tags from STAT1 ChIP-enriched DNA. STAT1 STAGE derived ditags were digested, self-ligated and gel extracted to generate tetratags. STAT1 tetratags were cloned into TA vectors and colonies were screened by PCR to check for inserts. Plasmids with inserts were sequenced by standard Sanger sequencing to confirm tetratag identity. As expected, the sequenced tetratag had the EcoR1 recognition sequence on one end and 4 tags of length 21 nucleotides delineated by CATG sequences. However, the tetratag technology was not applied on a wider scale as newer sequencing methods made it possible to sequence ChIP-enriched DNA fragments directly without having to perform any of the tedious steps involved in STAGE.

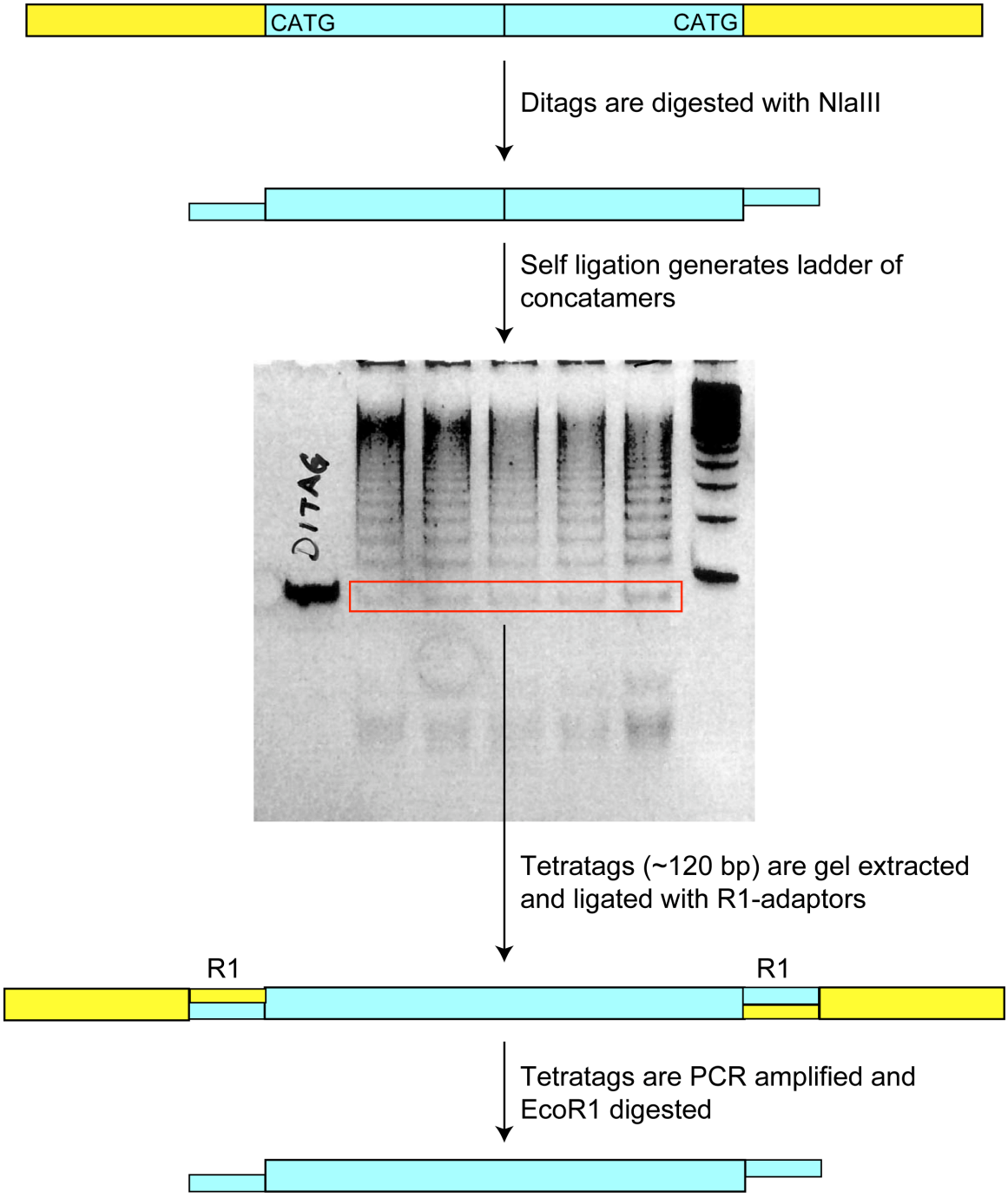


Figure 2.6 The tetratag strategy

Chapter 3: Mapping Genome-wide Transcription Factor Binding Sites Using ChIP-sequencing

INTRODUCTION

With the availability of ultra high throughput sequencing technologies such as Solexa and ABI SOLID, it was possible to generate genome-wide TF binding data with minimal manipulation of the ChIP-enriched DNA. Additionally, these technologies required small amounts of starting material (less than 50 ng of Chip enriched DNA) and were capable of generating several million short reads per run of sequencing. Solexa sequencing generates short reads from the ends of DNA fragments. DNA fragments to be sequenced are ligated to adapters and immobilized on a flow cell surface. Immobilized fragments are then subjected to solid phase bridge amplification to generate up to 1000 identical copies in close proximity to each other. Sequencing cycles are initiated by adding all four (A, C, G, T) labeled reversible terminators, primers and DNA polymerase. The emitted fluorescence due to laser excitation is captured and the first base recorded. This cycle is repeated several times to sequence short stretches from the ends of the DNA fragments. At the time of this study, the maximum possible length that could be sequenced with a minimal error rate was ~ 30 bp. Currently, its possible to generate 50-75 bp reads by Solexa technology. Solexa sequencing typically generates 2-4 million short reads per lane of the flow cell. We used the Solexa sequencing technology to map E2F4 binding sites in the human lymphoblastoid cell line, GM06990.

RESULTS AND DISCUSSION

Generating the raw data

E2F4 ChIP was performed by Bum Kyu Lee, a graduate student in the Iyer lab and ChIP-enriched DNA was subjected to Solexa sequencing. Two lanes of the flow cell

were used for the sequencing reaction to generate a total of 11.5 million short reads. Reads were mapped to the hg17 version of the human genome (May 2006 Build 36.1) using the alignment software Eland and only reads mapping to an unique location in the genome were retained. After applying this filter, 6.5 million reads were retained and used for peak calling (~2.5 million from lane1 and ~4 million from lane2). We also sequenced input DNA (Input) in a similar fashion and obtained 8.5 million mapped reads.

Identifying peaks in the sequencing data

Solexa sequencing technology generated 25-35 nt short reads from the ends of ChIP-enriched DNA fragments. These short reads were mapped back to genome using the ELAND algorithm. We obtained 6,508,011 reads from the E2F4 chip library and 8,474,489 reads from the input library. We applied a parzen window based algorithm to precisely define binding sites from short read sequencing data. The algorithm scores each base pair according to the distribution of reads around that base pair. Higher the score, greater is the likelihood that the given base pair is the binding site. The genome is thus converted into likelihood landscape of binding where local maximas define binding events (Fig. 3.1, Fig. 3.3A). Peak detection was initially performed on the plus and minus strands separately. Each read assigns a score to its neighboring nucleotide as a function of the read's distance from that nucleotide. The function used to assign scores is a Gaussian kernel with a defined standard deviation also termed a band-width. Local maxima on the plus and minus strands are defined as peaks. High scoring plus peaks that are upstream and within 500 bp of a minus strand peaks are considered paired and the distance between the paired plus and the minus peaks is calculated as the fragment length. The distributions of such fragment lengths for various datasets have been shown in Fig. 3.2. Interestingly, different datasets show significantly different distributions. These distributions do not seem to be factor dependent as can be seen by comparing data for

SRF ChIP-Seq performed in Jurkat cells or fibroblasts (Fig. 3.2C & 3.2D). The variations in fragment lengths observed across datasets may reflect the actual size of the DNA fragments that were gel extracted prior to sequencing. In order to make the estimation robust to outliers, we ignored the top and bottom 25-percentile and calculated the median from the remaining data. Half of the median fragment length was taken as the distance by which each read was to be shifted. Plus reads were shifted in the left to right direction while minus reads were shifted in the right to left direction. Re-positioned plus and minus reads were considered together and the peak-finding performed again to find local maximas across the genome thereby defining binding sites.

Peak detection was performed using reads derived from the E2F4 ChIP-enriched DNA and Input. Peaks in the Input dataset represent DNA fragments that are sequenced irrespective of any enrichment. This may happen due to effect of local chromatin structure, shearing bias or the presence of extensive repeat regions in the genome (Fig. 3.3B). In order to filter out such areas, E2F4 peaks were input corrected. If a E2F4 peak was within 500 bp of a Input peak, the given E2F4 peak score was divided by the input peak score. If a given E2F4 peak overlapped with more than one input peak, the higher scoring input peak was used for the correction. However, to leave promoter peaks intact, E2F4 peaks mapping to within 10,000 bp from any TSS of a gene were not input corrected.

To assign significance to E2F4 peaks, we performed random simulations to estimate the false discovery rate (FDR) of our procedure (Fig. 3.4A). We selected 6.5 million coordinates across the genome and peak finding was performed as described above. This process was repeated 20 times and the average number of random peaks detected at various score thresholds was compared to the number of peaks detected in the E2F4 dataset to give the false discovery rate. At a score threshold of 4.4, the estimated

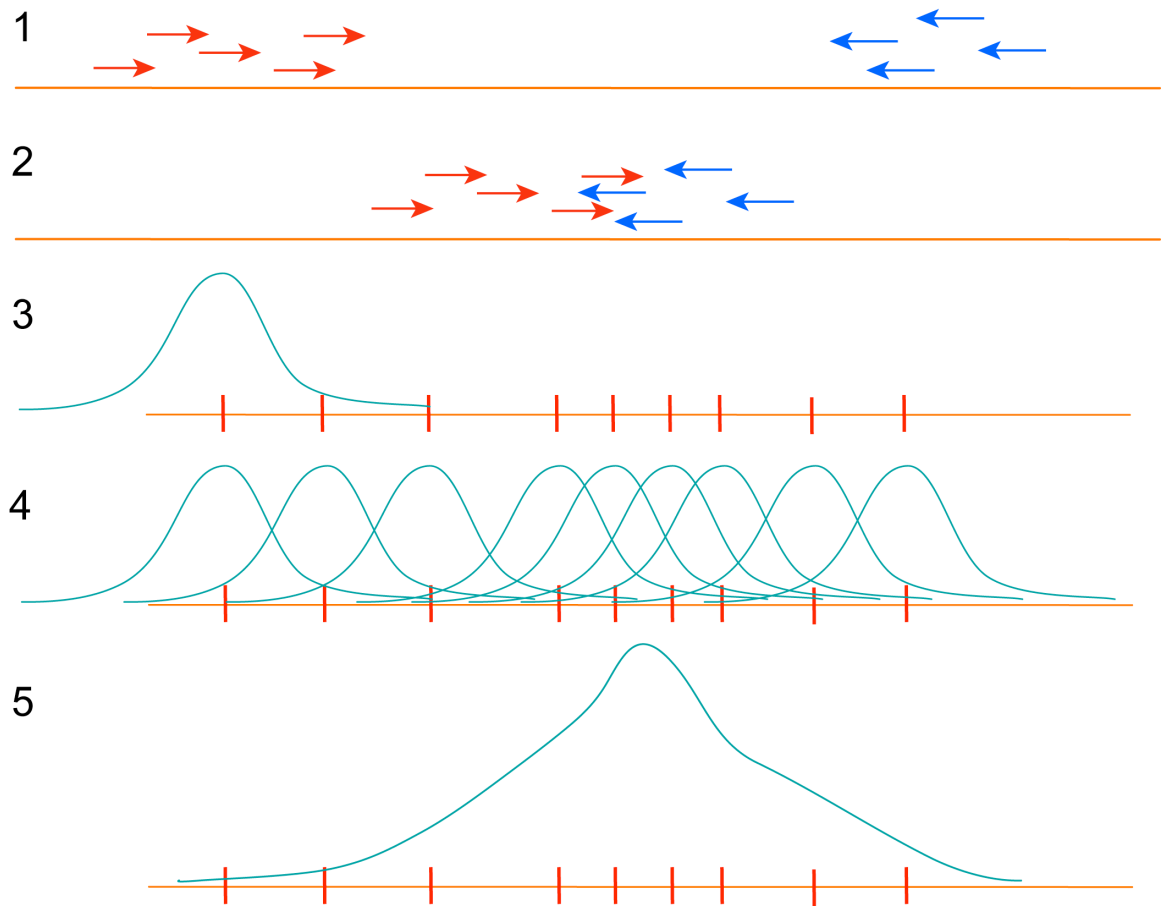


Figure 3.1 Peak detection algorithm

1) Reads mapped to the plus strand are shown in red while reads mapped to the minus strand are shown in blue. 2) Reads are shifted by a distance equal to half the fragment length. Plus strand reads are shifted towards the right while minus strand reads are shifted towards the left. 3) A Gaussian kernel is centered on the start coordinate of each read and base pairs in the neighborhood of that read are assigned a score that is a function of the nocc of that read and the distance of the base pair from that read. The farther the base pair from the read, the lower is the score assigned to it. The score is calculated in accordance to the Gaussian kernel. The neighborhood is defined as 4 times the standard deviation of the kernel on either side. 4) The process is repeated at every read. 5) If a base pair is assigned scores from several reads, the individual scores are added. This generates a likelihood landscape of TF binding across the genome where local maximas define the base most likely to be the binding site.

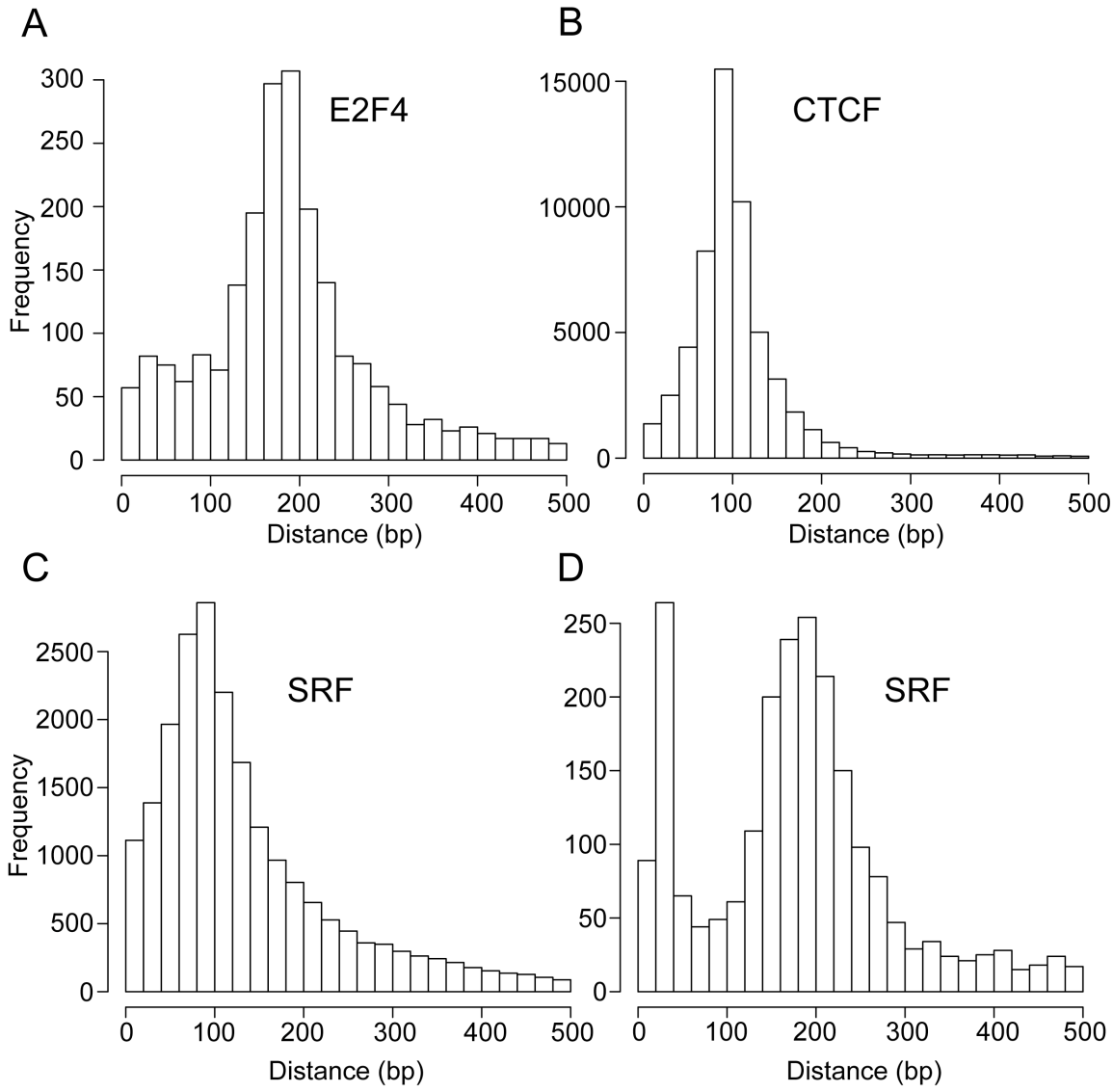


Figure 3.2 Distribution of plus to minus peak distances

Peak finding was performed on the plus and minus strand separately. A plus peak followed by a minus peak within 500 bp constituted a plus-minus peak pair. Histogram shows the distribution of the distances between the plus and minus peaks within the pair. The top and bottom 25-percentile values were dropped and the median distance was estimated from the remaining data. Reads were extended by half this distance. A) E2F4 data in GM06990 cells (ChIP performed by Bum Kyu Lee) . B) CTCF data in GM12878 lymphoblastoid cell line (ChIP performed by Bum Kyu Lee as part of the ENCODE project). C) SRF in Jurkat cells. D) SRF in fibroblasts (ChIP performed by Bum Kyu Lee).

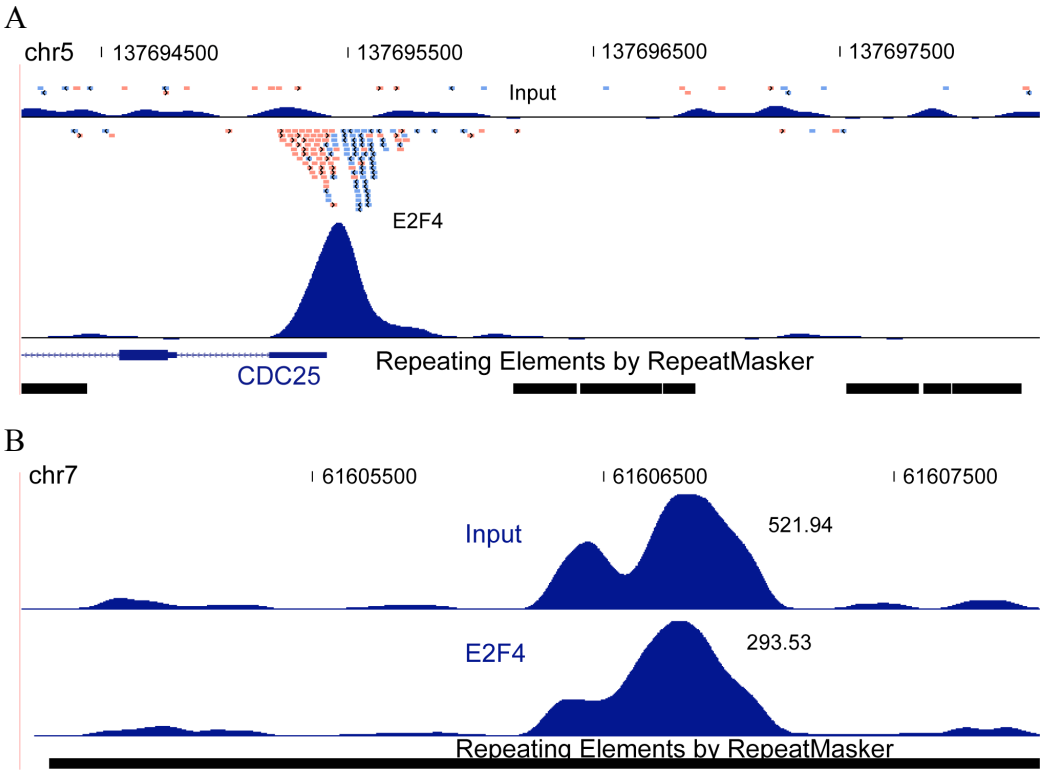
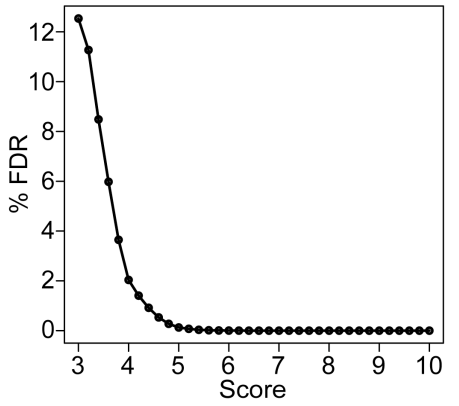
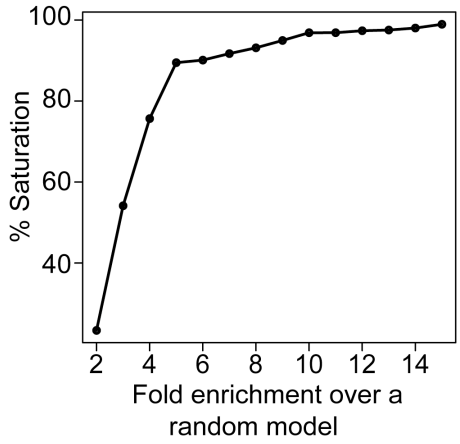


Figure 3.3 Peaks identified from sequencing data

(A) Detailed view of the CDC25C promoter is shown. Our analysis detects a strong peak in the E2F4 dataset while no corresponding peak is detected in the input. (B) A genomic locus on chromosome 7 showing a strong peak detected in the input region corresponding to a E2F4 peak. Peaks in the input region commonly map to repeat regions in the genome shown by the black bar.



(A) Estimating the false discovery rate (FDR). At a score of 4.4, the FDR was less than 0.1 % and corresponded to ~16,000 E2F4 peaks genome-wide.



(B) Saturation analysis
Percentage coverage of possible E2F4 binding sites across the genome has been plotted at different enrichments. Enrichments were calculated with respect to a null background. Total number of possible E2F4 binding sites across the genome at a given enrichment was estimated using a capture-recapture approach.

Figure 3.4 Estimating the quality and coverage of the sequencing results

FDR was < 1% and generated ~16,000 E2F4 peaks across the genome. For further validation, Bum Kyu selected 42 peaks and assayed these putative binding sites by quantitative PCR (qPCR). Out of these 42 peaks, 30 had a score greater than 4.4 and 12 had a score between 3 and 4.4. At a 1.5 fold enrichment cut-off for the qPCR, 90% of the 30 high scoring peaks (score greater than 4.4) were validated while only 41% of the low scoring peaks (score between 3 and 4.4) could be validated as true binding sites. This data showed that there were true binding sites amongst peaks with scores less than the cut-off of 4.4 but the false discovery rate was much higher. Hence, we chose all peaks with a score greater than or equal to 4.4 for further analysis.

Estimating binding site coverage

An important question in ChIP-sequencing data analysis is what is the percentage of binding sites that have been recovered for a given factor at a given depth of sequencing i.e. at what point does the sequencing reach saturation levels? To answer this question we used a capture-recapture approach. Capture-recapture methods have been used to estimate the population of animals in a given area [77]. The area under study is assumed to be closed such that no animals leave or migrate into the area. A random subset of animals is caught, tagged and released back into the population. The tagged animals are allowed to mix homogenously within the untagged population under the assumption that no tags are lost. A second capture is now performed and the number of tagged animals that are captured in the second round (recapture) is noted. The population of the area under study is given by:

(N1 x N2)/k where N1 and N2 represent animals captured in the first and second round of capture and k represents the number of animals recaptured i.e. the overlap between N1 and N2.

The sequenced reads from the E2F4 chip library were obtained in two sets or "lanes", the first set having 2,305,280 reads and the second set having 4,202,731 reads. To apply the capture-recapture method to estimate the total number of E2F4 binding sites across the genome, we binned the genome into 500 bp bins and counted reads that mapped to a given bin. This process was performed using the entire E2F4 library of reads and for the two lanes separately. Each lane was treated as a separate capture and the bins were considered as animals so that the total number of bins could now be calculated by the formula given above where N1 and N2 would be the number of bins in lanes 1 and 2 respectively while k is the overlap between the two sets. A drawback with the capture-recapture model is that it assumes every animal (or bins in our case) has an equal probability of capture. This is obviously not true as binding sites that are highly enriched in the ChIP sample will generate significantly more reads than background and hence will be "captured" at higher probability than background bins. Additionally, we also had to take into account the false positive bins as they could artificially increase or decrease the population estimate depending upon whether they are dominant in the numerator or the denominator. To address the issue of unequal probability of capture, we thresholded bins based on fold enrichment. Fold enrichment was calculated, as the ratio of the number of reads within a given bin divided by the average number of reads expected in a bin if occupied. All bins above a given threshold were considered equally likely to be selected by the sequencing process. This is not a perfect solution but merely an approximation. To address the issue of false positives, each bin was assigned a *P-value* based on a Poisson distribution to evaluate its significance. At a given threshold of fold enrichment (i.e. for a given minimum number of reads per bin), the corresponding p-value is calculated using the Poisson distribution. The expected number of false positive peaks is simply equal to the p-value multiplied by the total number of bins in the genome. This number is

subtracted from N1 and N2 and also from the denominator as shown in the following equation:

$$\frac{(N1 - fp1) * (N2 - fp2)}{\min\left[k(1 - \frac{fp1}{N1}), k(1 - \frac{fp2}{N2})\right]}$$

where,

N1: Number of bins in set 1 above a given fold enrichment

N2: Number of bins in set 2 above a given fold enrichment

k: Number of bins common to set 1 and set 2

fp1: Expected number of bins in set 1 according to a random Poisson model

fp2: Expected number of bins in set 2 according to a random poison model

fp1 and fp2 represent the expected number of false positives at each enrichment cut-off.

The false positive rate for each lane is subtracted from the overlap (k) in the denominator of the above equation and the minimum of the two values is used. This is so because, for example, if we are comparing 5 objects with 10 objects, then the maximum overlap possible between the two sets is 5. The expected total number of bins is calculated at each fold enrichment and the percentage saturation at each enrichment cut-off was calculated as the ratio of the observed number of bins in the entire E2F4 chip library (i.e. 6,508,011 reads) and the estimated number of bins at that enrichment from the capture-recapture approach. Estimation of the coverage at different thresholds generates a saturation curve that is shown in Fig. 3.4B. From the saturation curve, it can be seen that at a fold enrichment of 5 or more, we have recovered more than 90% of E2F4 binding sites in the genome while sites that are poorly enriched show much less coverage. Increasing the depth of sequencing should retrieve more of the poorly enriched binding sites. However, it must be noted that the background also increases with

increased sequencing depth and some binding sites that are very weakly enriched may not be reach complete coverage ever.

E2F4 binding characteristics across the human genome

It has been reported previously that many transcription factor binding sites tend to cluster close to the TSSs of genes indicating a strong preference for the core promoter [78, 79]. We binned the genomic region 10 kb upstream and downstream from the TSS of all genes into 50 bp size bins and mapped E2F4 peaks to these bins. We first wanted to investigate the relative distribution of all E2F4 peaks detected by our analysis. If the local maxima of a peak fell within a bin, the peak was said to map to that bin. We counted the number of peaks that mapped to each bin and then calculated the percentage of peaks mapped within each bin with respect to all E2F4 peaks within the 20 kb window centered on the TSS. The same procedure was used to generate transcription termination site (TTS) profiles. We found that majority of E2F4 peaks map within 1 kb of the TSS (Fig. 3.5A & 3.5B) while no such distribution was observed with respect to the TTS. We repeated the analysis with peak scores instead of peak numbers. If a peak mapped to a bin, the score of that peak was assigned to that bin. In case multiple peaks mapped to the same bin, the maximum score amongst the mapped peaks was assigned to the bin. All corresponding bins across all genes were averaged and smoothed by a moving window of 3 to generate an average peak score profile across TSSs. The same procedure was used to generate TTS profiles. As seen in Fig. 3.5C & 3.5D, the profile shows a sharp peak between -500 and +500 bp with the maxima positioned slightly upstream of the TSS. This indicates that E2F4 has a strong propensity to bind near TSS i.e. in the core promoter of a gene. This data agrees with previous analysis performed on tiling arrays covering 1% of the genome [64].

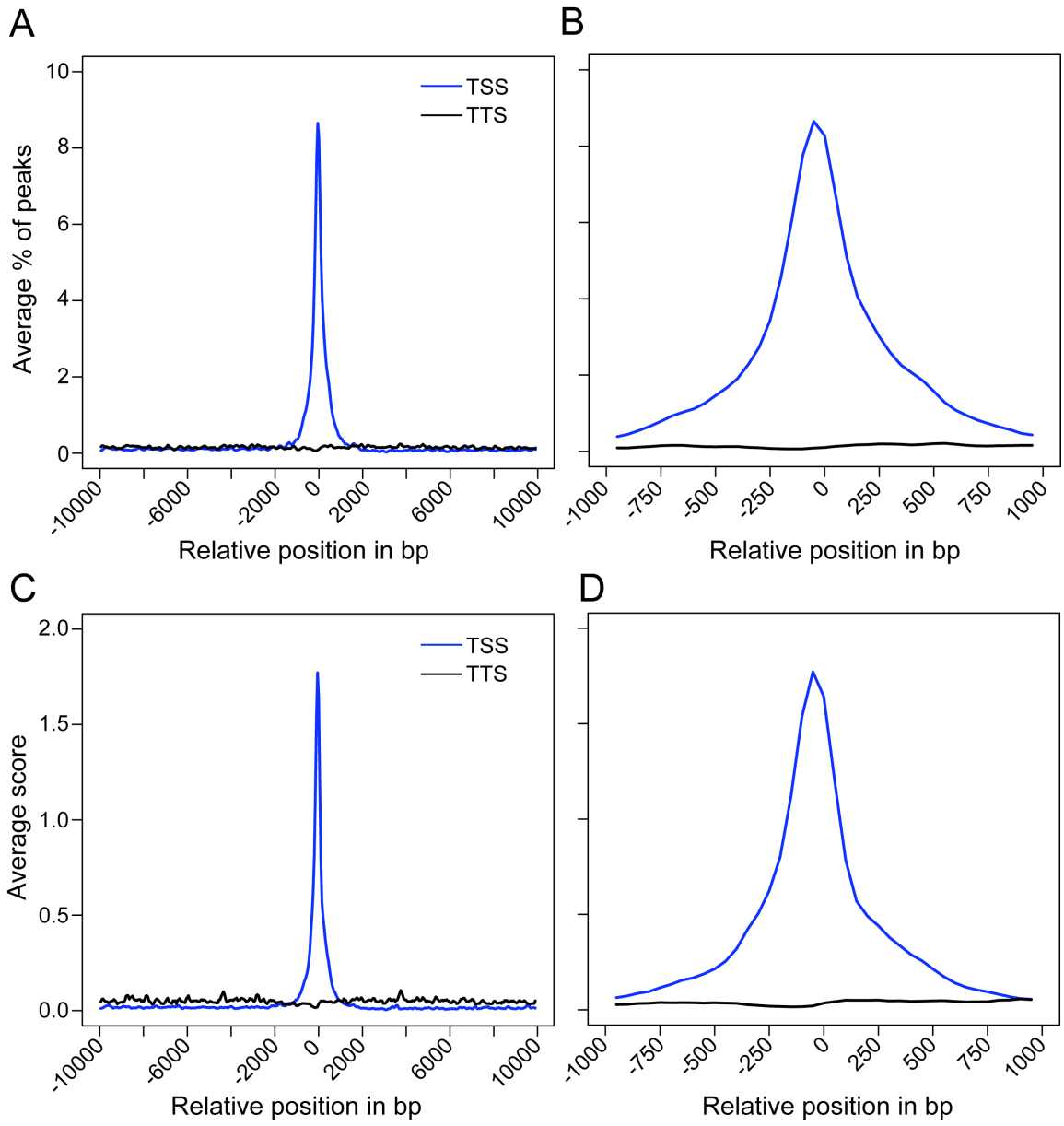


Figure 3.5 Distribution of E2F4 peaks at the TSS

(A) E2F4 peak distribution around the TSS (blue line) and TTS (black line) of all genes. (B) Close-up of the region between -1 kb to + 1kb in (A). (C) E2F4 peak score distribution around the TSS (blue line) and TTS (black line). (D) Close-up of the region between -1 kb to + 1kb in (C).

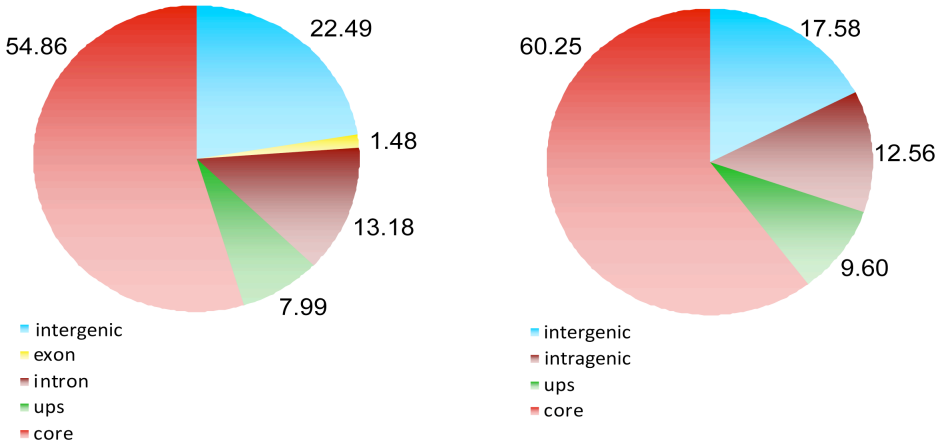


Figure 3.6 Distribution of E2F4 peaks in the genome

Left panel: E2F4 peak distribution in genomic regions defined using the RefSeq database. Right panel: Peak distribution in genomic regions defined using the expanded TSS dataset. This dataset did not have detailed information about introns and exons. Hence, instead of introns and exons, we used "intragenic" to define peaks that mapped between the transcription start and stop of genes.

Next, we divided the genome into 5 regions, namely 1) core promoter i.e. within 2kb upstream or downstream of any TSS, 2) upstream i.e. greater than 2 kb but less than 20 kb upstream from any TSS 3) exon 4) intron 5) intergenic i.e. greater than 20 kb upstream from the TSS of any gene. We mapped E2F4 peaks to these 5 regions such as each peak mapped to one and only one region. In case of a conflict, we set a hierarchy amongst the regions such core promoter > upstream > exon > intron > intergenic. For example, if a peak mapped within the core promoter of a gene on the plus strand but was within 20 kb of the TSS of a gene on the minus strand, that peak was assigned to the core promoter region. Of the total number of E2F4 peaks used in this analysis, ~55% mapped to core promoter regions, reconfirming our previous observation. Interestingly, a significant proportion of peaks (22.5%) could be mapped to intergenic regions even though it was the lowest on the hierarchy (Fig. 3.6 left panel).

To ensure that this number was not due to missing annotations, we used an expanded list of 60,000 TSSs that included predicted transcripts as well. Using the expanded TSS list the percentage of peaks mapping to intergenic region dropped to ~18% (Fig. 3.6 right panel). This indicated that a substantial portion of E2F4 binding sites occur far away from known TSSs. Long-range regulation of gene transcription by enhancers has important physiological roles in eukaryotes, especially mammalian cells and it may be possible that E2F4 can regulate gene expression by binding to enhancer sequences. On the other side these binding sites may also represent yet to be discovered gene TSSs. We also found that the number of E2F4 binding sites correlate very well with gene abundance on a per chromosome basis.

It has been previously observed that E2F4 binding motifs are over-represented in bidirectional promoters i.e. same promoter regions driving expression of divergent genes. However, it is not known whether E2F4 binding sites behave the same. We defined a bidirectional promoter as the region of the genome between two TSSs that are separated by less than 2 kb with the corresponding genes on opposite strands. Additionally, we required that the transcripts of these genes should not overlap. Based on these criteria, we identified 918 promoters corresponding to 1836 genes (~10%) amongst the 18,693 genes annotated in RefSeq. We mapped E2F4 peaks to these bidirectional promoters and found 572 genes to be regulated by E2F4. This represented a significant enrichment over background under a hypergeometric model (*P-value* < 2×10^{-20}).

Motif analysis

Previous structural data and in vitro SELEX analysis has revealed that the sequence TTTSSCGC where S stands for a G or C is a strong binding motif for E2F4 [80]. However, we found that only 5% of all our E2F4 binding sites had this canonical motif. This was in agreement with a previous E2F4 ChIP-chip study [64] and suggested

that there might be other significant motifs that are preferred by E2F4 to select binding sites. We extracted a 200 bp genomic sequence centered on each peak and performed de novo motif discovery using the freely available software Discovery of Ranked Imbalance Motif or DRIM [81]. Since, DRIM performs an exhaustive search based on enumeration of all possible motifs within criteria set by the user, it proved to be computationally expensive to perform the search using 16,000 sequences. Hence, we divided the sequences into strong, moderate and weak binding categories based on the peak scores such that each category had only 500 sequences and performed the motif discovery on these subsets. Motifs discovered in each subset were pooled together and duplicate motifs were removed to generate a final set of 5 motifs with *P-values* less than 10^{-4}.

Since these motifs were derived from a subset of sequences, we analyzed the prevalence and enrichment of these motifs in all 16,000 E2F4 binding sites. Fig. 3.7 shows the data for the analysis of all 5 newly discovered motifs as well as the canonical motif. All motifs showed enrichment in the entire set of E2F4 binding sites with the enrichment increasing with increasing binding strength. However, there were some subtle differences between the motif enrichments. The canonical motif (Fig. 3.7, Motif 1) showed the strongest enrichment reaching a plateau at a score of 15 but showed one of the lowest abundance in the binding sites. Enrichment for motif 2, on the other hand, continues to increase with increasing binding strength. This suggests that the canonical motif may not be the strongest preferred motif and may be used only in binding sites with moderate affinity. Motifs 3,4 and motifs 5,6 showed moderate and weak enrichments respectively. Enrichment for motif 4 reaches a peak at score 10 and then gradually decreases. This apparent reduction in enrichment is due to the fact that the motif is observed infrequently in E2F4 peaks above the score of 10 suggesting that different motifs may specify different binding affinities.

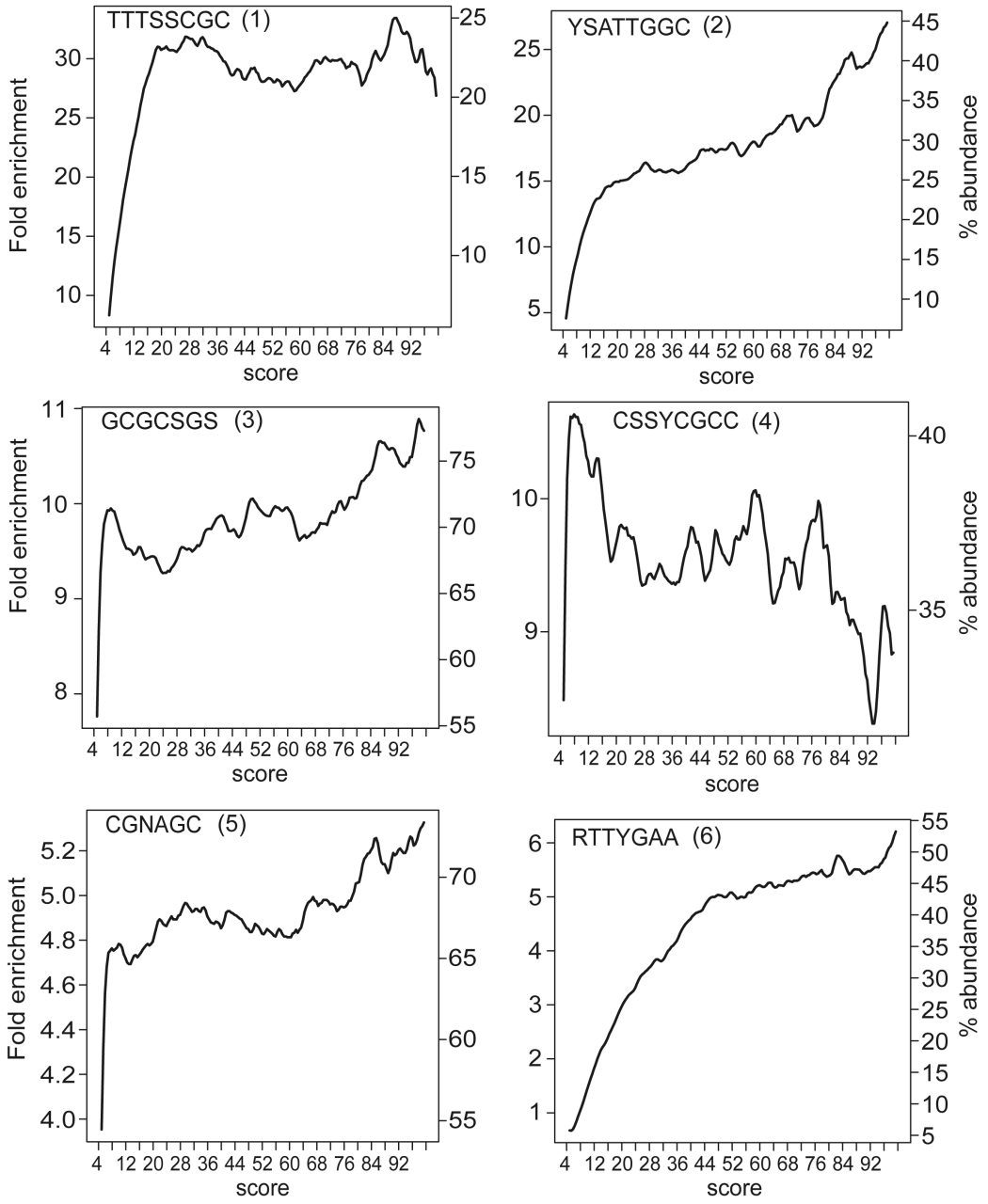


Figure 3.7 Motif analysis

For each of the 5 *de novo* discovered motifs and the canonical motif (motif 1), fold enrichment and the percentage abundance of the motif has been plotted at different peak score cut-offs. For a given panel, the left axis indicates the fold enrichment over background while the right axis indicates percentage abundance of the motif. The motif sequence is given at the top of each panel with the motif number in brackets.

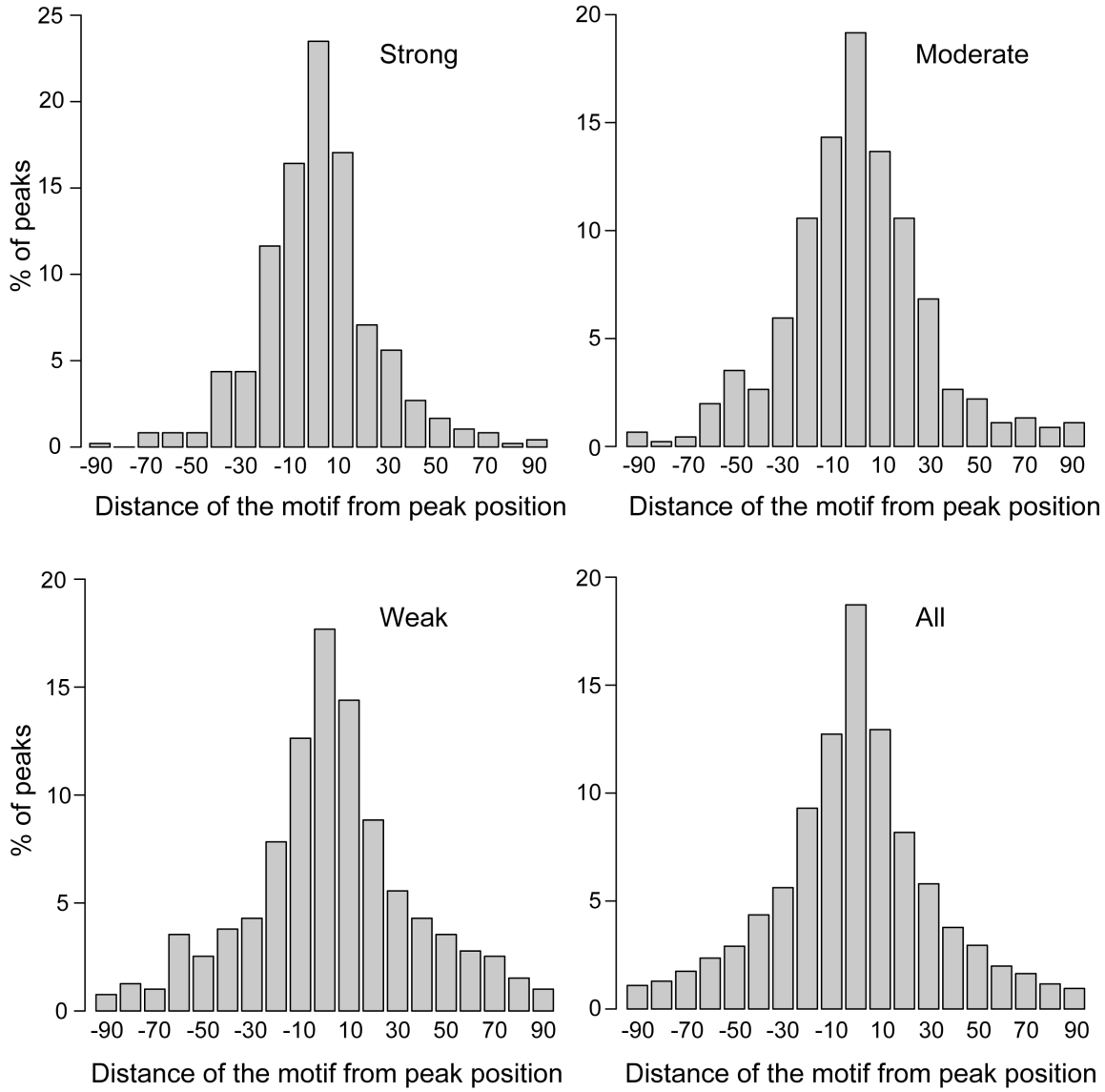


Figure 3.8 Distribution of motifs within peaks

Distances of the local maximas to the nearest mapped motif within a given peak is plotted as a histogram. Y-axis shows the percentage of peaks with a E2F4 motif within the specified distance from the local maxima for the strong, moderate and weak sets of E2F4 peaks (see text) and for all E2F4 peaks identified in this study.

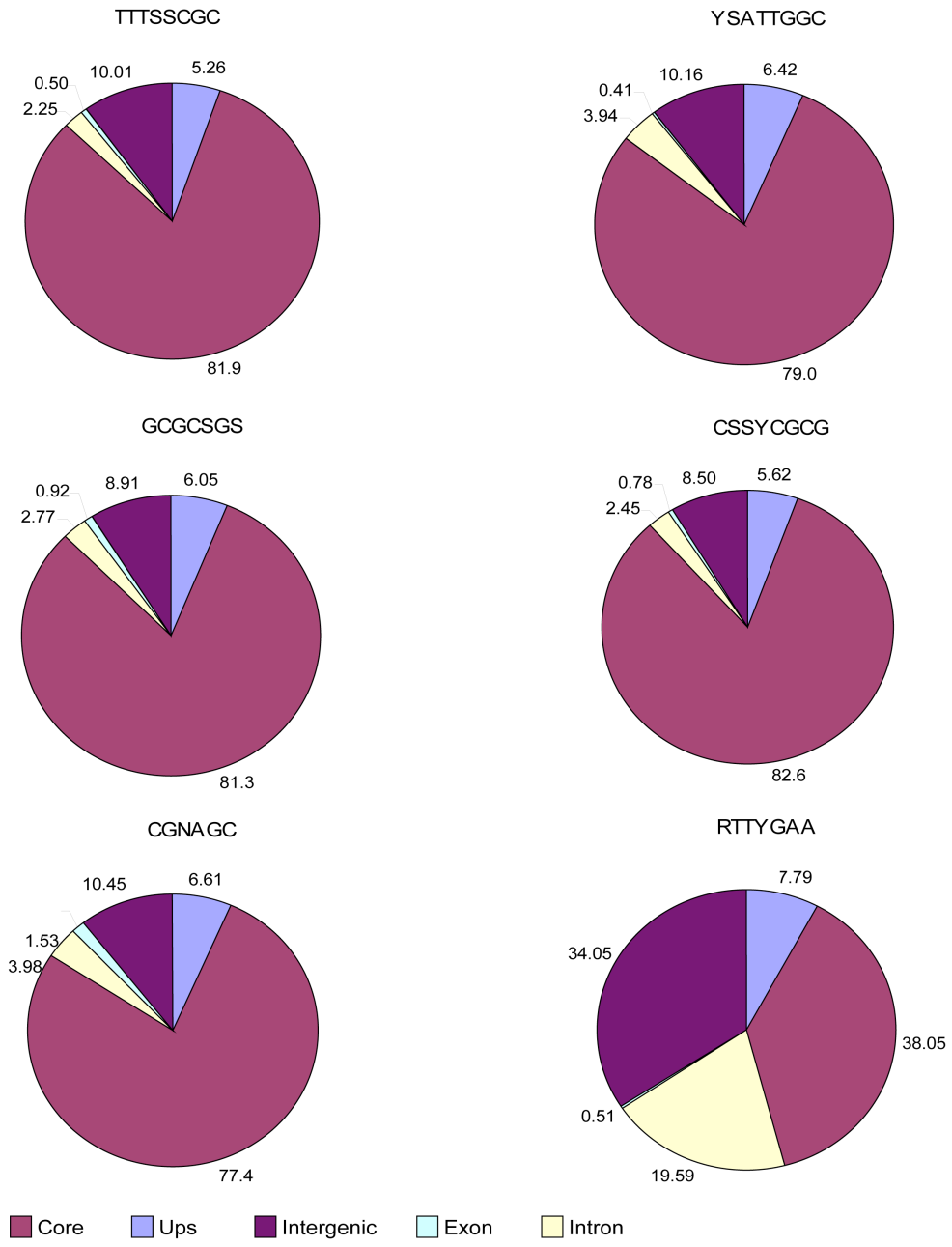


Figure 3.9 Distribution of E2F4 peaks within genomic regions

The motif under study is indicated at the top of each pie chart. Genomic regions have been color coded as indicated.

We wanted to investigate how these motifs were distributed within peaks detected by our analysis. We took the set of strong, moderate and weak sets of E2F4 peaks and estimated the distance of the local maximas within a peak to the nearest E2F4 motif and plotted the distances as a histogram. We repeated this analysis for all E2F4 peaks identified by our analysis (i.e. score >= 4.4). As shown in Fig. 3.8, E2F4 motifs were found within 20-30 bp of the maximas for almost 80% of the peaks in all sets. This indicated that we were able to detect E2F4 binding events with high precision. In order to investigate whether different motifs are preferred at different genomic locations, we estimated the occurrence of each motif in the above defined genomic regions i.e. core promoter, upstream, exon, intron and intergenic (Fig. 3.9). All motifs except RTTYGAA (Motif 6) were highly represented in the core promoter regions. Motif 6 on the other hand showed equal representation in the core promoter and intergenic regions as well as significant representation in the introns. This is in contrast to other motifs that showed minimal occurrences in the introns. Whether motif 6 containing binding sites form a special functional subset of E2F4 targets remains to be seen.

Transcription factors are known to display co-operative binding where binding of a single protein molecule to the binding site enables or recruits binding of additional molecules strengthening the regulatory potential. If our discovered E2F4 motifs are truly functional, we should expect to find multiple occurrences of these motifs in E2F4 peaks as compared to background. We selected E2F4 peaks that had at least one of the 6 E2F4 motifs and counted the total number of occurrences of these motifs in a given peak. The same analysis was performed on a random set of genomic sequences. As shown in Fig. 3.10, the number of motifs occurring in E2F4 peaks is significantly more than that found in the random dataset. Additionally, the number of motifs found per peaks correlated with

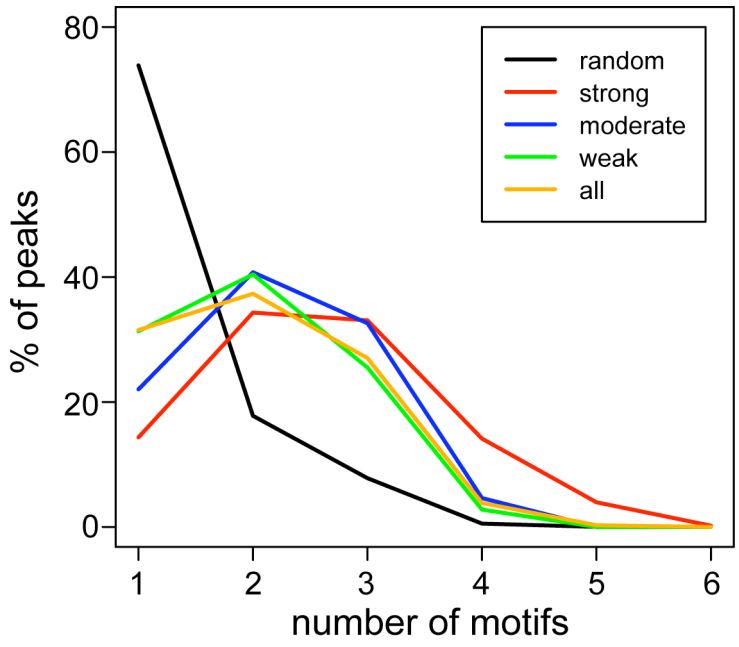


Figure 3.10 Frequency of motif usage in peaks

Only peaks that had at least one of the 6 E2F4 motifs were considered for this analysis. Motif frequency was calculated for the strong, moderate, and weak set of E2F4 peaks (see text) as well as for all E2F4 peaks above the cut-off of 4.4 and compared to a set of randomly generated peaks.

KEGG Pathway Term	Motifs
Cell cycle	1, 2, 3, 4, 5, 6
Ubiquitin mediated proteolysis	3, 4, 5
Pyrimidine metabolism	1, 3, 5
DNA polymerase	1, 3, 5
p53 signaling pathway	3, 5
Chronic myeloid leukemia	4
N-Glycan biosynthesis	3
Biosynthesis of steroids	2

Table 3.1 Associating E2F4 motifs with functional pathways

the strength of the binding i.e. the peaks in the strong binding set showed a higher occurrence of motifs than the peaks in the weak set.

Next, we investigated whether the discovered E2F4 motifs have specific functional significance. We took the set of genes that showed E2F4 binding within their core promoter regions and identified E2F4 motifs within these binding sites. We then performed functional pathway analysis on this set of genes (Table 3.1). If a given gene was assigned to a definite functional pathway, for example cell cycle, then all motifs associated with that gene were associated with the cell cycle pathway. A given motifs was allowed to be in multiple pathways. All E2F4 motifs used in this analysis showed an association with the regulation of cell cycle. The E2F4 motif RYTTGAA (Motif 6) was strongly enriched only in the cell cycle pathway associated genes. It must be noted that a very similar motif described previously was shown to function as a cell cycle repressor element and shown to recruit E2F4/RB complex [82, 83]. Some pathways were significantly enriched for a single motif. For example, motif 2 was significantly enriched only in the set of genes associated with steroid biosynthesis while motif 3 was strongly associated with the N-glycan biosynthetic pathway. These results suggest that different

E2F4 motifs may confer functional specificity by targeting E2F4 to pre-defined sets of genes.

Many transcription factors are known to work in concert or in a competitive manner to regulate gene expression. We wanted to investigate if there are other TFs that associate with E2F4 binding sites. We extracted 1000 bp sequences centered on the E2F4 peaks and identified motifs of other TFs that were statistically enriched around E2F4 binding sites as compared to genomic sequences randomly sampled from the genome. Several transcription factor motifs were found to be significantly associated with the E2F4 binding (Table 3.2) and some of these were previously validated associations. One such example is the constitutively expressed factor NF-Y that binds to several E2F4

AHRARNT	AP2	AP4	ARNT	ATF	ATF6	CREBP1
E2F	EGR1	EGR2	EGR3	ELK1	ER	MAZR
MAX	NF1	NF-Y	NRSF	P53	PAX2	PAX5
RFX1	SP1	SREBP1	STAT1	STAT3	TAXCREB	USF
XBP1	YY1					

Table 3.2 Conserved transcription factor motifs enriched within 500 bp of E2F4 peaks.

regulated cell cycle genes. Binding of NF-Y to the promoters of these genes results in recruitment of other activator proteins such PCAF and p300 that in turn activate the downstream gene. On repression by E2F4, NF-Y is displaced from these promoters implying opposing modes of regulation for E2F4 and NF-Y [84].

MicroRNAs regulated by E2F4

MicroRNAs (miRNAs) are newly discovered short non-coding RNAs that regulate gene expression post-transcriptionally by either translational repression or mRNA degradation [47]. It has been shown that many TFs and miRNAs are involved in

complex networks involving reciprocal regulation [53-55]. MicroRNAs are transcribed as long transcripts called pre-miRNAs and further processed into the mature duplex form [42]. However, the actual TSS of the pre-miRNA may be several kilobases away from the mature sequence [42]. We mapped E2F4 binding sites within 10 kb upstream of the start of the mature sequences of all known human miRNAs annotated in miRbase and identified several miRNAs as putative targets. The main problem with identifying miRNA targets in such a manner is that many miRNAs are intronic and are transcribed from the TSS of the host gene. Hence, for such miRNAs, assigning a TSS-proximal binding site unambiguously to the miRNA is difficult as the binding event may be regulating the host gene and not the miRNA. To avoid such issues, we first focused on miRNAs that mapped to intergenic regions and hence could be identified unambiguously. We detected strong E2F4 peaks upstream of the miR-17-92 and the let-7 clusters. The miR-17-92 cluster has been shown to promote cell proliferation and translationally represses E2F1 [50, 85]. E2F4 mediated regulation of miR-17 suggest a novel mode of regulation within E2F family members. The miRNA let-7 inhibits Myc protein expression [86] while E2F4 has been shown to be involved in the transcriptional repression of Myc in response to TGFβ receptor stimulation [87]. Accordingly, we also found a strong E2F4 binding signal (score 15.6) less than 70 bp from the Myc TSS. E2F4 mediated regulation of let-7a adds another layer of complexity to the interplay between the two oncofactors. Next, we included intra-genic miRNAs that shared their TSS with their host genes for this analysis. The miRNA miR-22 showed a strong peak less than 2 kb upstream from the TSS of its host gene C17orf91. As described in chapter 4, we found that miR-22 targets cell cycle arrest and apoptotic genes, thus facilitating cell growth. Thus we found E2F4 binding sites upstream of cell cycle promoting miRNAs i.e. miR-

17-92 cluster and miR-22. Since E2F4 has been described mainly as a repressor of the cell cycle, it would be interesting to investigate whether E2F4 represses these miRNAs.

Functional pathway enrichment analysis for E2F4 targets

If a E2F4 binding site could be mapped within the core promoter region of a gene i.e. (2 kb upstream or downstream from the TSS), that gene was termed a E2F4 target. Applying this criteria we identified 7,086 target genes that represent nearly a third of all annotated protein coding genes. We carried out functional analysis on the set of genes that showed a E2F4 binding site in their core promoter region using the web-based tool DAVID [88]. We recovered previously known functional annotations associated with E2F4 targets such as cell cycle regulation, DNA repair, RNA processing, stress response and apoptosis. We also found several novel interesting functional pathways such as protein transport, chromatin turnover and sterol biosynthesis. We also found a significant overlap between E2F4 targets and components of the p53 signaling pathway (*P-value* < 10^{-5}), which suggests a possible collaboration between the E2F4 and the tumor suppressor p53. We also found that E2F4 was bound to the core promoters of all E2F4 family members except E2F4 itself and E2F6. It must be noted that E2F6 acts redundantly with E2F4 as a repressor [89]. Additionally, we observed E2F4 binding sites in the core promoter regions of all three RB family proteins (pRB, p107 and p130) as well DP1, which are known E2F4 co-factors and required for its function as a repressor [89]. These results indicate that E2F4 auto regulates its own function, but whether it activates or suppresses its co-factors remains to be seen.

MATERIALS AND METHODS

E2F4 ChIP

E2F4 ChIP was performed by Bum Kyu Lee as described in [58].

Motif discovery

Since attempting to run motif discovery algorithms on all 16,426 peaks would have been computationally expensive, we divided peaks into strong (500 peaks, peak score: 24.93 to 250), moderate (500 peaks, peak scores 8.01 to 9.01) and weak categories (500 peaks, peak scores 5.6 to 5.75). A 200 bp region centered on each peak was extracted from the human genome assembly hg18. Motif discovery was performed using the software DRIM [81] on each category separately at a *P-value* cut-off of 1×10^{-5}. A random background was generated by sampling 200,000 sequences of 200 bp from the genome. 55.9% of E2F4 peaks lie within 2 kb upstream and downstream of TSSs of genes and this ratio was maintained in the random sample.

Co-enrichment

Conserved transcription factor binding site data for the human genome assembly hg18 was obtained from http://genome.ucsc.edu. This data contains transcription factor bind sites (TFBS) for 398 transcription factors from the TRANSFAC database that are conserved between human, mouse and rat species (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=118849564&c=chr13&g=tfbsConsSites). A TFBS was considered associated with a E2F4 peak if it was found within 250 bp of that peak. The frequencies of TFBS associated with peaks were calculated for E2F4 peaks and for the randomly generated peaks. The analysis was performed on the strong moderate and weak categories separately and p-values were calculated according to a binomial model. We filtered TFBS that were not enriched at p-value of < 1 x 10-6 and were associated with less than 4% of the E2F4 peaks under consideration.

Chapter 4: Nucleosome Remodeling Across a Eukaryotic Genome in Response to Transcriptional Perturbation

INTRODUCTION

How do you pack approximately 2 meters of genomic DNA into the small confines of a nucleus? The evolutionary solution to this problem is the nucleosome [17]. Nucleosomes are the building blocks of eukaryotic chromosomes. They consist of about 147 bp of DNA wrapped around a histone octamer protein core. The protein core comprises of two copies of the histones H2A, H2B, H3, and H4 around which is wrapped about 1.65 turns of helical DNA [18]. Nucleosomes are arranged at regular intervals along the genome and hence have a "beads-on-a-string" appearance [19]. The genomic region in between two nucleosomes is called the linker region and this maybe occupied by the histone H1 [17]. Linker regions are more accessible to nucleases like micrococcal nuclease or MNase that cleave in between two nucleosomes [90].

Nucleosomes limit or allow accessibility to transcription factor binding sites and are involved in multiple cellular processes [23]. However, the nucleosomal locations on a genome-wide scale are still poorly understood. There have been attempts to explain nucleosome positioning by the underlying sequence constraints but these predictions fare only about 15% better than random guessing [25, 91]. Studies mapping nucleosome positions using high-resolution microarrays have shown that most assayable nucleosomes are well positioned [26, 90]. In vivo nucleosome positions are influenced by a variety of chromatin remodelers as well as the transcriptional machinery [92]. Chromatin remodeling also involves the exchange of common histones in the nucleosomal core with histone variants that influence local chromatin structure [22].

Recently, ChIP-sequencing technology was used to map H2AZ variant containing nucleosomes across the yeast genome. This study mapped about 10,000 such nucleosomes that are mainly enriched near promoters [24]. Another study using high-resolution tiling arrays (the arrays had probes every 4 bp tiled across the entire yeast genome) mapped the positions of ~50,000 nucleosomes in the yeast genome [26]. However, the remodeling of individual nucleosomes in response to transcriptional reprogramming has not yet been studied in any eukaryotic.

We used ultra high-throughput sequencing methodology (Solexa/Illumina) to sequence the ends of nucleosome associated DNA and mapped individual nucleosomes on a genomic scale at a single bp resolution. We show that nucleosome density and stability over promoters and coding regions were well correlated with transcription rate rather than absolute transcript levels. Two distinct modes of chromatin remodeling were associated with transcriptional regulation. On transcriptional perturbation, gene activation was mainly accompanied by the eviction of 1-2 nucleosomes from the promoter, while gene repression was mainly accompanied by the appearance of nucleosomes with varying stability over the promoter. Our work constitutes the first study of dynamic single-nucleosome remodeling in response to transcriptional perturbation across an entire eukaryotic genome.

RESULTS

Identifying nucleosomes by ultra-highthroughput sequencing

Micrococcal nuclease (MNase) cuts chromatin in between nucleosomes at the linker region. We isolated chromatin from normally growing yeast cells and yeast cells that were subjected to heat-shock. Chromatin was then subjected to MNase digestion and DNA associated with mono-nucleosomes was gel extracted. Gel extracted DNA was then

sequenced by Solexa sequencing technology. Solexa sequencing generates 30-35 bp reads from the ends of DNA fragments. Reads were mapped back to the yeast genome using the eland alignment software. We used a Parzen window based clustering algorithm to define peaks on the plus and minus strand separately as described in chapter 2 (Fig. 4.1). A plus peak followed by a minus peak within 100-200 bp was defined as a nucleosome. A 146 bp window was then positioned around the center of the plus and minus peaks. We calculated two scores for each nucleosome. The first score (hereafter referred to as score) defined the strength or the confidence of detecting a nucleosome at the given position while the second score (hereafter referred to as the stability score) defined how well-positioned or fuzzy the detected nucleosome is.

The depth of sequencing was different for normal growth sequence library (~500,000) as compared to that generated from heat-shocked cells. In order to account for this difference we sought to normalize the peak scores so that scores across the two datasets are comparable. The normalization procedure generated score between 0 and 1 and has been explained in detail in the Methods section. Nucleosomes above a score cut-off of 0.25 showed good correlation between normal and heat-shocked samples and hence we chose 0.25 has a cut-off to define nucleosomes. At the indicated score cut-off, we detected 49,043 nucleosomes in normally growing yeast cells and 52,817 nucleosomes in heat-shocked yeast cells. The linker region in our datasets i.e. the distance between the end of a nucleosome and the start of the next one, was approximately 20-50 bp with a mode at 30 bp. Assuming a linker region of 50 bp, the distance between the starts of two adjacent nucleosomes cannot be less than 196 bp (146 + 50) or ~ 200 bp. Since the yeast genome is about 16 Mb in length, this suggests that

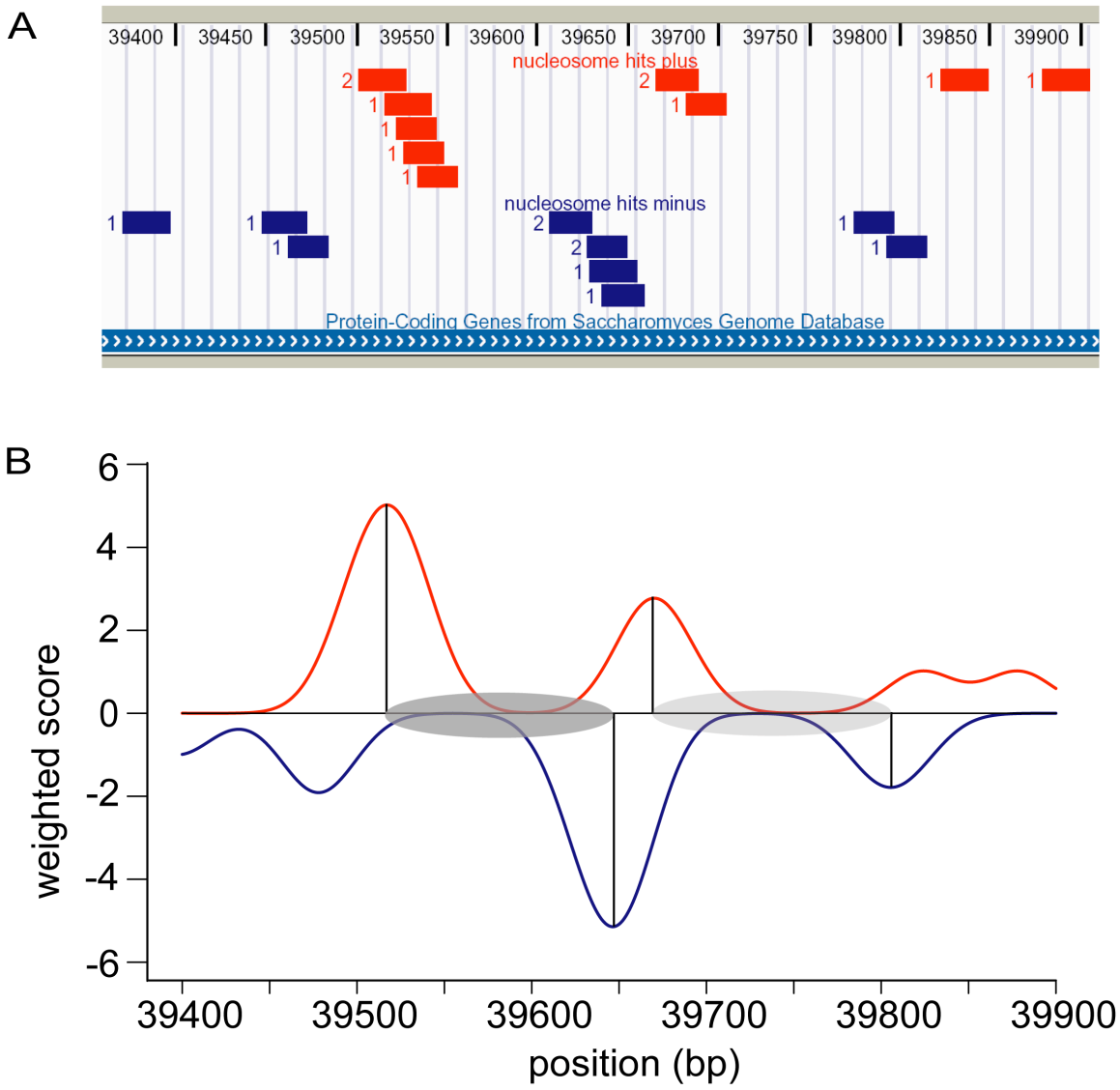


Figure 4.1 Nucleosome mapping algorithm.

(A) Reads mapped back to the plus strand are displayed as red horizontal bars while those mapped back to the minus strand are displayed in blue horizontal bars. (B) The Parzen scores at each bp are displayed across the same genomic region. Local maximas were detected as peaks. A plus strand peak followed by a minus strand peak within a distance of 100-200 bp was detected as a nucleosome. Reproduced from [93].

73% of the yeast genome is occupied by nucleosomes. In our analysis, only reads that mapped uniquely to the yeast genome were used for mapping nucleosome. Taking into account that a sizeable fraction of the yeast genome is repeated, we estimate that around 78% of the yeast genome is covered by nucleosomes.

Recapitulating previously known biological data

We assessed the accuracy of our data by comparing with known nucleosome positions across the PHO5 promoter (Fig. 4.2A). The yeast PHO5 gene is repressed during growth in rich media by well-positioned nucleosomes covering the binding site for the transcription factor PHO4 [94]. The nucleosome positions calculated by our detection method matched well with the three known nucleosome positions. The positions of these nucleosomes did not change between the normal and heat-shocked samples as expected since the PHO5 gene is unaffected by the transcriptional perturbation of heat-shock (Fig. 4.2B). Individual promoters of genes that respond to heat-shock showed clear evidence of nucleosomal remodeling. For example, SSA4 is a heat-shock activated gene [95] that shows well-positioned nucleosomes in the promoter under normal conditions (Fig. 4.3A). On heat-shock, SSA4 is activated and accordingly, nucleosomes were seen to be evicted from the SSA4 promoter. RPL21 is a ribosomal gene that did not show any nucleosome positioning in its promoter under normal growth (Fig. 4.3B). On heat-shock, RPL21 showed appearance of well-positioned nucleosomes in its promoter and accordingly is repressed [95]. A catalog of nucleosome positions in yeast mapped using high-resolution tiling arrays was recently published [26]. Our nucleosome positions correspond very well to this recent study (Appendix E). Thus our detected nucleosome positions mapped known positions and showed expected remodeling changes across promoters that were responsive to heat-shock.

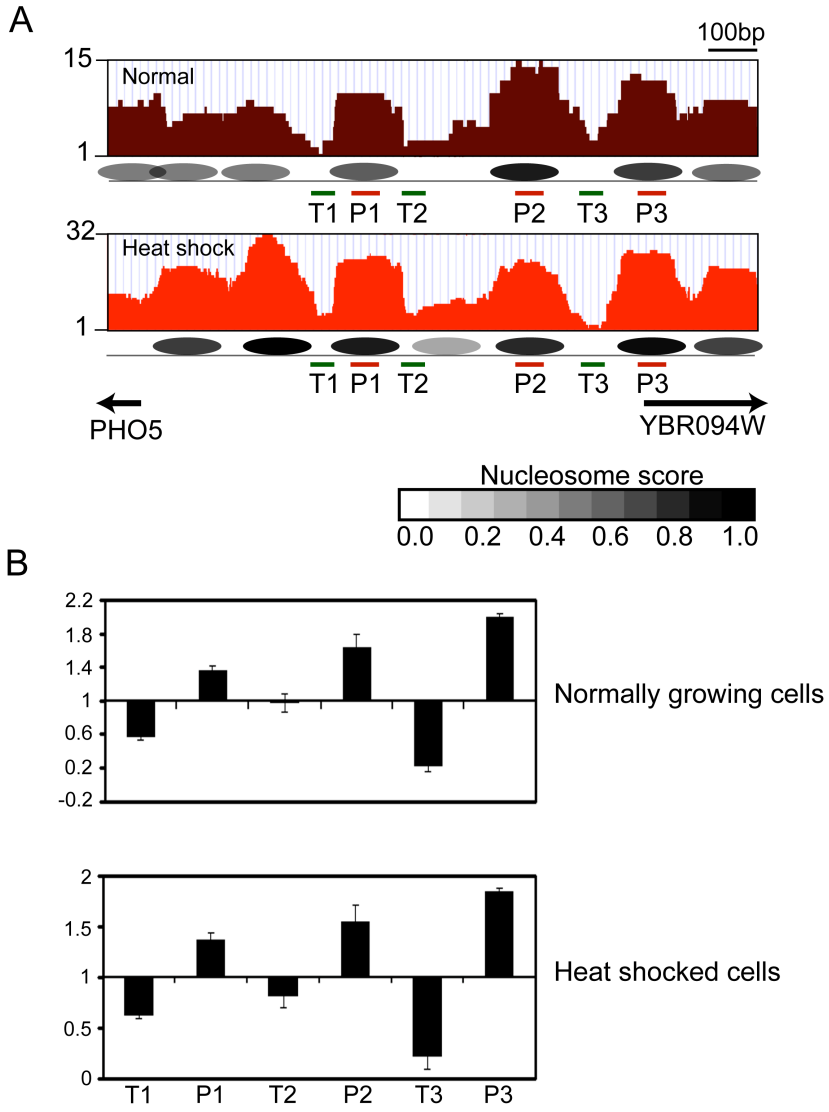


Figure 4.2 Nucleosomal positions recaptured by sequencing

(A) Detailed view of the *PHO5* locus showing raw sequence read densities under normal growth conditions (brown) and heat shock (red). Nucleosome positions detected by our algorithm are shown as ovals shaded according to their nucleosome scores as indicated. The regions analyzed by qPCR are shown in green (troughs) and red (peaks). Black arrows indicate the direction of genes in that locus. (B) qPCR data confirms the nucleosome (P1,P2,P3) and linker (T1,T2,T3) positions detected by our analysis. Assayed nucleosome positions remained unchanged between normal and heat-shock conditions. Reproduced from [93].

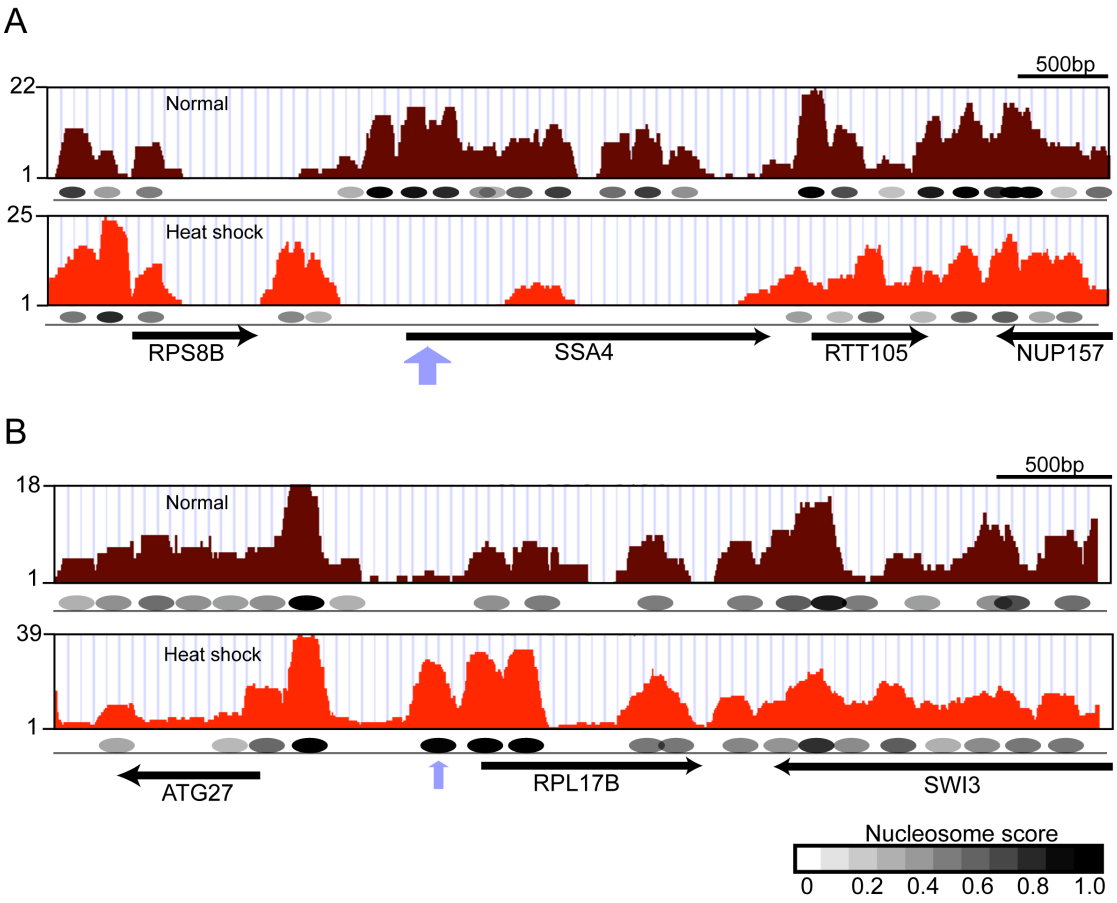


Figure 4.3 Nucleosome remodeling captured by sequencing

(A) Detailed view of the SSA4 promoter. SSA4 is a heat shock activated gene that is repressed under normal growth conditions. Accordingly, the SSA4 promoter and coding region is covered by an array of nucleosomes that are evicted upon heat shock (thick purple arrow). (B) Detailed view of the RPL17B locus. RPL17B is a ribosomal protein gene that is highly expressed under normal growth conditions but repressed upon heat shock. Accordingly, a single well-positioned nucleosome appears over the RPL17B promoter (thin purple arrow). The nucleosome positions calculated using our algorithm are indicated as shaded ovals while the score is indicated in grayscale. Reproduced from [93].

Nucleosome occupancy at promoters vs. coding regions

PCR microarrays have shown the promoters are depleted of nucleosomes as compared to coding regions [96]. Accordingly, we saw that the number and stability of nucleosomes was lower in promoters as compared to coding regions (*P-value* < 2.2 x 10^{-16}). We plotted the average nucleosome occupancy profiles across the promoters and coding regions of all yeast genes (Fig. 4.4A). First, as reported previously, promoters had lower occupancy than coding regions [90]. Second, the nucleosome free region upstream of the transcription start sites (TSS) is equivalent to a width of a nucleosome. Third, there was a strongly positioned nucleosome immediately downstream of the TSS that coincided with the previously mapped H2AZ nucleosome [24]. Fourth, nucleosomes are positioned at regular intervals but with decreasing strengths over the coding regions. These observations correlate well with what was observed on a single yeast chromosome using microarrays. Nucleosome data from heat-shocked samples gave similar results. Interestingly, we also observed a strongly positioned nucleosome at the 3' end of genes followed by a nucleosome free region (Fig. 4.4B). This nucleosome signal was not an artifact of closely positioned genes in the yeast genome as the signal persisted even in convergently transcribed genes that would lack a promoter immediately downstream of their 3' ends. This 3'end nucleosome signal showed a moderate association with long genes and with genes that were expressed at a lower rate.

Influence of the presence of a TATA box and transcription rate on nucleosome positioning

We reasoned that the average nucleosome profiles across gene TSSs might conceal distinct patterns of nucleosome occupancy. To reveal such sub-patterns, we clustered the

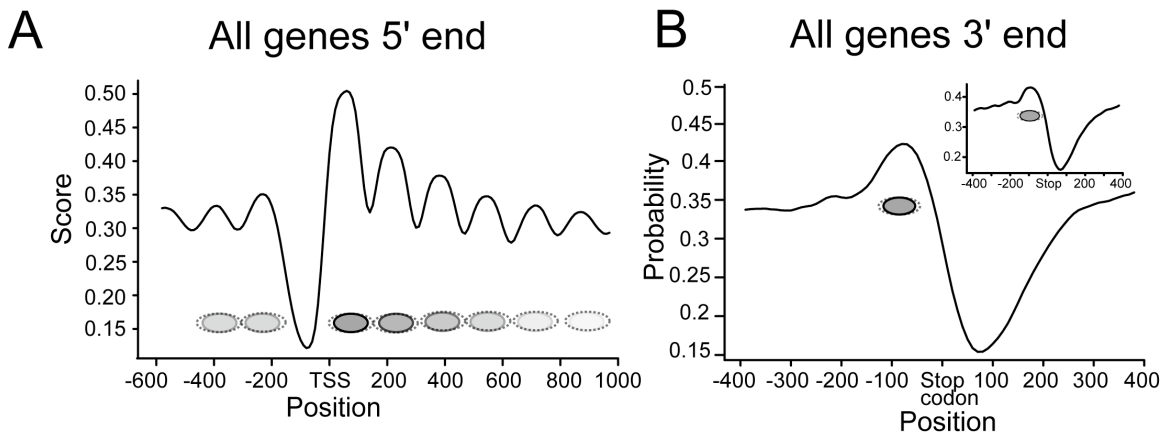


Figure 4.4 Nucleosome positioning around the TSS and stop codons

(A) Average nucleosome profiles across all yeast genes from -600 to +1000 from the TSS of yeast genes. Nuclesome positions are shown as gray ovals with scores as indicated in Fig. 4.2 and Fig. 4.3. The dotted ovals indicate the fuzziness of the nucleosomes. (B) The 3' end of genes is marked by a well-positioned nucleosome with a nucleosome-free region immediately downstream. The inset shows the 3' end of convergent genes indicating that the observed pattern is not due to an adjacent promoter. Adapted from [93].

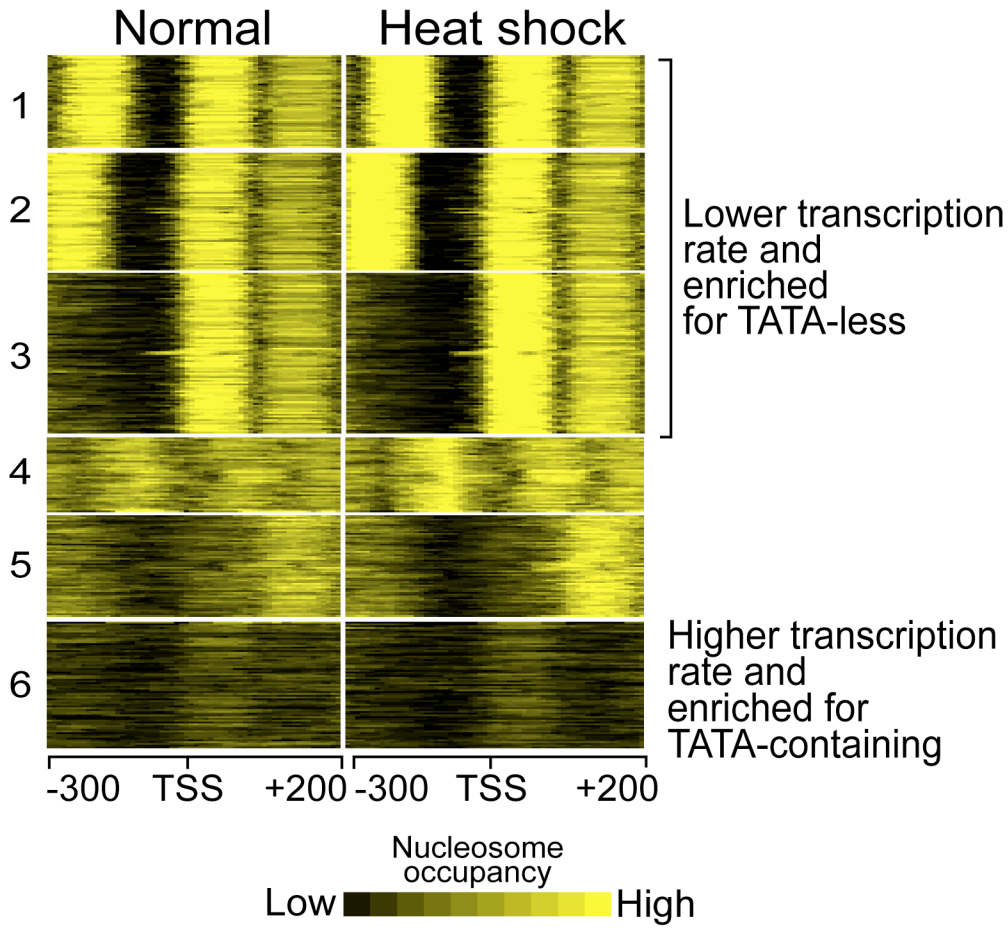


Figure 4.5 Nucleosome profiles across promoters

K-means clustering reveals distinct patterns of nucleosome positioning. The nucleosomes are color coded by their probability as indicated by the color scale. Clusters 1-3 were enriched for TATA-less promoters and showed lower transcription rate ($p <= 10^{-10}$). Clusters 6 were enriched for genes that were TATA-containing and showed higher transcription rate ($p < 10^{-10}$). Adapted from [93].

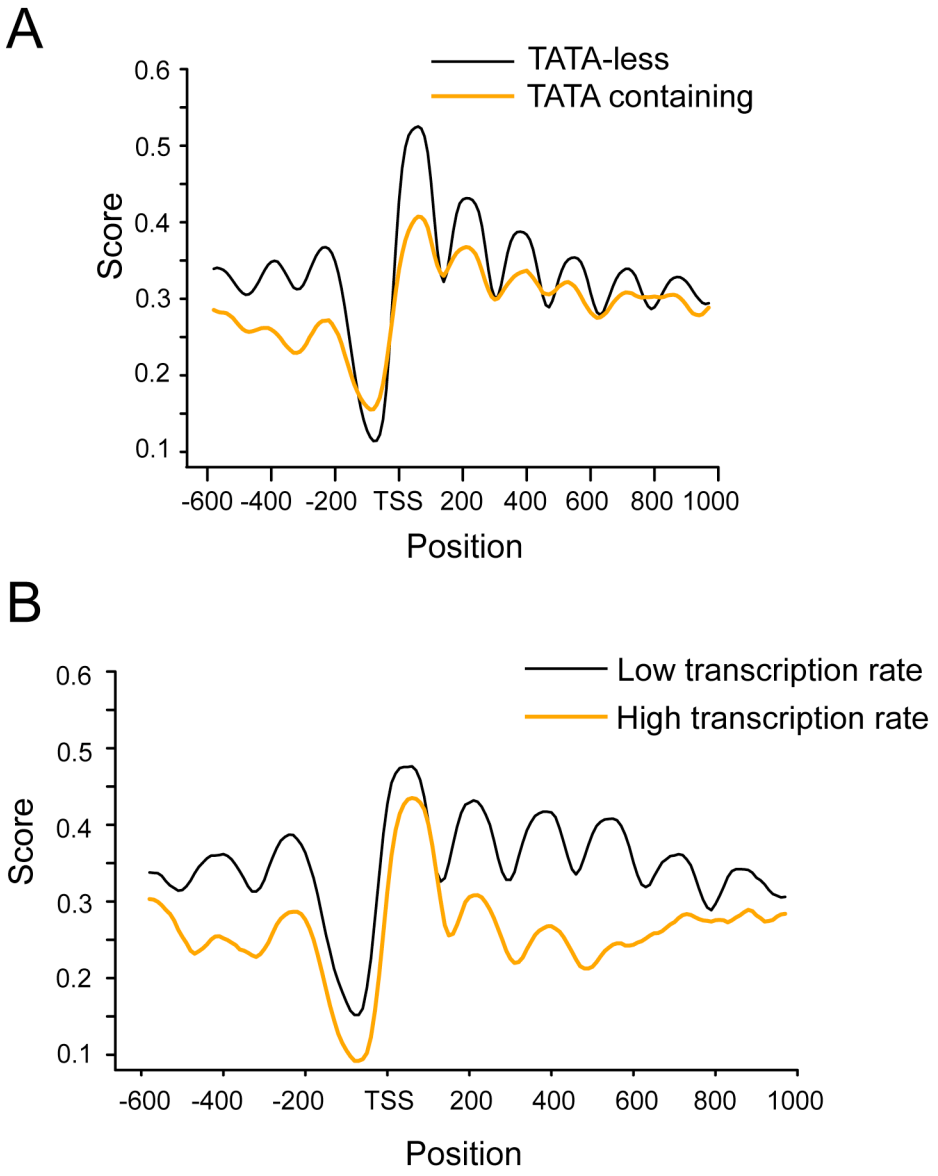


Figure 4.6 Dependence of nucleosome positioning on the TATA box and transcription rate

(A) Average nucleosome profiles across TATA-less (4965) and TATA-containing promoters (1074) aligned with respect to the TSS. (B) All genes in the yeast genome were sorted in a descending order according to their transcription rates. Average nucleosome profiles were generate for the top 500 genes (highly transcribed, orange) and bottom 500 genes (low transcription rate, black). Adapted from [93].

nucleosome profiles by k-means clustering. This analysis revealed several classes of nucleosome occupancy patterns (Fig. 4.5). There was no significant distinction between these different promoter classes with the respect to TBP occupancy or absolute transcript levels. However, promoters with nucleosome occupancy patterns showing well-positioned nucleosomes were enriched for TATA-less promoters and on average had a low transcription rate. Conversely, promoters with fuzzy nucleosome positioning were enriched for TATA containing promoters and higher transcription rates. We segregated promoters into TATA containing or not and then generated average nucleosome profiles (Fig. 4.6A). This analysis showed that the presence of a TATA containing box was strongly correlated with fuzzy nucleosome positioning while TATA-less promoters showed the stereotypical average patterns that was seen across all yeast genes. This pattern difference between TATA-less and TATA containing promoters was not due to differences in gene numbers between the two sets. Segregating promoters by transcription rate showed that promoters with higher transcription rates showed stronger nucleosomes positioning as compared to promoters with lower transcription rates (Fig. 4.6B).

Nucleosome positioning was largely retained between normal and heat-shocked cells. Approximately, 65% of positioned nucleosomes in normally growing cells were within 30 bp from nucleosomes detected in heat-shocked cells. At a threshold of 0.25, less than 10% of nucleosomes shifted their positions by more than 100 bp. In addition to nucleosome profile across the promoters, we also observed regular positioning of nucleosomes over the coding regions. We aligned nuclesomes over the coding regions of all genes by the first nucleosome immediately downstream of the TSS (Fig. 4.7) and assigned each profile a nucleosome positioning score (NPP). Higher the NPP, better positioned were the nucleosomes. We then sorted the nucleosome profiles by the NPP

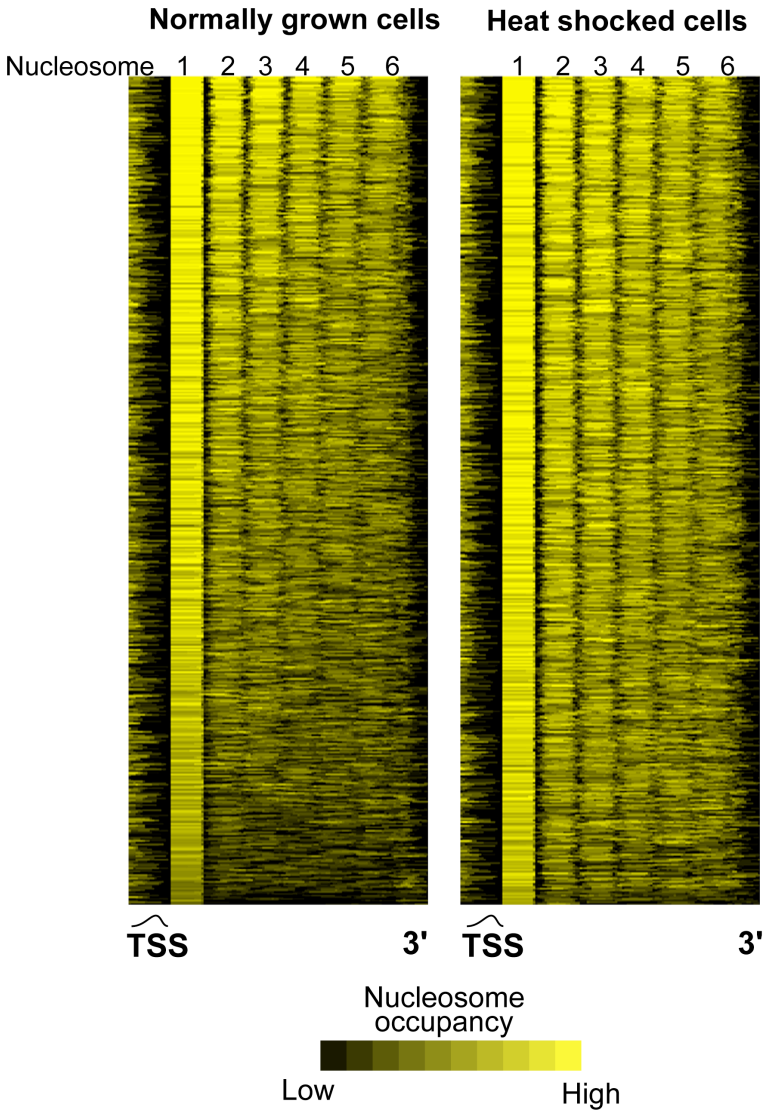


Figure 4.7 Nucleosome positioning over coding regions

Nucleosome positions in the coding regions of all yeast genes were aligned by the first nucleosome and assigned a nucleosome positioning score (NPP) based on how well nucleosomes were positioned. Higher NPP score indicated an array of well-organized nucleosomes. Genes were sorted in a descending order according to their NPP scores. Nucleosome profiles in the coding regions essentially remain the same between normally growing and heat shocked cells. The non-aligned TSS is approximated by the curve. Adapted from [93].

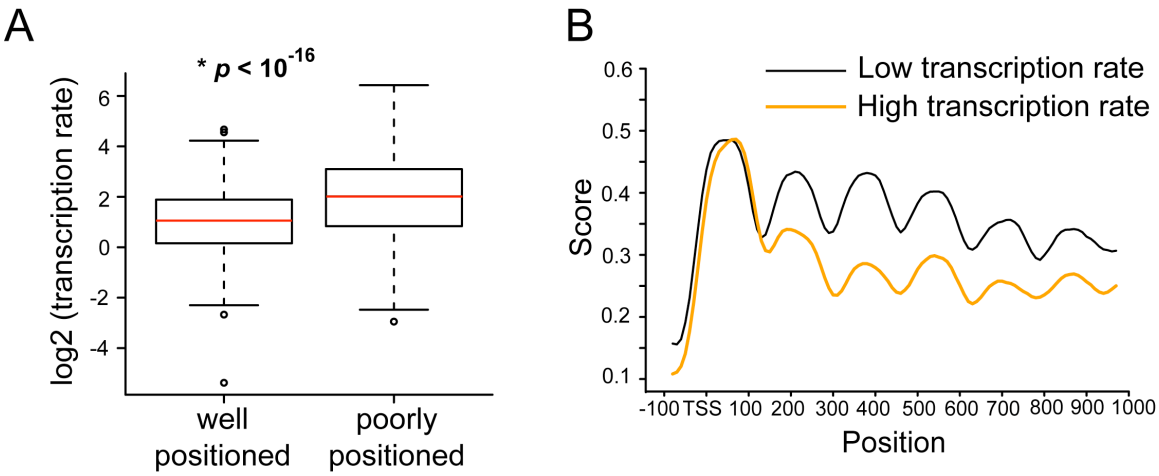


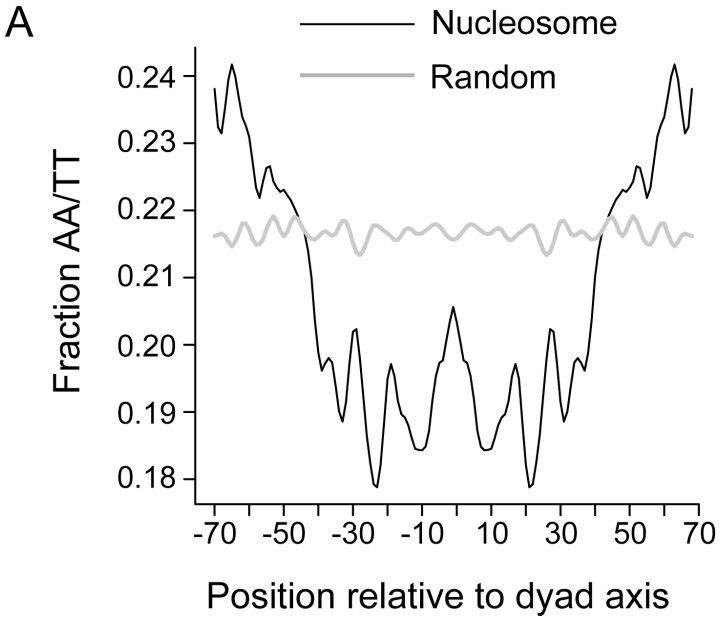
Figure 4.8 Influence of transcription rate on nucleosome positioning in coding regions.

(A) The transcription rate of genes with well-positioned nucleosomes (high NPP score) is lower than that of genes with poorly positioned nucleosomes (low NPP score). The red line indicates the median transcription rate, the boundaries of the box indicate the 1st and 3rd quartiles, the horizontal line connected to the box with dashes indicate the data dispersion while the circles indicate the outliers. (B) Gene were sorted in a descending manner according to their transcription rates and the average nucleosome profiles across the coding regions were plotted for the top 500 genes (high transcription rate, orange) and bottom 500 genes (low transcription rate, black). Adapted from [93].

score in a descending fashion. We did not observe any correlation of the NPP scores with absolute transcript levels. However, genes with higher NPP scores had lower transcription rates as compared with genes that had lower NPP scores (Fig. 4.8A). Conversely, genes that had lower transcription rates had higher NPP scores i.e. showed well-positioned nucleosomes in the coding regions as compared to highly transcribed genes (Fig. 4.8B). Interestingly, the first nucleosome immediately downstream of the TSS was unaffected by the transcription rate (Fig. 4.8B). Overall, it seemed that nucleosome positioning in yeast was not random but was strongly influenced by transcription rate.

Sequence dependent nucleosome positioning

Previous analyses of DNA sequences have revealed that nucleosome positioning is strongly influenced by underlying DNA sequence. However, predictions of in vivo nucleosome positioning in yeast have not met with much success. It is still not clear to what extent the underlying DNA sequence contributes to nucleosomal positioning. One possibility we entertained was that a nucleosome is positioned by its underlying sequence, and then the subsequent nucleosomes stack with respective to the already positioned nucleosome. In particular, the regular array of nucleosomes that we observed in the coding regions might result from nucleosomes stacking up against the first well-positioned nucleosome, possibly the H2AZ. To test this hypothesis, we examined the sequence dependence of nucleosomal consecutive nucleosomal positions in the array that was observed in the coding region. We generated a "reference" profile of the relative frequencies of AA/TT dinucleotides from the sequence underlying the top 500 strongly positioned first coding nucleosomes (Fig. 4.9A). Previous sequence analyses of nucleosomal DNA have demonstrated a 10 bp periodicity in the AA/TT dinucleotide frequency pattern and such a repeating pattern was observed in our analysis as well. Although the information content in our pattern is not as high as observed before, our profile is significantly different from a profile generated from aligning random sequences. We then correlated the reference profile to AA/TT dinucleotide profiles underlying subsequent positioned nucleosomes and calculated an average correlation for each position. As expected, the first position showed the highest correlation with the reference profile. Interestingly, the subsequent positions showed significantly high correlations though the value were lower than that observed for the first position (Fig. 4.9B). This indicated that although the underlying sequence makes a modest contribution to nucleosome positioning, this contribution is



B

Nucleosome	1	2	3	4
Correlation	0.90	0.75	0.62	0.67

Figure 4.9 Contribution of sequence to nucleosome positioning.

(A) AA/TT dinucleotide frequencies at each position in the DNA sequence associated with the top 1000 well positioned first nucleosomes in Fig. 4.7 were averaged and smoothed by a 3 bp moving average (black). The same analysis as performed on a random set of sequences from the yeast genome (gray). (B) AA/TT dinucleotide frequency profiles were generated for each indicated nucleosome position in the coding region and correlated with the AA/TT profile generate in Fig. 4.9A. Table shows the correlation coefficients for the respective positions. All correlation values were significant. Adapted from [93].

maintained even in nucleosomal arrays formed adjacent to strongly positioned nucleosomes.

Nucleosome positioning is strongly influenced by dynamic changes in transcription

In order to examine how transcriptional changes affect global nucleosomal positioning, we generate nucleosomal remodeling profiles across all yeast promoters by subtracting nucleosome profiles in heat-shocked cells from normal yeast cells bin-wise as described in Materials and Methods. Thus a positive value at any given bin indicated that a nucleosome present under normal conditions was evicted after heat-shock while a negative value indicated the appearance of a nucleosome after heat-shock. We performed k-means clustering on the subtracted profiles to generate distinct patterns of remodeling across yeast promoters.

We first analyzed remodeling across promoters that were activated at least 2-fold in response to heat-shock (Fig. 4.10). Within the activated class, we could define two distinct groups of promoters (groups 2 & 4) where a nucleosome covered the promoter under normal growth conditions thereby preventing access to the transcriptional machinery but was evicted upon heat shock. Of these group 2 showed a significant enrichment for the targets of the activator Msn4 (*P-value* = 0.02) [16]. Promoters in group 1 had a nucleosome-free region between -200 and 0 under normal and heat-shock conditions. This group was enriched for targets of the transcriptional activator Hsf1 (*P-value* < 0.02) [95]. Group 3 showed enrichment for the remodeler Swi5 (*P-value* = 0.002) . The difference in nucleosomal remodeling associated with the two heat-shock responsive transcriptional activators points to two different modes of regulation by the two factors. Hsf1 is pre-bound to many of its target promoters even under normal growth conditions

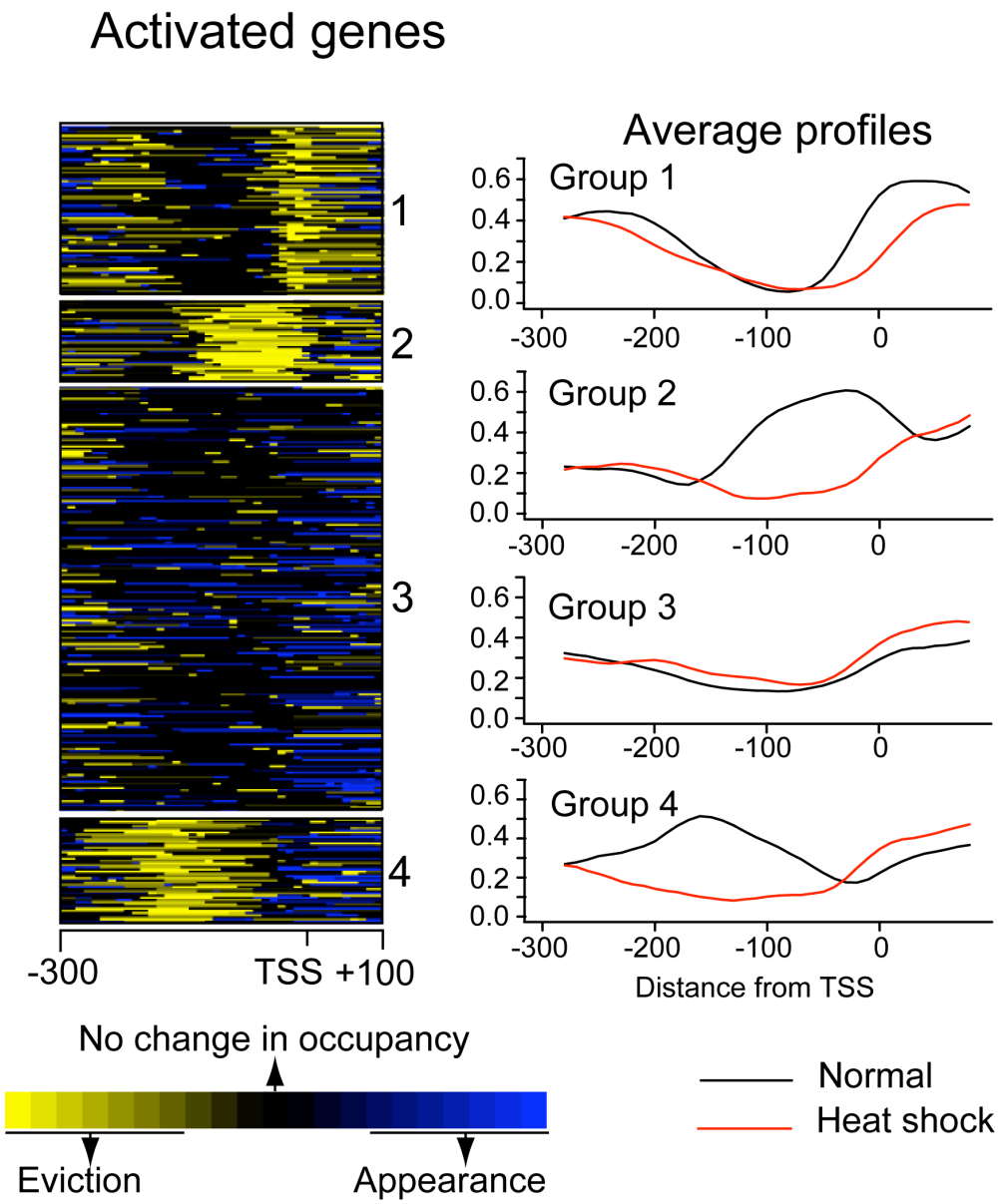


Figure 4.10 Nucleosome remodeling across activated promoters

Remodeling profiles of genes activated more than 2-fold upon heat shock were aligned with respect to the TSS and clustered by k-means clustering. Nucleosomes present during normal growth but evicted upon heat shock are represented by yellow while nuclesomes appearing after heat shock are indicated by blue. The line graphs show the average profiles per group. K-means clustering was performed by using data from -200 to TSS, but profiles displayed are shown for -300 to +100. Adapted from [93].

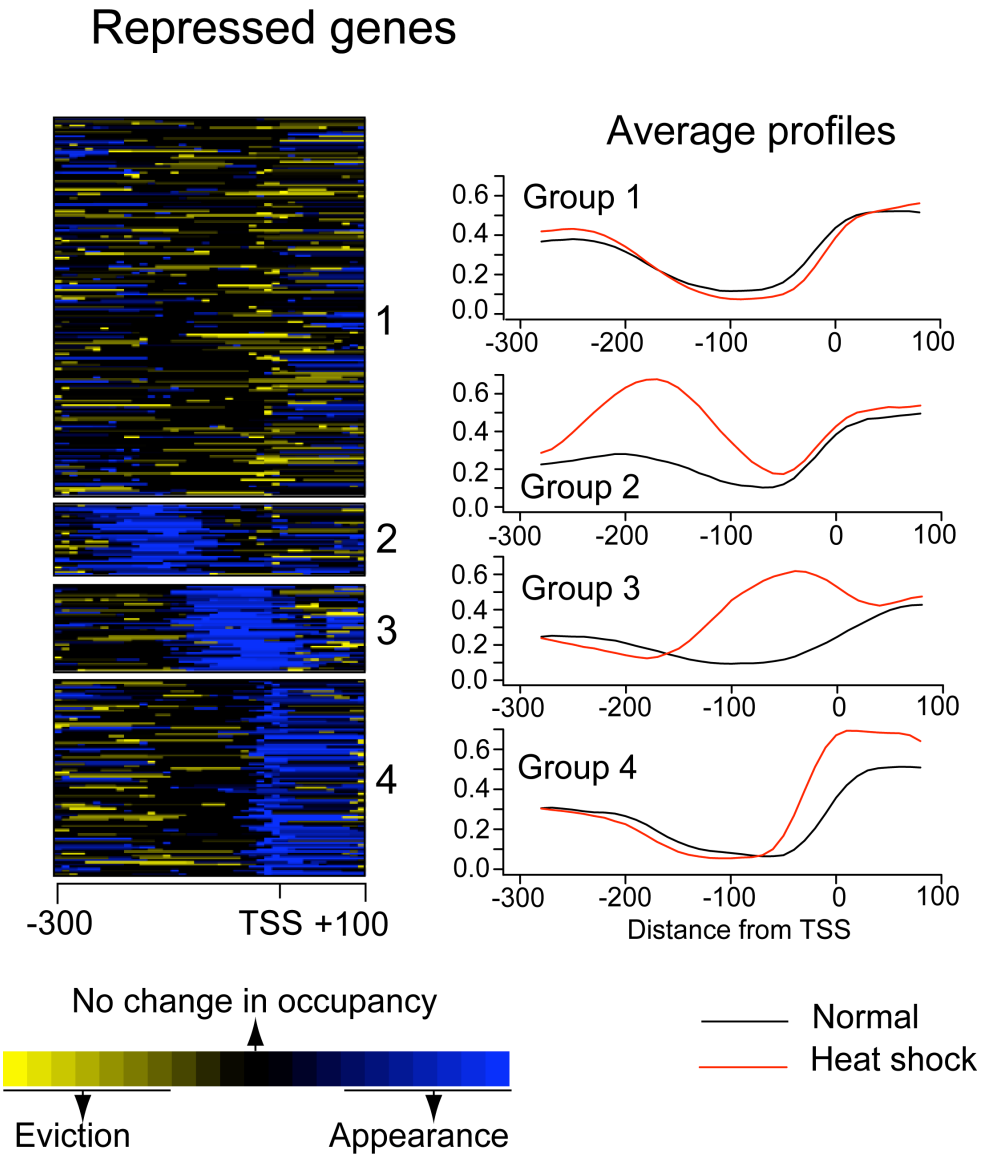


Figure 4.11 Nucleosome remodeling across repressed promoters

Remodeling profiles of genes repressed more than 2-fold upon heat shock were aligned with respect to the TSS and clustered by k-means clustering. Data was generated as described in Fig. 4.10. Adapted from [93].

[95] and this may explain the nucleosome-free region observed both before and after heat-shock. On the other hand Msn4 translocates to the nucleus after heat-shock to activate downstream promoters [97] and this may explain the eviction of the nucleosome observed in group 2.

Promoters that were repressed more than 2-fold after heat shock were clustered into four groups based on their nucleosome remodeling profiles (Fig. 4.11). Promoters in group 2 had a nucleosome-free regions between -200 and -100 bp upstream of the TSS under normal growth conditions, which was occupied by a single nucleosome after heat shock. Group 3 repressed promoters showed a similar profile as group 2 with the exception that the nucleosome appearing after heat shock covered an area between -125 and +50 bp relative to the TSS. Group 1 and group 4 sets of promoters both showed a nucleosome-free region between -200 and 0 bp relative to the TSS irrespective of the transcriptional status. Group 4 showed a moderate increase in nucleosome occupancy immediately downstream of the TSS. Group 3 was significantly enriched for targets of the transcription factors Rap1, Sfp1, Esa1, Fhl1, and Gcn5, all of which are involved in regulating ribosomal protein genes under normal growth conditions [98-101] (Appendix F). Accordingly, we found a significant enrichment for ribosomal genes in group 3 ($p = 2.6 \times 10^{-5}$). Group 1 was depleted for the targets of all above mentioned transcription factors and was also significantly depleted for ribosomal protein genes ($p = 6.5 \times 10^{-4}$). We also looked at the set of ribosomal promoters for remodeling changes. Under normal growth conditions, there is a well-positioned nucleosome immediately downstream from the TSS. Upon heat-shock, we observed nucleosomes appearing upstream of or covering the TSS across most ribosomal promoters (Fig. 4.12).

We calculated a nucleosome remodeling score for each promoter that estimated the extent of nucleosome eviction as well as appearance. In general, activation was

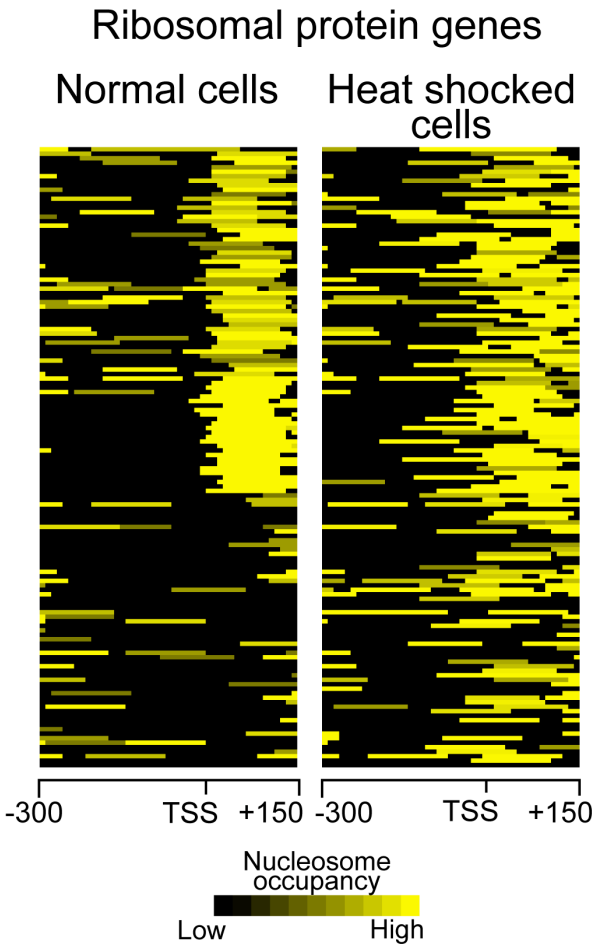


Figure 4.12 Nucleosome profiles across ribosomal promoters

Nucleosome profiles were generated across ribosomal promoters under normal growth conditions and heat shock. Data used to generate the clusters was derived from -200 to +100 bp but profiles displayed are between -300 and +150. Adapted from [93].

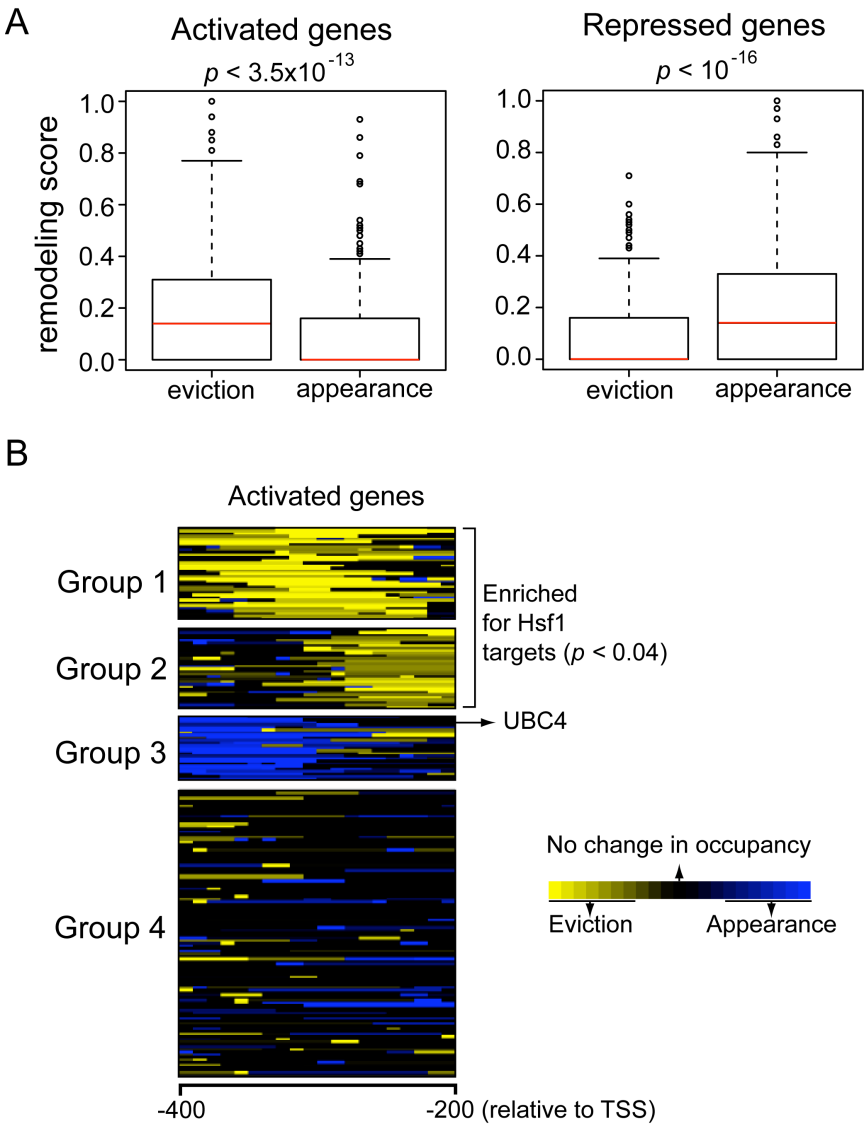


Figure 4.13 Remodeling paradigms

(A) Eviction and appearance of nucleosomes across promoters was quantified using a remodeling score (Materials and Methods) and calculated for activated and repressed genes. Box plot shows that activated genes showed higher levels of eviction while repressed gene showed higher levels of appearance. (B) Activation is not always associated with eviction. Remodeling profiles across activated genes (greater than 2-fold) between -400 and -200 were clustered. Group 3 shows a nucleosome appearing distal to the TSS. Nucleosome positions are color coded as in Figure 4.2. Adapted from [93]

associated with nucleosome eviction while repression was associated with nucleosome appearance (Fig. 4.13A). Although these trends are expected, we noticed that if we cluster the remodeling profiles based on promoter regions -400 to -200 bp upstream of the TSS, we observed nucleosome appearance events in activated promoters (Fig. 4.13B). Nucleosome eviction proximal to a promoter may coincide with nucleosome appearance more distally, as would be associated with translational repositioning of nucleosomes and this effect was observed at a smaller subset of promoters.

The above analysis of nucleosomal changes at promoters of genes showing significant transcriptional changes suggests that chromatin remodeling events accompanying transcriptional perturbation are restricted to a few discrete patterns involving the eviction, appearance or repositioning of one or two nucleosomes. Remodeling over larger domains was observed only for a small subset of promoters. We also analyzed remodeling events occurring at genes that showed less than 1.2 fold change in expression levels upon heat shock. Surprisingly, specific patterns of remodeling were observed in spite of no evidence of change in transcriptional status. This suggests that there is a background level of remodeling activity that is superimposed on specific remodeling changes.

Nucleosome remodeling influences accessibility of transcription factor binding sites

Nucleosome positioning can influence the accessibility of the core promoter as well as binding sites for sequence-specific transcriptional regulators. About 90% of the sites occupied by transcription factors on chromosome III under normal growth conditions were depleted of nucleosomes. Examination of single nucleosome remodeling at promoters that were activated or repressed by heat shock in our data revealed instances where the accessibility of the TSS and of experimentally defined transcription factor binding sites was indeed affected by remodeling. For example, at the UBC4 promoter

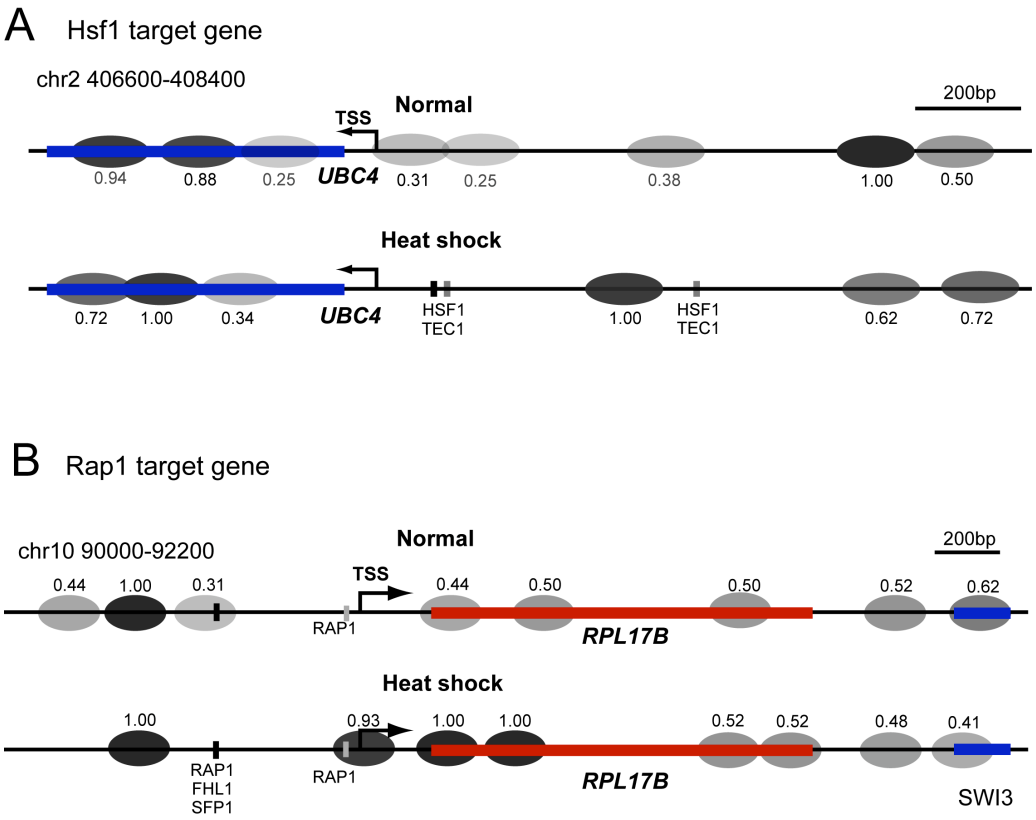


Figure 4.14 Regulating TSS and promoter accessibility

(A) Nucleosome eviction observed at the heat shock activated UBC4 promoter (blue line). Nucleosome defined by our sequencing data are shown as ovals shaded according to their scores. The positions of TFs binding sites are derived from (ref) and shaded according to their confidence. (B) Nucleosome appearance observed at the RPL17B promoter and coding regions (red line). Adapted from [93].

which is activated by heat shock, three moderately positioned nucleosomes covering two distinct Hsf1 binding sites as well as the TSS were evicted, while a single well positioned nucleosome appeared between the two Hsf1 binding sites (Fig. 4.13B and Fig. 4.14A). Conversely, at the RPL17B promoter, which is repressed by heat shock, one well-positioned nucleosome appeared after heat shock to cover the TSS and a low confidence proximal Rap1 binding site (Fig. 4.14B). Interestingly, another moderate nucleosome

upstream was evicted, exposing a higher confidence distal Rap1 binding site as well as an Fhl1 site. Such eviction and appearance of nucleosomes at adjacent sites could either reflect translational repositioning or independent events; our experiments cannot distinguish between these two possibilities.

Based on these observations and other computational predictions of whole genome nucleosome positions [4], we hypothesized that chromatin remodeling upon transcriptional perturbation could result in changes in the accessibility of the functional binding sites of stress-related transcription factors. To test this hypothesis, we measured the change in accessibility of transcription factor binding sites upon heat shock, by comparing the overlap between functional binding sites for transcription factors measured by ChIP-chip [17] and nucleosome positions before and after heat shock (Fig. 4.15). Of the 101 factors tested, 46 had fewer than 20 functional binding sites each in the genome and we therefore excluded them from this analysis. The remaining 55 transcription factors could be stratified into three classes based on the change in accessibility of the functional binding sites after heat shock: factors whose binding sites showed an increase in accessibility after heat shock (Fig. 4.15A), factors whose binding sites showed no significant change in accessibility (Fig. 4.15B), and those that showed decreased accessibility after heat shock (Fig. 4.15C). As hypothesized, most of the transcription factors involved in mediating the stress response belonged to the first group. The functional binding sites for several key stress-related transcription factors such as Hsf1, Msn2, Msn4, and Aft1 showed some of the strongest increases in accessibility because of nucleosome repositioning upon heat shock. In addition, binding sites for transcription factors Abf1 and Cbf1 that are involved directly or indirectly in chromatin remodeling [102, 103] showed increased accessibility. Surprisingly, we also observed increased accessibility for transcription factors involved

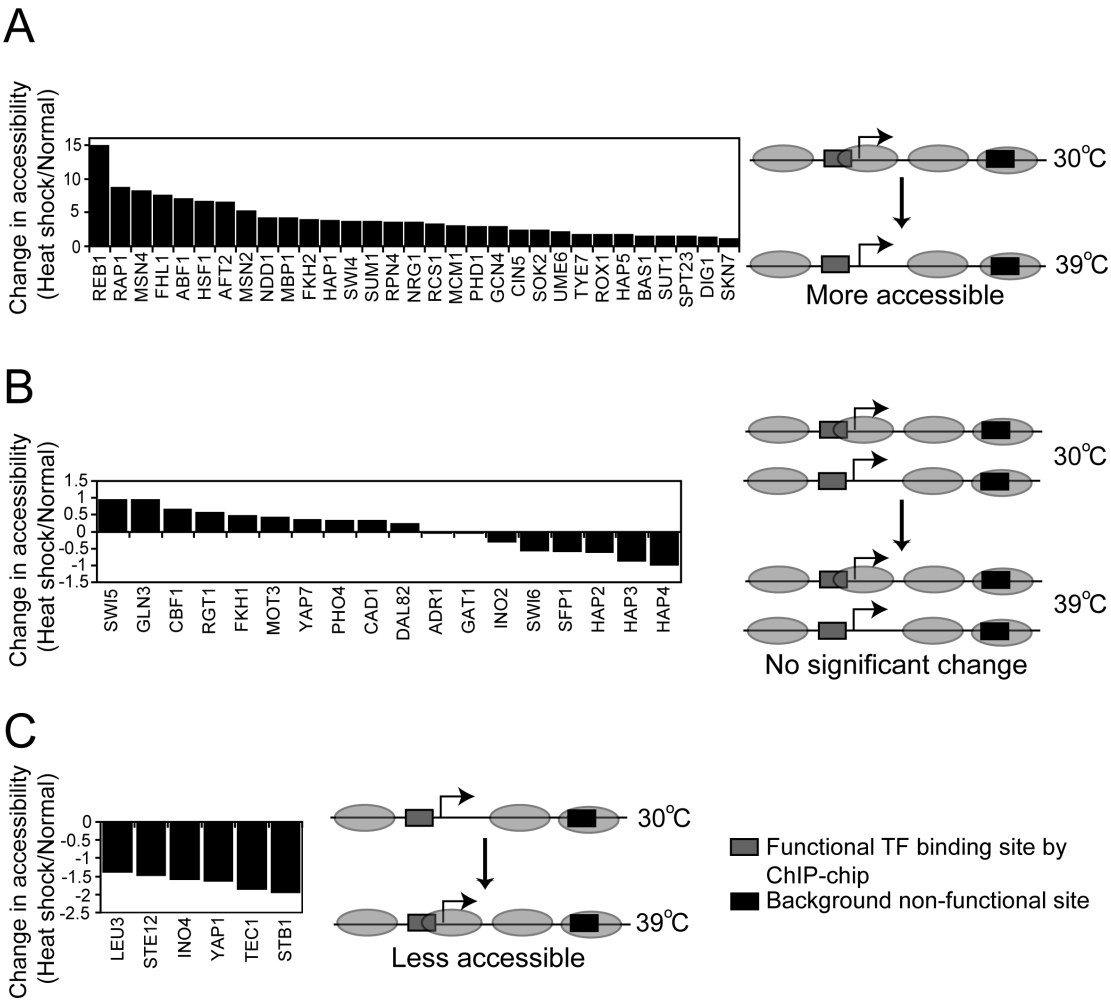


Figure 4.15 Remodeling regulates TF binding site accessibility

Transcription factors were grouped into three classes based on the change in the accessibility of their binding sites after heat shock. Changes in accessibility are represented on the Y-axis in arbitrary units. (A) TF binding sites that show increase in accessibility. (B) TF binding sites that show minimal or no change in accessibility. (C) TF binding sites that show decrease in accessibility. The schematics on the right show models representing each change. Adapted from [93].

in ribosomal protein gene transcription such as Rap1 and Fhl1 (see Fig. 4.14B for an example). These two transcription factors continue to occupy ribosomal gene promoters even during transcriptional repression [104, 105], raising the possibility that their occupancy of the promoter under such conditions, facilitated by the increased chromatin accessibility that we observed, could be related to a repressive function. Transcription factors whose binding sites did not show a significant change in accessibility were mainly those involved in the regulation of genes in metabolic pathways.

DISCUSSION

We have mapped the dynamic remodeling of most nucleosomes in the yeast genome during a transcriptional perturbation using a combination of micrococcal nuclease digestion, isolation of mononucleosome associated DNA and Solexa sequencing. Using a Parzen window based approach, which is a generally applicable method to analyze all similar datasets derived from ultra high-throughput sequencing, we defined the dynamic remodeling of approximately 50,000 nucleosomes at single nucleotide resolution in normally growing cells and in cells that were transcriptionally perturbed by heat shock for 15 minutes. Our study independently confirms expectations about nucleosomal positioning based on previous smaller scale and lower resolution studies, but also reveals novel features about chromatin structure and transcriptional activity, especially given that previous studies have not examined the dynamic repositioning of nucleosomes in response to genome wide transcriptional reprogramming.

Our results showed that in addition to a positioned nucleosome at the transcription start site, genes in general tend to also contain a well-positioned nucleosome at the 3' end of the coding region. Yeast genes are thus demarcated by a well-positioned nucleosome at each end of their transcribed regions, with a nucleosome-free gap just beyond. This could potentially reflect chromatin organization that facilitates RNA polymerase

initiation as well as termination. Most coding regions also showed strongly and regularly positioned nucleosomes, although the strength of the nucleosome positioning was weaker in genes transcribed at high rates. Interestingly, the first well positioned boundary nucleosome downstream of the TSS, which is likely to be an H2AZ variant containing nucleosome based on previous studies [24], showed similar stability in genes transcribed at high and low rates (Fig. 4.8B), suggesting that this chromatin landmark is important for demarcating promoters.

Upon transcriptional perturbation, the majority of nucleosomes did not change positions, either at promoters or within coding sequences (Fig. 4.5 and Fig. 4.7). Gene specific remodeling was restricted to the discrete eviction, appearance or repositioning of one or two nucleosomes localized to promoters. Remodeling events at genes that were activated or repressed upon heat shock could be classified into distinct patterns, indicating that there is no simple rule for nucleosome remodeling at promoters to activate and repress genes. Thus, although activation was generally and quantitatively associated with nucleosome eviction and transcriptional repression with nucleosome appearance (Fig. 4.13A), there were cases where strongly positioned nucleosomes appeared at activated promoters (Fig. 4.14). Translational repositioning of nucleosomes would seem like eviction and appearance at different spots in the same promoter. These observations suggest that nucleosome remodeling at promoters is not a trivial consequence of transcriptional activity appearing as overall openness of chromatin at activated promoters and obstruction at repressed promoters, but rather, that the precise placement of individual nucleosomes at promoters mechanistically regulates transcription by modulating access of trans-acting factors to specific sites.

In addition to chromatin remodeling specifically at regulated promoters, many promoters however showed dynamic single nucleosome remodeling during the

physiological perturbation even in the absence of any resulting transcriptional change, indicating that selective, activity-specific remodeling was accompanied by a certain number of background, non-specific remodeling events. We speculate that these background single nucleosome remodeling events poise promoters for rapid future transcriptional activity, by either assembling partial pre-initiation complexes [106], or by exchanging core histones with one or more histone variants [107]. A recent study showed that nucleosomes are globally positioned by Isw2 acting at the boundary between genes and intergenic regions, and that some of the Isw2-dependent remodeling occurs independent of transcription [108]. Therefore, the background remodeling seen in the absence of transcriptional changes in our study could potentially reflect non-specific remodeling by ISW-like complexes.

We classified transcription factors into three classes based on change in accessibility of their binding sites upon transcriptional perturbation. All the prominent stress related transcription factors belonged to the category showing a strong increase in accessibility upon transcriptional perturbation. In addition, we found that Rap1 and Fhl1 binding sites showed an increase in accessibility even though the majority of their target genes, namely the ribosomal protein genes, showed a decrease in transcription upon heat shock stress. When transcription of the ribosomal protein genes is repressed by heat shock, osmotic shock or inhibition of the TOR pathway by rapamycin, it is known that Ifh1 leaves the promoter but Rap 1 and Fhl1 remain bound [104]. It is possible that Rap1 and Fhl1 play a role in recruiting chromatin remodelers to bring about a repressive chromatin structure at the ribosomal protein genes. Previous studies have indicated that the primary discriminant between a functional and a non-functional transcription factor binding site in vivo is the presence of stably positioned nucleosomes covering the latter [24, 25]. Our results above indicate that superimposed on this, there is a second mode of

regulation at functional binding sites of stress-related transcription factors brought about by a stimulus-dependent remodeling of one or two nucleosomes, making the site more accessible for stable binding of transcription factors. Alternatively, binding of the transcription factor(s) could result in the remodeling of nucleosomes via the help of chromatin remodelers.

The work described here is the first study of genome-wide dynamic nucleosome remodeling events at single base resolution. More such studies in yeast and higher eukaryotes will shed light on the relationship between epigenetic changes at high resolution and the global regulation of gene expression.

MATERIALS AND METHODS

Preparation of mononucleosomes

Mononucleosomal DNA from yeast chromatin was extracted by Dr. Sushma Shivaswamy, a postdoctoral fellow in the Iyer lab. Yeast S288C cultures were grown in rich medium and subjected to 15 min heat shock as described previously [95]. At the end of 15 min, control and heat shocked cells (200 ml each) were treated with formaldehyde to a final concentration of 1% for 30 min. The reaction was stopped by adding glycine to a final concentration of 125 mM, and cells were harvested by centrifugation. Cells were washed in 2 X in PBS and resuspended in 20 ml of Zymolyase buffer (1 M sorbitol, 50 mM Tris pH 7.4, and 10 mM β- mercaptoethanol). Cells were spheroplasted by treating with 25 mg of 20T Zymolyase, and incubated for 40 min at 30C with shaking at 200 rpm. The remainder of the steps were carried out using a modified protocol described in [90]. Briefly, cells were spun down, washed 1 X with 5 ml Zymolyase buffer, and re-suspended in 2 ml of NP buffer (1 M Sorbitol, 50 mM NaCl, 10 mM Tris pH 7.4, 5 mM MgCl2, 0.075% NP 40, 1 mM β-mercaptoethanol, and 500 μM spermidine). CaCl2 was

added to a final concentration of 3 mM, and micrococcal nuclease digestions were carried out at concentrations ranging from 100 U/ml to 600 U/ml for 10 min at 37C. The reactions were stopped by adding 100 µl of 5% SDS and 50 mM EDTA. 3 µl of 20 mg/ml proteinase K was added to each tube, and incubated at 65C overnight. The DNA was purified by phenol-chloroform-isoamyl alcohol (25:24:1) extraction, and precipitated using ethanol. The DNA was treated with DNase-free RNase, re-extracted with phenol-chloroform-isoamyl alcohol, precipitated with ethanol, and resolved on a 1.25% agarose gel alongside a 100 bp ladder. The mononucleosome size band (approximately 150-200 bp) was excised and purified using the Invitrogen Pure-Link quick gel extraction kit. The purified DNA was sequenced using Solexa sequencing technology.

RNA isolation and expression profiling

Microarray expression profiling of yeast cells under normal and heat shocked growth conditions was performed by Dr. Sushma Shivaswamy, a postdoctoral fellow in the Iyer lab. S288C cells from 50 ml cultures before and after heat shock at 39C for 15 min were re-suspended in 8 ml of AE buffer (50 mM sodium acetate pH 5.2, 10 mM EDTA, 1.7% SDS). RNA extraction, cDNA labeling, and microarray manufacture and hybridizations were done as described previously [13]. For absolute expression analysis, sheared genomic DNA was labeled with Cy3 and cDNA was labeled with Cy5. For relative expression change analysis, cDNA from heat shocked cells was labeled with Cy5 and cDNA from normally grown cells was labeled with Cy3. The labeled cDNAs were mixed and hybridized onto DNA microarrays for 12-16 hrs. The arrays were washed, dried, and scanned with a Axon 4000B scanner (Molecular Devices). Cy5/Cy3 ratios were quantitated using GenePix Pro software and analyzed using Acuity microarray informatics software after filtering to exclude bad spots.

Quantitative PCR validation

Quantitative PCR data was generated by Dr. Sushma Shivaswamy, a postdoctoral fellow in the Iyer lab. Primer pairs used in Fig. 4.2B were designed to cover three peaks and three troughs in the promoter of PHO5 just upstream of the known Pho4 binding and DNaseI hypersensitive site [94]. Control primers used for normalization were designed in the region between YCR023C and YCR024C. qPCR was performed using SYBR green chemistry on an ABI 7900 instrument. Enrichment of target loci in the ChIP sample relative to sonicated genomic DNA was calculated for both unstressed cells and cells subjected to heat shock.

Nucleosome position detection

Solexa sequencing reads were mapped back to the Oct 2003 yeast genome assembly obtained from the SGD (http://www.yeastgenome.org/) and only reads that mapped uniquely to the genome were considered in the majority of our analysis. We generated 514,803 and 1,036,704 uniquely aligning reads for the normal and heat-shock growth conditions respectively. Reads mapping to the plus and minus strands were processed separately. Reads were clustered using a Parzen window based approach. Essentially, a Gaussian kernel was centered on each base-pair in the genome and a weighted score was calculated at that position. The mean of the Gaussian was taken as the position under consideration, with the standard deviation (smoothing bandwidth) set at 20 bp. Each read contributed to the mean position based on its kernelised distance from the mean. The weighted score indicated the likelihood of finding an edge of the nucleosome at the position. Thus, the entire genome was converted into a likelihood landscape, which was further processed to find local maxima. These maxima were then treated as centers of a cluster. Membership of a read in a cluster was based on its relative contribution to the weighted score of the center. The number of reads assigned to a

cluster was defined as the unweighted score of that cluster. We reasoned that a stable nucleosome would be expected to result in a denser clustering of the reads than an unstable one. The denser clustering of the reads results in better concordance of the unweighted score to the weighted score. Hence, each cluster was assigned a stability score that was calculated as the ratio of the unweighted score to the weighted score. Nucleosomes were identified as a plus cluster followed by a minus cluster within 100 – 200 bp. The nucleosome score was calculated as a sum of the plus and minus cluster unweighted scores. The nucleosome stability score was calculated as a weighted average of the individual stability scores of the participating clusters.

Overlap between unstressed and heat shock stressed cells

Whole genome maps for unstressed and stressed cells were filtered to exclude nucleosomes that had a normalized score less than 0.2 (see normalization procedure below). For each nucleosome in unstressed cells, the distance to the nearest nucleosome after heat shock was calculated. This data is reported in Supplementary Table 1. Similar analysis was used to determine the overlap between nucleosome positions determined in this study and those from previous studies [25, 26, 90].

Random simulations to generate a normalization factor

Reads equal in number to those we obtained from normal and heat shocked cells were selected at random from the yeast genome assembly Oct 2003 and peak finding was done as described. This process was iterated 20 times. The average maximum score obtained in the simulations was used as a scaling factor to normalize nucleosome peak scores for cells grown at 30°C. Normalization was done by dividing nucleosome peak scores by the scaling factor. We then calculated a scaling factor for the heat shock data by multiplying the scaling factor for the 30°C data by the ratio of the median peak scores for

39°C to the peak scores for 30°C. This was done to correct for differences in sequencing depth for the two samples, thus enabling quantitative comparison of nucleosome profiles across the two conditions.

Average nucleosome profiles for TATA-containing and TATA-less genes and separation by transcription rates

The upstream - 600 bp to downstream + 1000 bp of each uncharacterized and verified ORF in SGD was binned at 10 bp and nucleosomes were mapped to each bin. The zero point was the TSS. A nucleosome was said to map to a given bin if it completely overlapped with the 10 bp bin. Each bin was assigned the score of the overlapping nucleosome. In the cases where our algorithm detected overlapping positions for a nucleosome, and more than one nucleosome mapped to a single bin, the bin was assigned the highest score. Genes were separated into TATA-containing or TATA-less [34] and the average nucleosome profiles were generated for each group by averaging the scores for the bin across all the genes (4695 and 1074 promoters, respectively). Genes were similarly separated into the top 500 or bottom 500 with respect to transcription rates [109] and average profiles were plotted for these classes.

Nucleosome Positioning Periodicity (NPP) score and dinucleotide positioning profile

The NPP score was generated by calculating the similarity of the experimentally derived nucleosome profile over the coding region of every gene to an artificially generated profile where six nucleosomes of score 1.0 were regularly placed with 30 bp linker lengths. In general, genes with well-positioned nucleosomes had profiles that were most similar to the synthetic profile and hence, had a high NPP score. The first (+1) nucleosome downstream of the TSS is adjacent to a gap and is likely to be more strongly sequence dependent for positioning than a nucleosome that is flanked by other nucleosomes. We therefore derived AA/TT profiles from the sequence underlying the

first nucleosome. To derive high-confidence sequence profiles, we aligned all genes to the first nucleosome as shown in Fig. 4.7. We selected all +1 nucleosomes with a score ≥0.9 for the input set. Since nucleosomes show a dyad symmetry in terms of positioning over DNA, the reverse complement of each sequence in the input set was also included before calculating the profile. We calculated frequency profiles for the dinucleotides AA and TT, and summed and smoothed them using a 3 bp moving average. This high-confidence AA/TT profile was then correlated with the AA/TT profiles derived from all nucleosomes at the +1, +2, +3 and +4 positions.

Generation of nucleosome remodeling profiles and remodeling score

Genes that did not have 200 bp long promoter region were excluded for this analysis. For all of the genes that passed this filter, the difference between the nucleosome scores in normally grown cells and cells after heat-shock were calculated bin-wise from -400 bp upstream to +200 bp downstream of the start codon. For the plots and clusters shown in Fig. 4.10 and Fig. 4.11, we then created subsets of this data that included either genes that were activated by at least 2-fold, or genes that were repressed at least 2-fold by heat shock. To calculate the remodeling score, a 7-bin window, corresponding to a distance of 70 bp (approximately half of a nucleosome), was scanned along each profile and the individual bin scores were averaged for each window. The maximum window score in the positive direction across the entire profile was assigned as the remodeling score for nucleosome eviction while a similar maximum in the negative direction was assigned as the remodeling score for nucleosome appearance.

Increase in accessibility of TF binding sites after stress

TF motifs were mapped across the entire genome using position weight matrices derived from [17] by Patser [35] at a *P-value* cut-off of 0.01. These were considered the

putative binding sites while the functional ("true") binding sites were derived from published ChIP-chip data [16, 95]. A functional motif was considered to be occupied, and therefore not accessible, if it overlapped with a nucleosome that had a score ≥ 0.5. The occupancy of the chip-chip binding sites was compared to that of the putative motif binding sites and a hypergeometric distribution was used to calculate p-values. This analysis was done with data from both normal and heat-shock conditions. To calculate the significance of the change in binding site occupancy upon heat-shock, the *P-values* for the heat-shock nucleosome data were divided by the *P-values* derived from the normal condition data.

Chapter 5: MicroRNA Regulatory Networks in the Transition from Quiescence to Proliferation

INTRODUCTION

The G0 stage of the cell cycle, also termed quiescence, is a state of reversible cell cycle arrest wherein cells are poised to enter the cell cycle on external physiological stimuli The transition of mammalian cells from quiescence to proliferation is an important aspect of normal as well as cancer cell biology and is accompanied by the differential expression of hundreds of genes [110-112]. This reprogramming of gene expression is regulated by many transcription factors (TFs), many of which are known oncogenes or tumor suppressors [113]. Primary human fibroblasts provide an excellent model for studying the global gene expression programs regulating the transition from quiescence to proliferation. Gene expression programs active in fibroblasts exiting quiescence are very similar to programs active in wound healing and cells undergoing malignant transformation [114, 115]. It has also been shown that expression profiles of proliferating fibroblasts are good predictors of cancer progression [116].

The transition of quiescence to proliferation is mediated by many immediate-early TFs including Myc [111]. Myc is an established oncogene and, as an heterodimer with Max, regulates gene transcription by binding to E-box elements [117, 118]. The biological processes under regulation by Myc are diverse including cell-cycle progression, transformation and apoptosis [70, 119]. In addition to the transcriptional response accompanying proliferation, we have recently shown that the exit from quiescence activates several microRNAs (miRNAs) [111]. miRNAs are ~22 nucleotide endogenous non-coding RNAs that regulate gene expression at the post-transcriptional level by inhibiting translation or marking mRNAs for degradation [30, 49]. Thousands of

mammalian genes have been predicted to be under miRNA control and miRNAs have been implicated in many biological processes such as development, differentiation, proliferation, apoptosis, and tumorigenesis [31, 48, 54, 120]. It has been proposed that microRNAs are involved in fine-tuning the broader gene expression programs established by sequence-specific TFs [52].

Several studies have established that miRNAs and TFs are functionally linked in complex regulatory networks that govern cell proliferation and cancer [50, 53-55]. Recent studies have provided evidence that Myc may modulate its target gene expression at the post-transcriptional level by regulating several miRNAs. For example, Myc activates the mir-17-92 cluster that modulates E2F1 expression and Myc activated microRNAs are able to promote angiogenesis in tumors [50, 55].

Previous work in the Iyer lab had shown that the transition of quiescence to proliferation is accompanied by the activation of several miRNAs and many of these miRNAs are Myc responsive. Previous microRNA profiling and analysis performed by a former graduate student in the lab, Patrick Killion, had shown that the miRNA miR-22 was activated as early as 5 minute after serum stimulation and as also activated by Myc in HeLa cells. Additionally, ChIP-Seq data for Myc in HeLa cells (performed by Dr. Roger Liu) had revealed Myc binding sites upstream of the miR-22 start site. Patrick's analysis further showed conserved E-boxes overlapping with the ChIP-Seq signal as well as the presence of conserved binding motifs for other TFs such as SRF and STAT5. Both SRF and STAT5 are activated when quiescent fibroblasts are stimulated by serum. Though there were other miRNAs that showed similar profiles as miR-22, we focused on miR-22 for further characterization as there was very limited literature available on miR-22 and it was one of the strongest responders to serum stimulation.

RESULTS

Myc activates miR-22 on serum stimulation

We asked the question whether Myc activated miR-22 during the transition from quiescence to proliferation. We knocked-down Myc in proliferating fibroblasts and then induced them into quiescence. Quiescent fibroblasts were serum-stimulated and RNA was harvested 20 min and 2 hours after stimulation. Quantitative RT-PCR (qRT-PCR) for miR-22 expression showed that miR-22 activation is suppressed after Myc knock-down at both time-points as compared to the negative control (Fig. 5.1). This clearly showed that Myc, at least in part, is responsible for activating miR-22 during the transition from quiescence to proliferation.

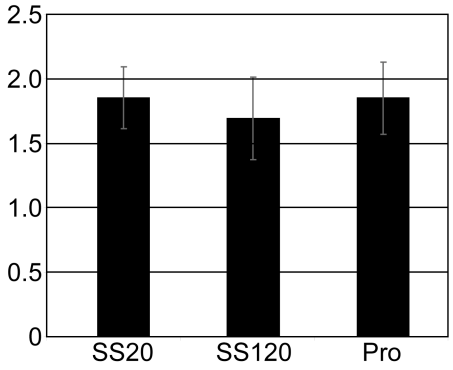


Figure 5.1 Myc regulates miR-22

Bar plot shows quantitative RT-PCR data for miR-22 expression. SS20: 20 minutes after serum stimulation, SS120: 120 minutes post serum stimulation, after Myc knockdown. Pro: miR-22 expression 24 hours after Myc knockdown in normally proliferating cells. RPL21 was used as a loading control. Values shown are averages for 4 technical replicates. Error bars indicate 2 standard deviations from the mean.

miR-22 targets in proliferating primary fibroblasts

Since microRNAs mainly cause suppression of their target genes subsequently resulting in further degradation of the transcripts, genes that show downregulation under

elevated levels of a given microRNA are likely to be direct targets and can be detected by microarrays [121-124]. To identify miR-22 targets in proliferating fibroblasts, we transfected synthetic miR-22 duplexes into human foreskin fibroblasts that were growing under media supplemented with 10% serum. MicroRNA duplexes were transfected at a final concentration of 100 nM and total RNA was harvested 24 hrs post-transfection, amplified, labeled and hybridized on custom cDNA microarrays that had ~47,000 probes corresponding to ~31,000 unique Unigene clusters. In order to investigate whether our experimental setup was able to detect miR-22 targets, we wanted to see whether the miR-22 "seed" i.e. nucleotide 2-8 from the 5' end of the microRNA was enriched in the downregulated transcripts. Additionally, we wanted to extend this analysis to all possible contiguous 6mers in the miR-22 mature sequence. We first ranked genes downregulated in the miR-22 transfected sample in order of maximum to minimum repression. We then sampled genes from the top of the list iteratively such that the first iteration would have the top 50 genes, the second iteration would have the top 100 genes, the third iteration would have the top 150 genes and so forth. In each iteration, we calculated the enrichment for all 6mers over background assuming a hypergeometric null model, where the background considered was all genes represented on the array. We also performed the same experiment and analysis in HeLa cells. In both cell lines, there was a clear enrichment of the miR-22 seed match in downregulated 3'UTRs over background (Fig. 5.2A and 5.2B). Moreover, this enrichment over background was statistically significant (*P-value* < 10^{-8} assuming a binomial model, Fig. 5.3A and 5.3B) in both cell lines. Thus, genes whose mRNAs were downregulated in response to high levels of miR-22 contained a significant proportion of direct targets of miR-22. Functional annotation analysis of the top 150 downregulated genes showed significant enrichment for genes involved in apoptosis (*P-value* < 0.003 for fibroblasts and < 0.03 for HeLa cells). This was consistent

with the expected pro-proliferative role of miR-22 and suggested that the observed activation of miR-22 as early as 5 min after serum stimulation may have definite functional significance.

miR-22 suppresses the interferon response under quiescence

An interesting hypothesis concerning the early activation of miR-22 after serum stimulation is that miR-22 is responsible, at least in part, to the inactivation of the quiescent state and thus enabling the cell to exit G0. Overexpression of miR-22 under quiescence would be expected to repress cell cycle inhibitory genes and promote genes favoring cell cycle exit. To investigate this hypothesis, we transfected miR-22 duplexes into fibroblasts that were induced into a state of quiescence by serum deprivation and profiled gene expression changes by cDNA microarrays as described above. The seed enrichment analysis described above showed that a significant proportion of the repressed genes were indeed direct targets (Fig. 5.2C & 5.3C). However, the enrichment was much lower than that observed for genes that were repressed under proliferation, indicating a much higher proportion of indirect targets. We found that 25 out of the top 50 genes downregulated by miR-22 in quiescent fibroblasts were interferon inducible genes (Fig. 5.4A). Most of the repressed genes that belonged to the interferon response pathway did not have any seed match with miR-22. This indicated that the observed repression of these genes was through an indirect mechanism. Accordingly, the seed enrichment improved considerably if the interferon pathway genes were excluded from the analysis (Fig. 5.4B).

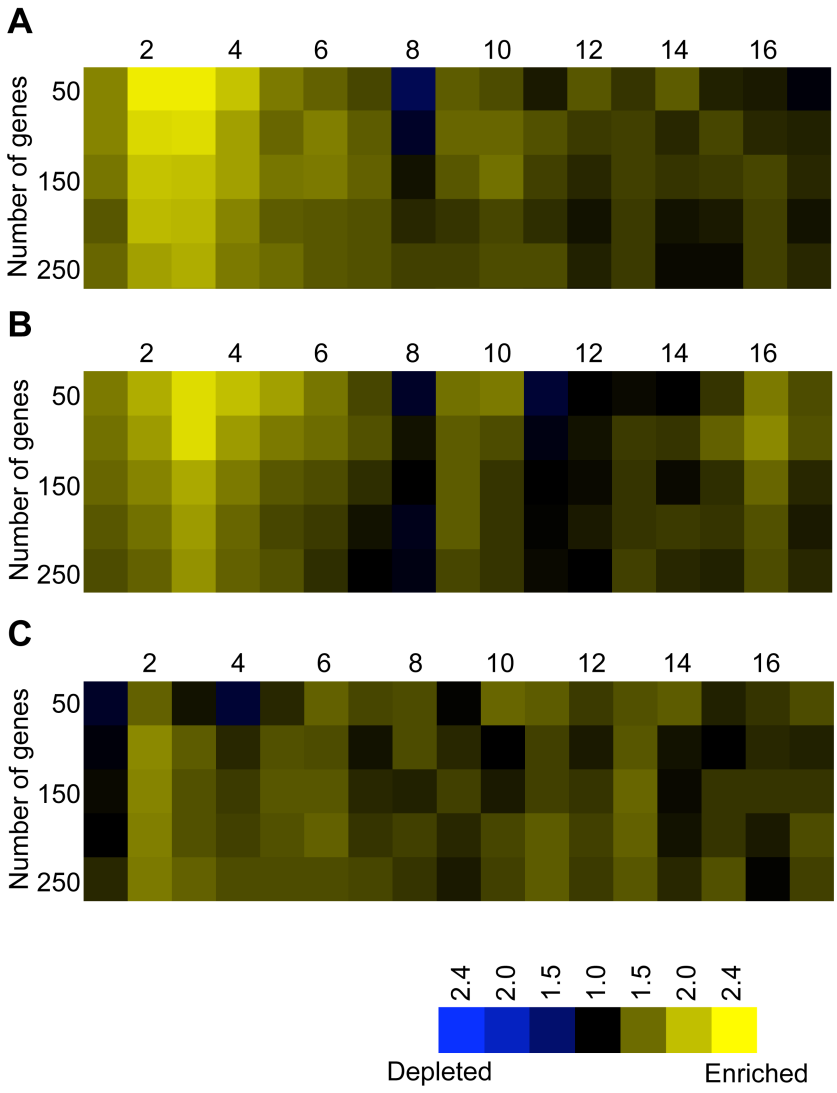


Figure 5.2 Seed enrichment analysis of miR-22 in repressed genes

Seed enrichment analysis for genes repressed by miR-22 in (A) proliferating fibroblasts (B) HeLa cells and (C) quiescent fibroblasts. miR-22 transfections were carried out at 100 nM final concentration and gene expression assayed by cDNA microarrays. Genes were ranked from most to least repressed and miR-22 seed enrichment was calculated in a cumulative bin size of 50 genes. Enrichment values were calculated for each 6 mer along the microRNA with yellow indicating enrichment while blue indicating depletion. Each value on the horizontal axis indicates the start position of a 6 mer from the 5' end of the microRNA. 6 mers starting at positions 2-4 were enriched over background.

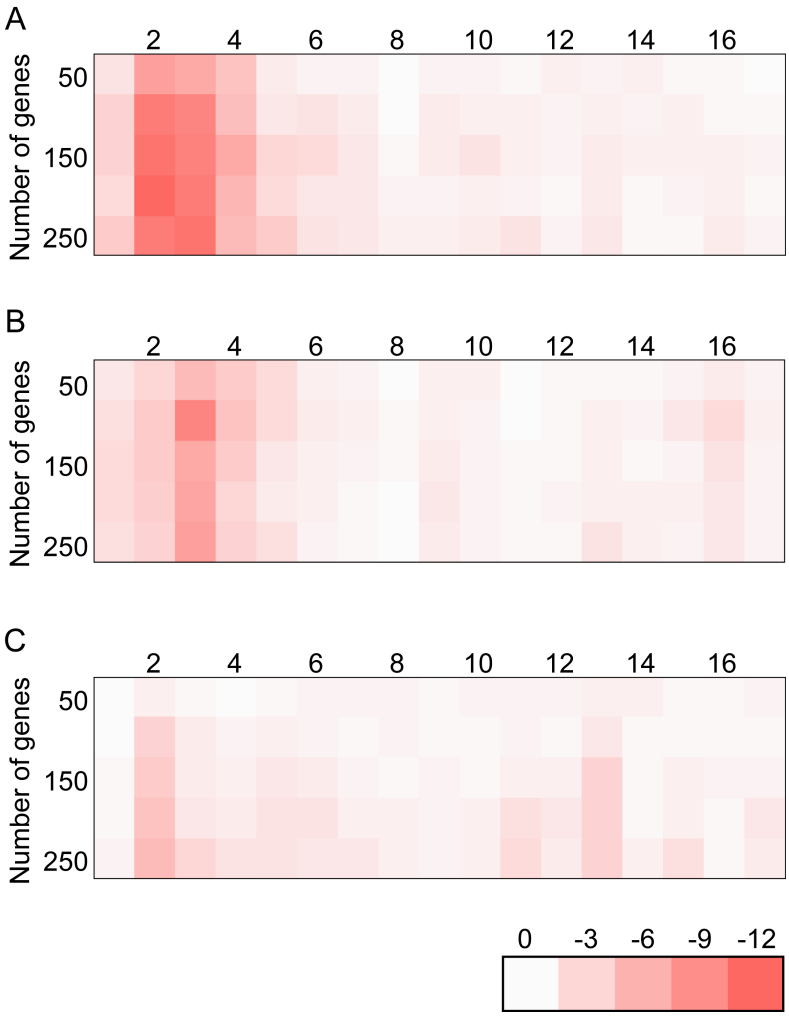


Figure 5.3 Seed enrichment analysis showing *P-values*

Significance for the enrichment of each 6-mer was calculated assuming a binomial distribution. Values reported are -1*log10(*P-values*) and are color coded as indicated.

The results so far suggested that an interferon response was active under quiescence and miR-22 was able to repress it. Previous studies profiling gene expression in quiescent fibroblasts have documented an interferon response that is activated as primary fibroblasts enter quiescence [112, 125]. To further confirm that an interferon response is activated under quiescence in our studies, we compared mRNA expression of a subset of interferon-stimulated genes (ISGs) by RT-PCR between proliferating and quiescent fibroblasts. As shown in Fig. 5.4C, all the assayed ISGs were upregulated under quiescence.

miR-22 suppresses poly I:C mediated type I IFN response

To further characterize miR-22 mediated suppression of the interferon pathway, we sought to artificially induce the IFN response in HeLa cells using poly I:C. Poly I:C is a double stranded RNA polymer composed of riboinosinic and ribocytidilic acid subunits that simulates a viral infection and induces a strong type I IFN response [126]. We co-transfected miR-22 duplexes and poly I:C molecules into HeLa cells and harvested cells at 6 hrs and 12 hrs post-transfection. Co-transfection of poly I:C with a control siRNA from Ambion that was not supposed to target any gene in human cells was used as a negative control. Induction of type I interferon response was measured by assaying interferon-beta (IFNB1) mRNA levels by RT-PCR and STAT1 phosphorylation by western blot. Co-transfection of miR-22 caused decreased IFNB1 transcription and concomitant decreased STAT1 phosphorylation as early as 6 hrs post-induction (Fig. 5.5A and 5.5B). The suppression was even more pronounced at 12 hrs post-induction. Additionally, miR-22 transfected cells clearly showed decreased apoptosis as compared to the negative control (Fig. 5.5C). These results strongly indicate that miR-22 is able to suppress the poly I:C induced type I interferon pathway.

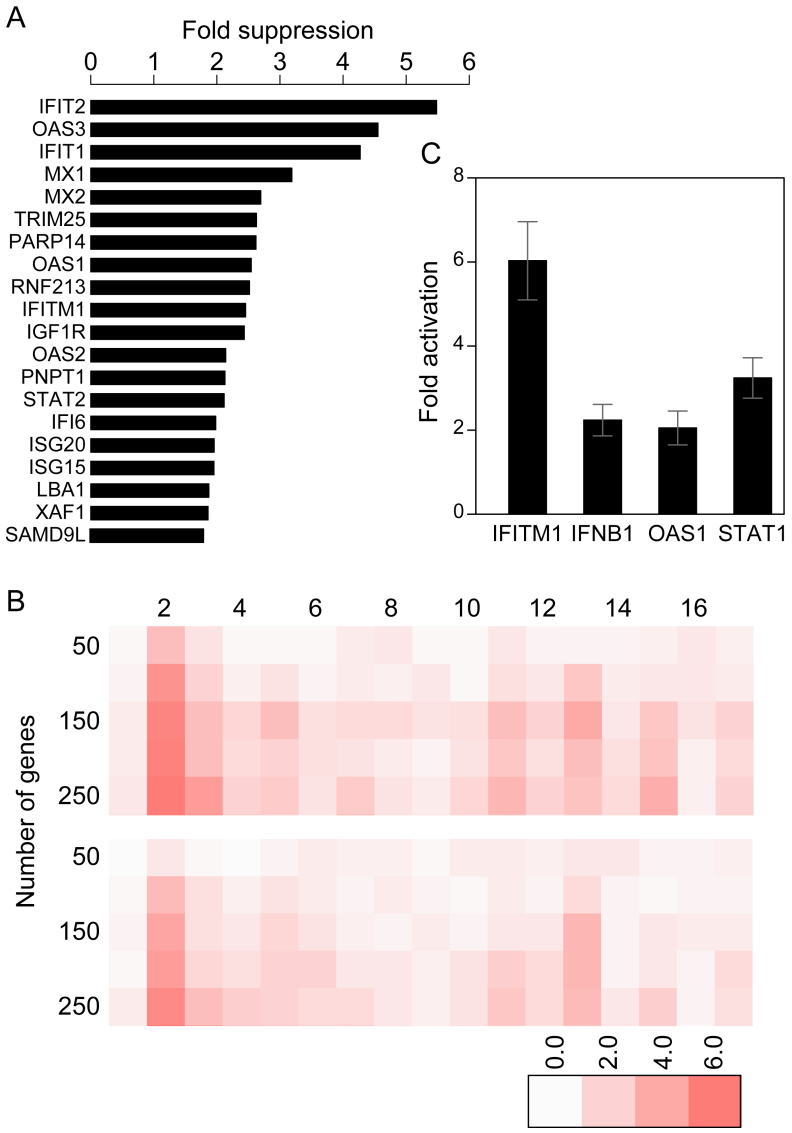


Figure 5.4 miR-22 inhibits the IFN pathway under quiescence.

(A) Bar chart shows microarray data for the top 20 interferon-stimulated genes (ISGs) that were repressed by miR-22 under quiescence. (B) Seed enrichment for the set of genes suppressed by miR-22 under quiescence without excluding the IFN responsive genes (lower panel) and after excluding IFN responsive genes (upper panel). Removal of interferon response genes enhanced the seed enrichments for 6 mers at positions 2-4. This suggested that the widespread interferon response suppression was largely an indirect effect. (C) The interferon pathway is active under quiescence. Quantitative RT-PCR data for interferon-beta and other ISGs shows upregulation of all transcripts under quiescence. Values shown are averages for 2 technical replicates. Error bars indicate 2 standard deviations from the mean.

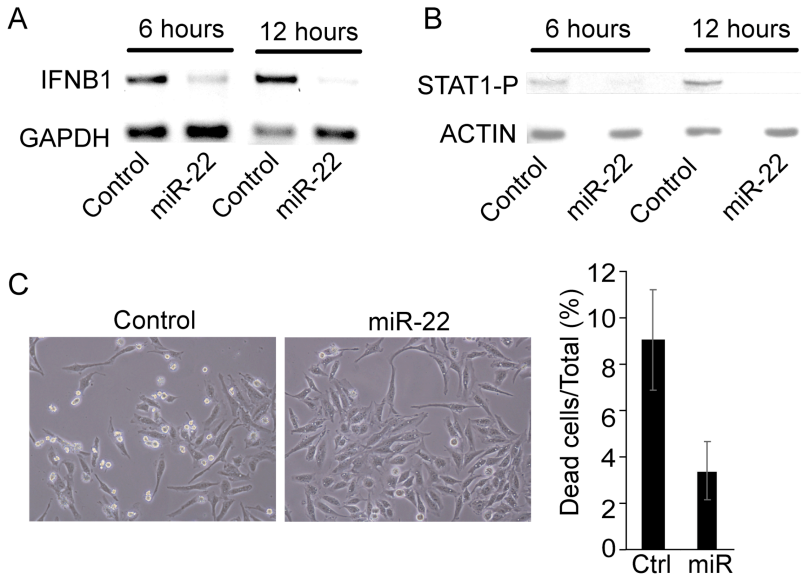


Figure 5.5 miR-22 suppresses poly I:C mediated interferon response

Poly I:C was used to elicit a type I interferon response in HeLa cells and the effect of miR-22 on the induced interferon response was investigated by co-transfecting either miR-22 or a negative control synthetic duplex siRNA that does not target any known human gene. miR-22 suppression of the poly I:C induced interferon pathway was assayed by measuring (A) inhibition of interferon-beta (IFNB1) transcription, using RT-PCR. GAPDH was used as a loading control and (B) Inhibition of STAT1 phosphorylation by miR-22, measured by immunoblotting. Actin was used as a loading control. (E) Inhibition of interferon-induced cell death by miR-22. The graph on right shows quantitation, where dead cells were counted using Trypan blue dye exclusion. "Ctrl" refers to the control siRNA. Values shown are averages for 3 independent transfections. Error bars indicate 2 standard deviations from the mean.

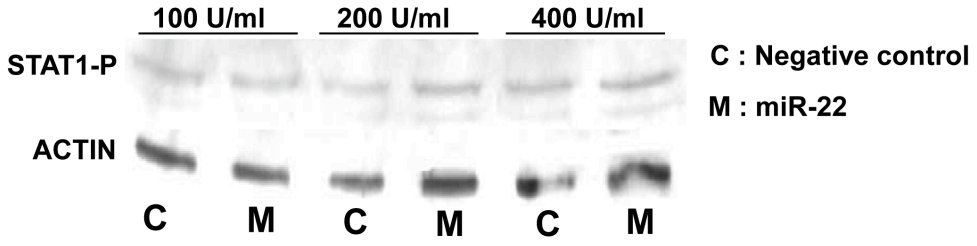


Figure 5.6 miR-22 act upstream of the JAK-STAT pathway

Poly I:C mediated type I interferon response amplification loop was bypassed by directly stimulating HeLa cells with varying amounts of recombinant type I interferon. miR-22 transfection was unable to prevent STAT1 phosphorylation indicating that miR-22 mediated blockade of the interferon signaling cascade was upstream of IFNB1 transcription.

Mechanism of miR-22 mediated IFN suppression

Activation of the IFNB1 promoter by poly I:C requires the co-coordinated activation of IRF3/IRF7, ATF2/c-Jun and NF-kappaB transcription factor complexes [127]. IFNB1 is secreted out of the cell, binds to the IFN receptors and activates the JAK-STAT pathway in a paracrine and autocrine fashion. Activation of the JAK-STAT pathway upregulates IRF7, that along with IRF3 participates in an amplification loop that contributes to further activation of the IFNB1 promoter [128]. We considered two possible scenarios for the mode of action of miR-22 mediated suppression of IFNB1 transcription. One possibility was that miR-22 acts upstream of IFNB1 while another possibility, which was not mutually exclusive, was that miR-22 acts downstream of IFNB1, and the downregulation of IFNB1 is due to inhibition of the feedback loop. To distinguish between these two possibilities, we sought to bypass the poly I:C induced IFNB1 transcriptional activation by directly stimulating the JAK-STAT pathway. HeLa cells were first transfected with miR-22 duplexes and 6 hrs post-transfection, cells were

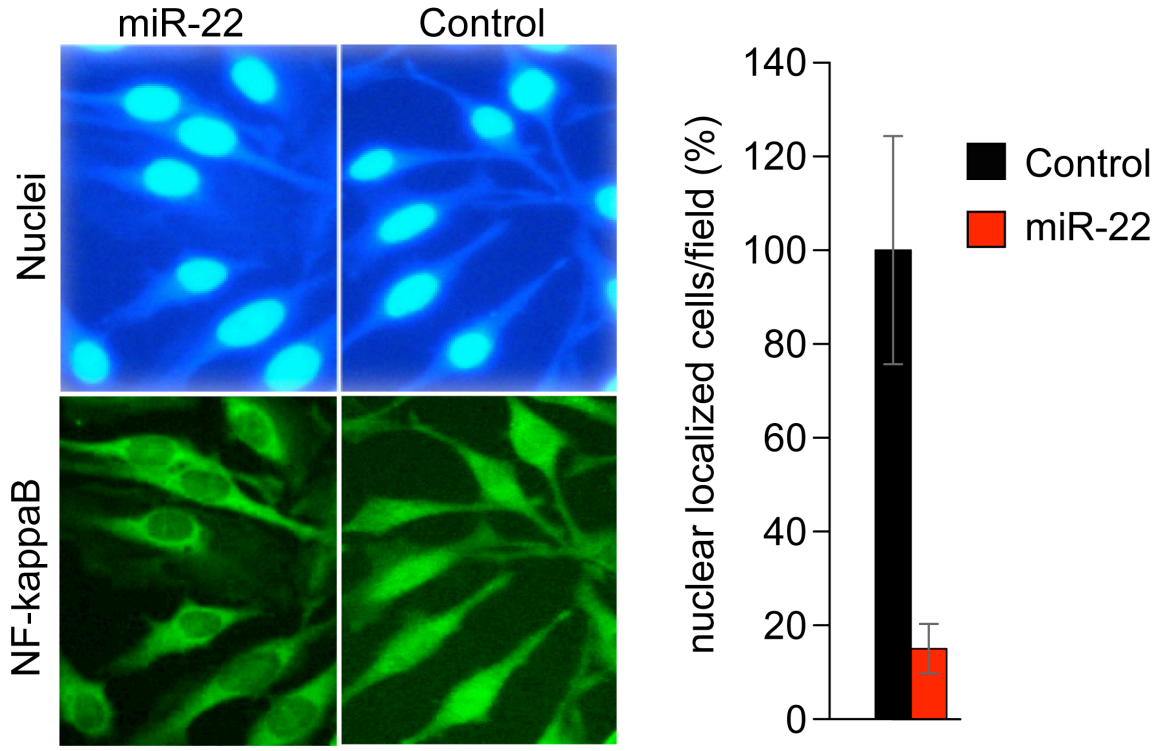


Figure 5.7 miR-22 inhibits NF-kappaB nuclear localization

(A) miR-22 impairs NF-kappaB nuclear localization. Nuclear localization of NF-kappaB nuclear was assayed by immunofluorescence for the p65 subunit. Bar plot shows the average percentage of cells per field showing nuclear localization. Averages were calculated from three randomly chosen fields and normalized to values obtained for the negative control (100%). Error bars show 2 standard deviations from the mean.

stimulated with increasing amounts of recombinant universal type I IFN for 12 hrs. JAK-STAT pathway activation was assayed by performing western blots for STAT1 phosphorylation. As shown in Fig. 5.6, miR-22 was unable to suppress STAT1 phosphorylation when the JAK-STAT pathway was directly activated. This strongly suggested that miR-22 acted upstream of the IFN receptor. However, we cannot rule out the possibility that miR-22 also acts downstream of the JAK-STAT pathway.

As stated above, 3 distinct events precede and are required for IFNB1 transcriptional activation, namely 1) IRF3 and IRF7 are phosphorylated by the IKBKE/TBK1 kinases, dimerize and translocate to the nucleus [127]. 2) The AP-1 family transcription factors ATF2 and c-JUN get phosphorylated via the MAP kinase pathway, dimerize and are retained in the nucleus [129, 130]. 3) The NF-kappaB inhibitory IKB molecules are marked for degradation by the inhibitor of NF-kappaB kinase protein complex releasing NF-kappaB that localizes to the nucleus [131]. Binding of all above mentioned transcription factor complexes is required for activation of the IFNB1 promoter. We investigated each one of the three events individually by assaying for IRF3 and ATF2 phosphorylation by western blot (WB) and NF-kappaB nuclear localization by immuno-fluorescence (IF) 3 hrs and 6 hrs after we co-transfected HeLa cells with miR-22 and poly I:C. Western blot assay showed no change in IRF3 and ATF2 phosphorylation between miR-22 transfections and negative control (data not shown). However, NF-kappaB IF showed a significant impairment of nuclear localization in miR-22 transfected cells as compared to negative control at 3 hrs (Fig. 5.7) and 6hrs (data not shown) after poly I:C stimulation. This showed that miR-22 suppressed the type I IFN by preventing NF-kappaB nuclear localization. The degradation of IkB proteins, which is required for NF-kappaB nuclear translocation, is subsequent to their phosphorylation by the kinase complex consisting of IKBKB, IKBKA and IKBKG [132]. The kinases IKBKB and

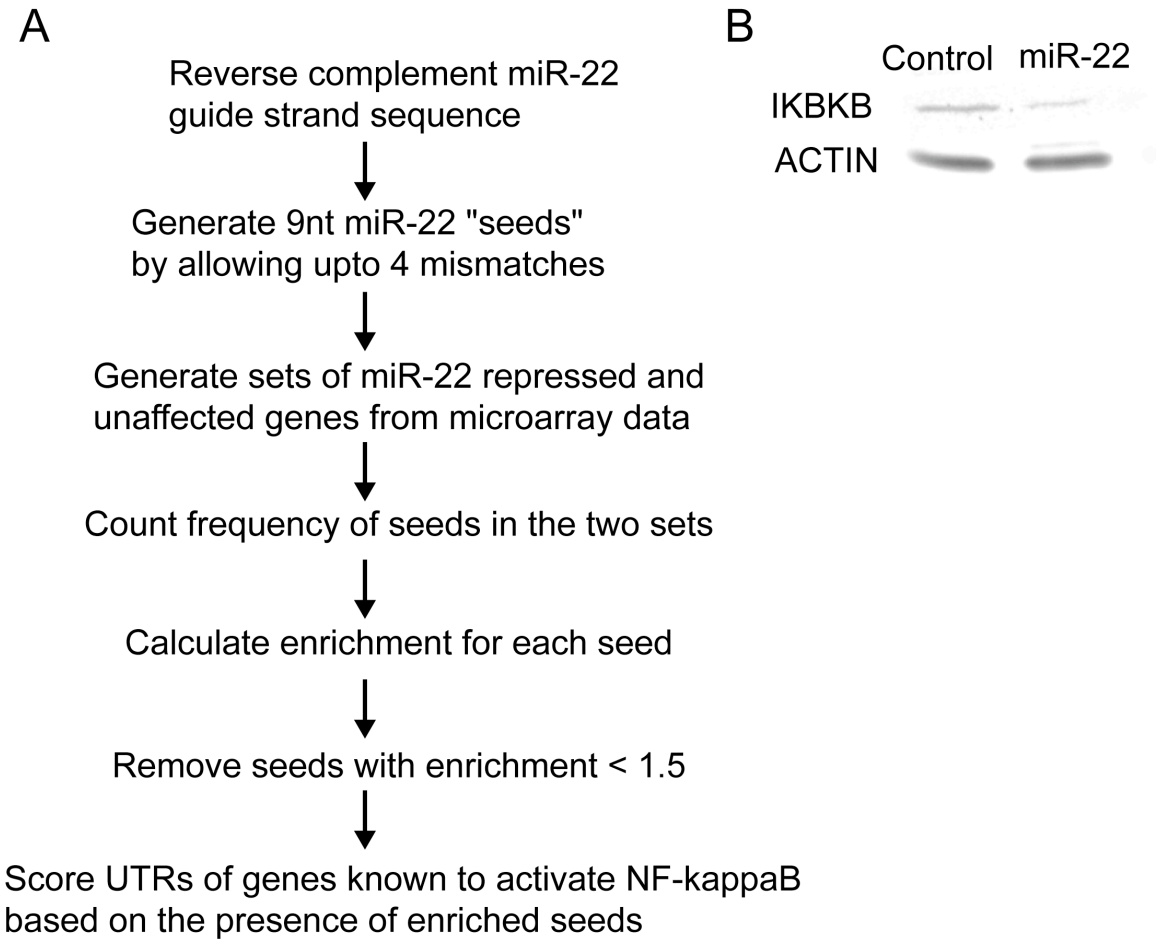


Figure 5.8 miR-22 inhibits IKBKB expression

(A) Flow chart to score 3' UTRs as potential miR-22 targets. (B) Western blot for IKBKB. HeLa cells were transfected with miR-22 duplexes or negative control siRNA and cells were harvested 24 hours later for immunoblotting. Actin was used as a loading control.

IKBKG showed significant seed matches to miR-22 (Fig. 5.8A) (see Discussion). Western blot assays showed that miR-22 transfection caused a modest decrease in protein expression of IKBKB (Fig. 5.8B). However, we were unable to confirm a direct interaction between miR-22 and the IKBKB 3'UTR by luciferase assay.

miR-22 targets cell cycle arrest inducers and pro-apoptotic genes

In addition to the interferon pathway, miR-22 also downregulated cell cycle arrest and apoptotic genes, which could account in part for its anti-apoptotic effects. In our microarray data, TP53INP1 showed greater than 1.5 fold downregulation by miR-22 in fibroblasts while DDIT4 showed a similar effect in HeLa cells. TP53INP1 is a downstream transcriptional target of p53 and has a functional role in causing cell cycle arrest [133] whereas DDIT4 is a DNA-damage inducible protein known to be involved in apoptosis and is also a direct target of p53 [134]. We confirmed both TP53INP1 and DDIT4 as being direct targets of miR-22 by luciferase assays (Fig. 5.9A).

miR-22-Myc feedback network

MXD4 is a transcriptional repressor of Myc and is in turn repressed by the Myc-MIZ1 complex [135, 136]. We found MXD4 gene expression to be repressed in miR-22 transfected HeLa cells as well as primary fibroblasts and luciferase assays showed MXD4 to be a direct target of miR-22 (Fig. 5.8A). This reveals a novel feedback loop where Myc activates miR-22 to suppress MXD4, which in turn downregulates Myc expression (Fig. 5.9B).

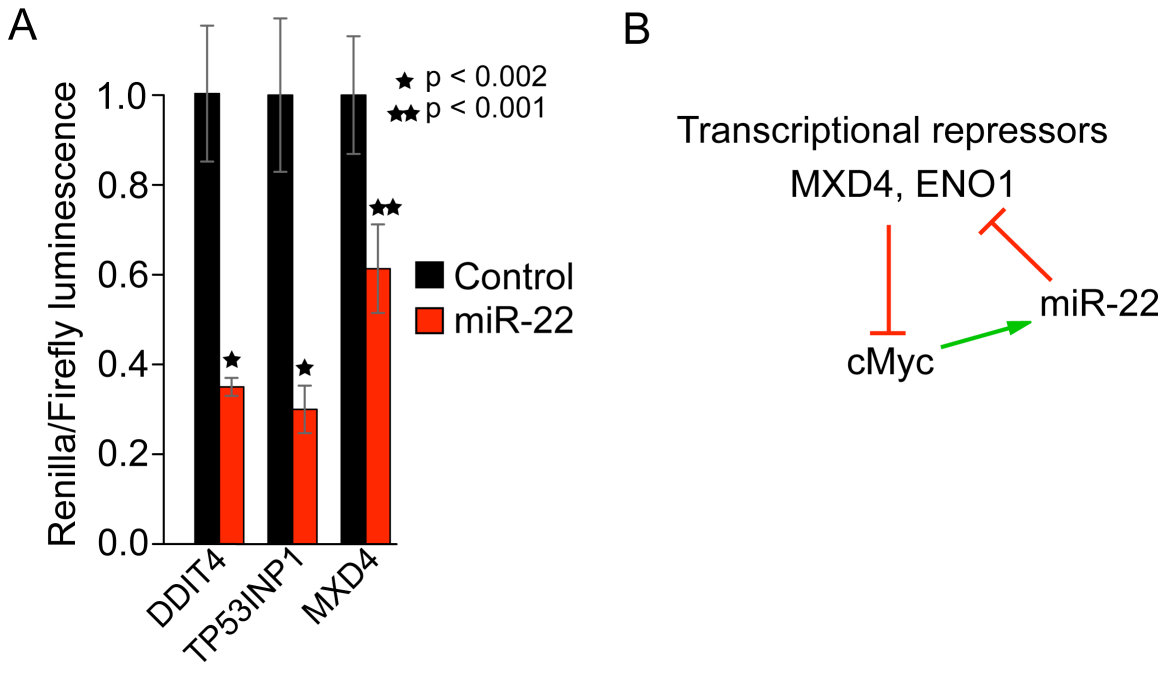


Figure 5.9 miR-22 targets cell-cycle arrest and pro-apoptotic proteins

(A) Luciferase assay shows that miR-22 downregulates TP53INP1, DDIT4 and MXD4. Target gene UTRs were cloned downstream of Renilla luciferase and luminescence values were normalized to Firefly luciferase. (B) Schematic of the Myc-MXD4-miR-2 feedback module.

DISCUSSION

A TF-miRNA module for re-entry into the cell-cycle

The transition of primary human cells from quiescence to proliferation is hallmarked by the activation of multiple early response transcription factors (TFs) like Myc, SRF, E2F and others, many of which are oncogenic factors [113]. Earlier work in the Iyer lab had shown that accompanying this strong transcriptional response is a concomitant expression program of microRNAs [111]. Recently, it has been shown that microRNAs can regulate, and in turn are regulated by, sequence-specific TFs to generate complex TF-miRNA regulatory networks [137, 138]. We have characterized a TF-

miRNA network that is activated as primary fibroblasts prepare to exit quiescence and enter a stage of rapid proliferation. Our results show that the oncogenic TF Myc activates miR-22 on serum stimulation (SS) of quiescent primary fibroblasts (Fig. 5.1). This activation is a direct consequence of Myc binding upstream to the miR-22 start site as shown by ChIP-seq binding data. In spite of the fact that Myc mRNA levels increase almost 1 hour after cells exit quiescence [110], Myc mediated activation of miR-22 occurs as early as 5 min and peaks at 20 min post-SS [139]. This suggests that Myc may be bound to the miR-22 promoter under quiescence and is poised to activate the promoter on receiving the appropriate stimulus. The early activation of miR-22 as cells prepare to renter the cell cycle may be explained by its effect on the interferon response. It has been observed previously [112, 125] and in our data that the quiescence of primary fibroblasts is accompanied by an activation of the interferon response, namely by upregulating TFs like IRF1, IRF7 and STAT1 and activation of interferon stimulated genes (ISGs). Though the exact role of the interferon response under quiescence remains unknown, it is possible that it maybe required for maintaining or inducing a state of cell cycle arrest. It has been suggested previously that quiescence is not simply a program of cell-cycle arrest and there may be signaling pathways engaged to maintain the cell in a viable and reversibly arrested stage [125]. Consequently, in order to exit G0, it is to be expected that the cell must somehow overcome this inhibitory influence on proliferation. We have shown that one of the means by which primary fibroblasts may accomplish this is by activating a suppressor of the IFN pathway i.e. the microRNA miR-22. miR-22 overexpression (by transient transfection) in cells induced to activate the type I IFN pathway showed a marked decrease in IFNB1 transcription and STAT1 phosphorylation (Fig. 5.5A and 5.5B), known indicators of IFN pathway activity. Consistent with its role as an interferon blocker, miR-22 expression was downregulated in psoriasis, a chronic inflammatory skin

disease [140]. Recently, viruses have been found to encode miRNAs that help evade the immune response and establish successful viral replication in the host [141, 142]. Elevated levels of miR-22 have been observed in hepatitis-B and hepatitis–C virus infected liver cells as compared to uninfected cells [143, 144]. Whether miR-22 activation under viral infection is merely a stress response or is required to bypass the interferon pathway remains to be seen. Additionally, miR-22 was found to target pro-apoptotic and cell cycle arrest proteins (Fig. 5.9A). Accordingly, miR-22 has been found to be highly expressed in self-renewing mammary progenitor cells, which supports its role as an anti-apoptotic molecule [145].

A 2008 study investigating Myc regulated miRNAs showed that Myc inhibits miR-22 expression in immortalized human B cells [146]. We have shown that Myc activates miR-22 in the primary foreskin fibroblast cell line 2091 while previous work in the Iyer lab has shown that miR-22 is activated by Myc in HeLa cells [139]. It is possible that Myc regulation of miR-22 is context dependent and cell-type specific. Recently, it was shown that miR-22 indirectly promotes upregulation of IL1B, a known inflammatory cytokine, in chondrocytes [147]. This data, seemingly in contradiction to our own, may be due to differences at the tissue level. It cannot be ruled out that miR-22 maybe capable of suppressing the IFN pathway but indirectly inducing other inflammatory cytokines like IL1B. Physiologically, this is possible in the context of fine-tuning the immune response as an over-enthusiastic secretion of inflammatory cytokines maybe deleterious at the tissue level.

Mechanism of action of miR-22

IFNB1 promoter activation requires the co-coordinated activity of IRF3/IRF7, ATF2/c-JUN and NF-kappaB complex of transcription factors. Our data shows that miR-22 specifically impairs the nuclear translocation NF-kappaB (Fig. 5.7) while leaving

IRF3/IRF7 and ATF2/c-Jun activation intact. How does miR-22 suppress NF-kappaB nuclear localization? To answer this question, we compiled a list of genes that can potentially mediate NF-kappaB nuclear localization. Most of these genes were from a previously published gain-of-function screen where the authors overexpressed full-length human and mouse cDNAs and assayed for NF-kappaB activation by reporter assays [148]. We also performed extensive literature surveys and added genes shown to be required for NF-kappaB activation that were not identified in the above-mentioned screen. We scored each gene as a potential miR-22 target depending on the presence of functional miR-22 seed sequences in the 3'UTR. We generated a list of miR-22 seed sequences where we allowed a maximum of 4 mismatches in the first 9 nucleotides of the miR-22 guide strand. Examples of seed sequences in our list would be: TNGCAGCTN, TGGNAGNTT etc. We counted the frequency of these 9mers in the 3'UTRs of the top 150 miR-22 repressed genes (C_rep) and genes that did not show any change in expression after miR-22 transfection (C_unchg). The ratio of C_rep/C_unchg was calculated as the enrichment of each 9mer seed and seeds that had an enrichment of less than 1.5 were eliminated as not significant. Each of the 21 remaining 9mer seeds was assigned a score that was calculated as 1 – 1/enrichment. Higher the enrichment, closer the score is to 1. We matched the significant seed sequences to 3'UTRs strictly in the 5 to 3' direction allowing for multiple matches. If two seed sequences were less than 22 nucleotides apart, the weaker enriched seed was removed. The 3'UTR was now assigned a score that was calculated as:

$1 - \Pi(1 - Si)$ where S indicates the score of a given seed sequence and i iterates over the total number of seed sequences found in the 3'UTR.

The top 20 genes included known NF-kappaB inducers like IKBKB, IRAK1, RIPK1 and members of the MAPK signaling pathway like MAP3K3 and MAP3K1 (Appendix F).

Nuclear localization of NF-kappaB is dependent on the degradation of IkB proteins, which in turn is subsequent to their phosphorylation by the kinase complex consisting of IKBKB, IKBKA and IKBKG. Though IKBKB showed significant seed matches to miR-22 by our analysis scheme and IKBKB protein expression was reduced by miR-22, IKBKB is not directly targeted by miR-22 as was shown by luciferase assays. Upstream of the IKBKB/IKBKG/IKBKA complex is the receptor interacting protein kinase-1 (RIPK1) kinase that interacts with several other proteins such as IRAK1, FADD and members of the MAPK pathway to activate downstream signaling. All the above-mentioned proteins were assigned significantly high scores by our scoring scheme indicating that these genes possessed multiple enriched miR-22 seed sequences in their 3'UTRs, indicating that these maybe likely targets. Interestingly, MAP3K3 was predicted as a miR-22 target by two miRNA target prediction algorithms, TargetScan Release 4.2 (April 2008) [149] and PicTAR [150]. Whether MAP3K3 is a direct target of miR-22 remains to be seen. In this study, we were unable to determine whether miR-22 effects its suppression of the IFN pathway by inhibiting any of the above mentioned proteins, most of which have been shown to be critical for NF-kappaB activation [151].

Cross-talk and feedback loops

In addition to directly targeting the IFN pathway, miR-22 also suppresses genes that can induce cell cycle arrest or apoptosis. TP53INP1, DDIT4, ZDHHC16, PMAIP1, ENO1 [133, 134, 152-154] are examples of proteins that are pro-apoptotic or function as cell cycle inhibitors and that were downregulated on miR-22 transfections as seen by

expression arrays. Interestingly, TP53INP1, DDIT4 and PMAIP1 are transcriptionally activated by the tumor suppressor p53 [133, 134, 153]. Previous studies have shown a

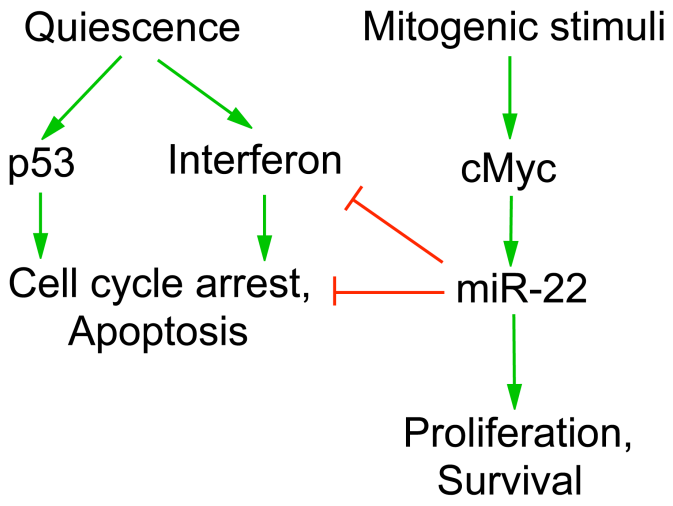


Figure 5.10 Myc-miR-22 proliferation network

Mitogenic stimuli (i.e. serum stimulation) activate Myc, which in turn activates miR-22 that suppresses the p53 and interferon mediated cell cycle arrest pathways enabling quiescent cells to exit G0.

p53 dependent suppression of Myc at the transcriptional level [155-157]. Our data, thus, reveals a novel cross-talk between the p53 and Myc regulatory networks. We have also shown that miR-22 participates in feedback loops that involve Myc. Myc is repressed by the transcription factor MXD4 which in turn is activated by another transcription factor MIZ1. Myc forms an inhibitory complex with MIZ1, thereby preventing MIZ1 mediated activation of MXD4 [136]. Our results show that, in addition to transcriptionally repressing MXD4, Myc activates miR-22 to suppress MXD4 at the post-transcriptional level. Another example of such feedback control involved the protein ENO1, a transcriptional repressor of Myc [158], which was also found to be downregulated after miR-22 transfection as assayed by expression arrays.

Our results describe a Myc-miR-22 driven proliferation network that is activated as primary cells prepare to exit quiescence (Fig. 5.10). As cells enter quiescence, inhibitory inputs from the p53 and interferon anti-proliferative pathways render the cells into a reversibly arrested stage. Mitogenic stimuli induce Myc to activate miR-22 that antagonizes the inhibitory effects on proliferation thereby enabling quiescent cells to re-enter the cell cycle. This study provides evidence that microRNAs upregulated during the transition of quiescent fibroblasts into a proliferative state have a defined functional role in reprogramming gene expression to enable the transition of G0 arrested cells into the cycling G1 stage. Further, our results provide additional evidence of the complex interplay between TFs and miRNAs to transduce extracellular signals into physiological responses.

MATERIALS AND METHODS

Normal cell culture conditions

HeLa cells and primary fibroblasts were maintained in DMEM (Dulbecco's Modified Eagle's Medium) supplemented with 10% FBS at 37C under 5% CO2. 2091 fibroblasts were made quiescent by first growing them under normal conditions until 50% confluent, then replacing medium with DMEM supplemented with 0.1% FBS and growing them for further 48 hours.

Rendering primary fibroblasts quiescent

Fibroblast cell cultures were maintained under normal cell culture conditions until 40% confluent. Medium was removed and cell cultures were washed 3x with PBS (Phosphate Buffered Saline). Replacement medium was DMEM supplemented with 0.1% FBS. Cell cultures were further maintained at 37°C for 48 hours.

Myc knock-down

Myc-specific siRNA and negative control siRNA was purchased from Dharmacon. HeLa cell cultures were grown under normal cell culture conditions. 6-well plates were seeded with 1.5 x 105 cells / well. Cell cultures were allowed to grow for 24 hours. Cell cultures were transiently lipotransfected with Invitrogen Lipofectamine 2000 according to the manufacturer's protocol (for siRNA transfection). Cell cultures were grown under normal cell culture conditions for 48 hours and then harvested for total RNA.

Real Time miRNA PCR

Patrick Killion performed the real time PCR for miR-22 expression. Quantitative real-time PCR was performed for miR-22 using Applied Biosystems TaqMan MiRNA Assays according to the manufacturer's protocol. miRNA-specific primers were provided in separate kits by the manufacturer for each of the miRNAs assayed as well as RPL21 the endogenous loading control.

Quantitative Reverse-Transcription PCR

RNA from quiescent and proliferating fibroblasts was extracted with the TRIzol reagent (Invitrogen) and reverse transcribed using random hexamers and the Superscript II system from Invitrogen. PCR was performed using the SYBR GREEN PCR Master Mix from Applied Biosystems. The target gene mRNA expression was normalized to the expression of GAPDH and relative mRNA fold changes were calculated by the ΔΔCt method. Primer sequences have been given in Appendix D.

miR-22 transfections

miR-22 guide and anti-guide mature sequences were obtained from miRBase (http://microRNA.sanger.ac.uk/sequences/) and the corresponding RNA oligos were

ordered from Invitrogen. Guide and anti-guide oligos were annealed in RNA annealing buffer (20 mM HEPES, pH 7.3, 50 mM KCl, 2 mM MgCl2) and the RNA duplex was transfected at a final concentration of 100 nM using lipofectamine 2000 according to manufacturer's instructions. Poly I:C was obtained from Sigma Aldrich and co-transfected at a final concentration of 200 ng/ml with miR-22 duplexes as described above. Sequences for siRNA against GFP were obtained from (Katome et al. 2003) and the RNA oligos were purchased from Invitrogen. siRNA guide and anti-guide strands were annealed and transfected as described above.

IFN stimulation

Recombinant universal type I interferon was obtained from PBL interferon source. HeLa cells were first transfected with miR-22 duplexes as described above and 6 hours post-transfection, cells were stimulated with 100 U/ml, 200 U/ml and 400 U/ml IFN. Cells were harvest 12 hours after IFN treatment for western blot analysis.

Western blots

Cell lysates were separated on 10% SDS-PAGE gels and proteins were transferred onto PVDF membranes. Membranes were blocked with 5% milk in TBST (25 mM Tris pH 8.0, 150mM NaCl, 0.05% Tween-20) and probed with corresponding primary antibodies against specific proteins (phosphorylated STAT1: Cell Signaling Technology, IKBKB: Santa Cruz biotech, MAP3K3: Epistomics). HRP-conjugated secondary antibodies were used to detect primary antibodies and proteins were visualized by chemiluminescence.

Immunofluorescence

HeLa cells were grown to 70% confluency under normal growth conditions and co-transfected with miR-22 and poly I:C as described above. 3 hours and 6 hours post-

transfection cells were fixed in 2% formaldehyde for 10 min, washed with PBS and permeablized with ice-cold methanol by incubating at -20 C for 15 min. After washing with PBS, cells were blocked with 5% donkey serum in PBS containing 0.03% Tween-20 (PBS-T) and incubated with an antibody against the p65 subunit of NF-kappa B (Santa Cruz biotech) diluted in PBS-T. After an overnight incubation at 4C, cells were washed and incubated with fluorophore-conjugated secondary antibody. Nuclei were visualized by Hoechst nuclear staining and images were captured by fluorescence microscope (Nikon).

Luciferase assays

Entire 3UTRs, if possible, or at least 0.8 to 1.2 kb around the predicted miR-22 site in 3UTR was cloned into a Renilla vector under a CMV promoter. Another vector containing the Firefly luciferase under a CMV promoter was used as a normalization control. HeLa cells were plated in 12-well plates at 8 x 104 cells/well and Renilla and Firefly vectors were co-transfected at 50 ng each along with 100 nM final concentration of miR-22. siRNA against GFP was used as a negative control. Cells were harvested 24 hours post-transfection and luciferase activity was measured using the Promega Dual Luciferase kit according to manufacturers instructions. For each construct assayed, we performed at least 2 biologically independent experiments with 3 technical replicates per experiment. Fold suppression was calculated as the ratio of Renilla to Firefly values for miR-22 normalized by the mean of the Renilla to Firefly ratios for the siRNA against GFP.

Chapter 6: Identifying miRNA Targets by Ago2 IP

INTRODUCTION

The precise criteria through which miRNAs recognize their target genes still remain unclear. It is known for certain that some amount of complementarity between the 5' end of the miRNA and the target transcript is essential for recognition. MicroRNAs commonly target the 3' UTRs and effective target suppression requires a perfect match at the seed sequence i.e. nucleotides 2-8 from the 5' end of the miRNA [30]. However, many experimentally verified target show that a few mismatches are tolerated in the seed sequence and some sites are found in the coding regions as well as the 5' UTRs. Additionally, it has been shown that extensive complementarity between the 3'end of the miRNA can compensate for a non-ideal seed match. Also, the sequence context within which the miRNA binding site resides can affect accessibility of the RISC machinery thereby influencing miRNA regulation [49].

MicroRNA target prediction algorithms mainly work by focusing on the seed match between the miRNA and the target 3' UTR [30]. To increase specificity, some algorithms include site conservation criteria while others use free energy calculations of the miRNA-mRNA duplex to select the best possible sites [149, 159, 160]. Experimentally, the most common strategy is to overexpress the miRNA either by transient transfection or expression off a viral vector and assay changes in transcript abundance by expression microarrays or protein levels by mass spectrometry [122, 124]. Transcripts or proteins that are downregulated and have seed matches are potential targets. The drawback with using microarrays to assay changes in transcript levels is that genes that show suppression at the protein level but no change at the transcript level will be missed as false negatives while transcripts showing expression changes due to

secondary effects will be classified as target generating false positives. The mass spectrometry approach can potentially identify all proteins that change expression due to the miRNA but cannot distinguish direct from indirect effects. Another emerging technique is to use Ago2 immunoprecpitation to identify direct miRNA targets [161, 162]. Ago2 IP is carried out in cells that are transfected with the miRNA of interest or mock transfected. An increase in Ago2 occupancy of any given transcript after miRNA transfection would indicate a direct target. So far, most Ago2 IPs have been carried out with tagged Ago2 proteins, either with transient transfections or creating stable cell lines. Transient transfection of plasmids limits the procedure to a few cell lines that are easily transfectable while creating stable cell lines is a laborious and time consuming process. Recently, Chi et. al. used Ago2 immunoprecipitations followed by deep sequencing to identify targets of the miRNA miR-124. However, the monoclonal antibody they used is not available commercially [163].

RESULTS AND DISCUSSION

We wanted to immunoprecipitate endogenous Ago2 to bypass the above mentioned issues. Hence, we tried a native antibody against Ago2 from Abcam that had been shown to work for western blots. As shown in Fig. 6.1A, the antibody was able to immunoprecipitate the endogenous Ago2 protein. We performed the IP in HeLa cells, essentially following the procedure used by Hendrickson et al [162]. To assay whether the IP worked, we performed quantitative RT-PCR on 4 genes, namely TP53INP1, DDIT4, E2F1 and IKBKB. TP53INP1 and DDIT4 were confirmed as miR-22 targets from our previous work, while E2F1 was shown to be a target of the miR-17-92 cluster [50]. Since both miR-22 and miR-17-92 miRNAs are highly expressed in HeLa cells, these miRNA targets would be expected to be occupied by Ago2 and hence show enrichment in the IP. IKBKB was chosen as a gene of interest. Since we found that

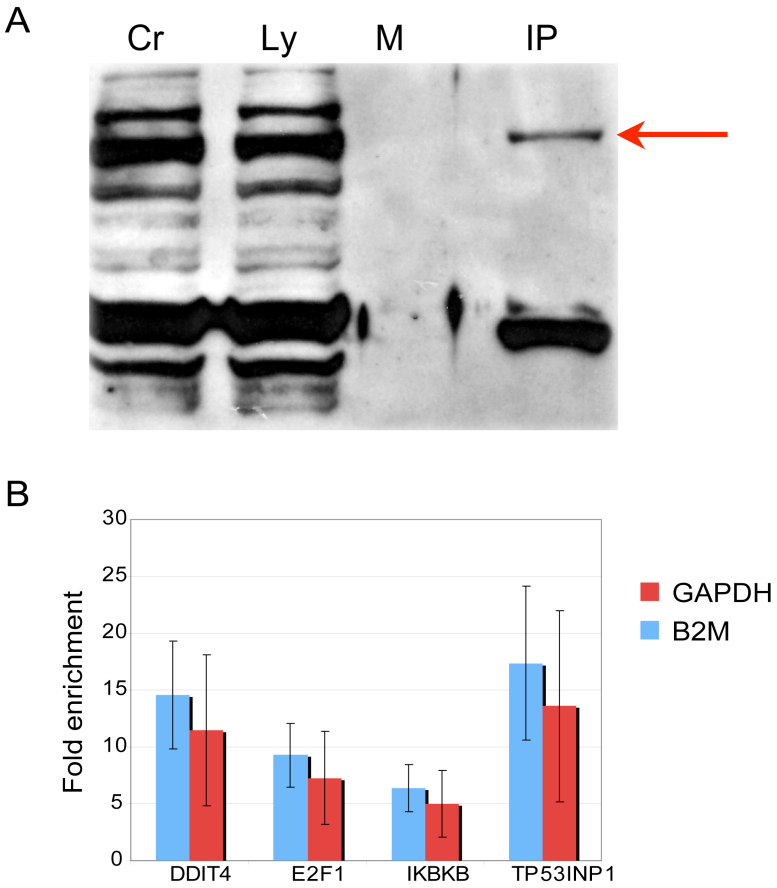


Figure 6.1 Ago2 IP

(A) Western blot for Ago2 pull-down. Cr: Crude cell lysate, Ly: Cleared lysate, M: Mock IP, IP: Ago2 IP. Film had to be over exposed to visualize the Ago2 IP band (indicated by the red arrow). (B) Quantitative RT-PCR for Ago2 IP. Enrichment values shown are the averages of 2 technical replicates. Error bars indicate 2 standard deviations from the mean.

IKBKB levels were suppressed by miR-22, we wanted to know if IKBKB was under miRNA control. Enrichment was measured using GAPDH and B2M as negative controls. Initial attempts produced modest enrichments (data not shown) and hence we decided to optimize the IP further. We introduced a pre-clearing step where the lysate is incubated with agarose beads before adding the antibody. The idea behind the pre-clearing step is that any transcripts that are "sticky" or adhere non-specifically to the beads will be cleared off before the addition of the antibody. Secondly, we compared the effect of increasing the number of wash steps. HeLa cells growing in a 15 cm diameter plate were lysed in 1ml of lysis buffer and split equally into two tubes. Each half received the same amount of antibody and beads. One tube was washed 3 times with lysis buffer while the other tube was washed 8 times. Fig. 6.1B shows the enrichment values for experiment involving 8 washes. All assayed genes showed a clear enrichment with respect to both B2M and GAPDH. The 8 washes IP showed higher enrichment as compared to the 3 washes IP (Fig. 6.2A). This increase in enrichment was maintained irrespective of the negative control used i.e. GAPDH or B2M. Finally, we compared the effect of adding 0.1 SDS to the washes. The lysate was prepared and split as before. The first tube was washed 8 times with the lysis buffer while the 2nd tube was washed 7 times with lysis buffer containing 0.1% SDS and then once with normal lysis buffer. As shown in Fig. 6.2B, the SDS washing caused a slight decrease in enrichments. Thus the final protocol included a pre-clearing step and 8 washes with the NP-40 only lysis buffer. We chose to use the Ago2 IP method to identify direct targets of miR-22. We transfected HeLa cells with the mature miR-22 duplex at a final concentration of 100 nM and

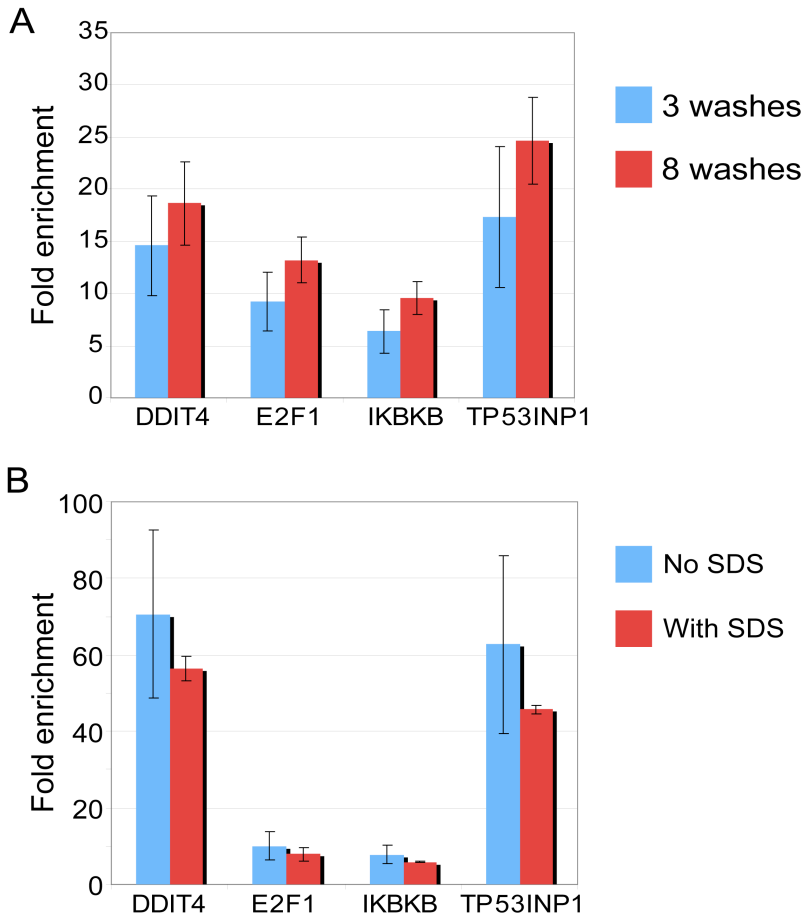


Figure 6.2 Optimizing the IP.

(A) Comparing 3 washes vs. 8 washes with lysis buffer. Enrichment increased slightly with 8 washes. (B) Effect of SDS on the enrichment. Including 0.1% SDS in the wash decreased the enrichment for all genes assayed.

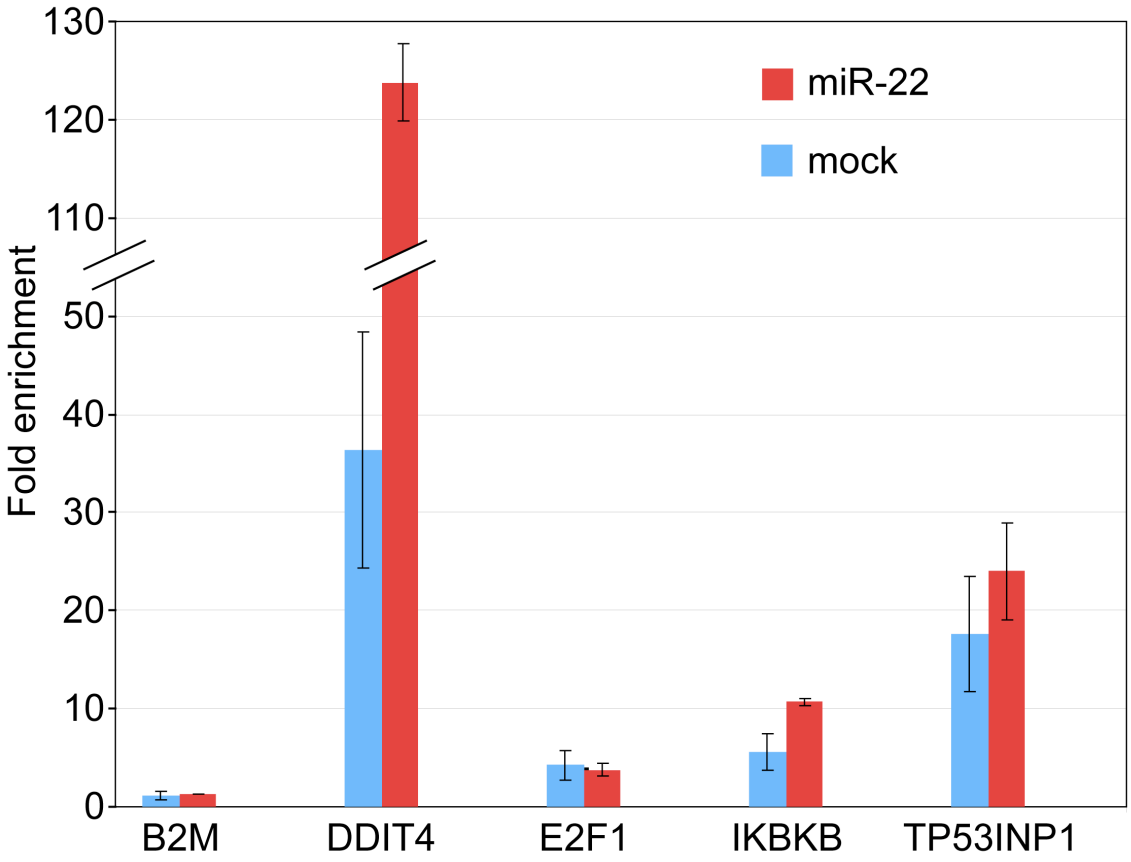


Figure 6.3 Ago2 IP under miR-22 transfection.

Ago2 IP was performed in miR-22 and mock transfected HeLa cells. Enrichment was calculated with respect to GAPDH using the ΔΔCt method. The enrichment for DDIT4 under miR-22 transfection is shown on a broken y-axis.

performed Ago2 IP as described above. We also performed a parallel Ago2 IP in mock transfected HeLa cells. Direct targets were expected to show a significant increase in Ago2 occupancy and hence a corresponding increase in enrichment in the IP performed under miR-22 transfection as compared to the IP performed in mock transfected cells. We chose to screen the following genes by quantitative RT-PCR as proof-of-principle for the assay: DDIT4, TP53INP1 and IKBKB were our expected true positives while B2M and E2F1 were the expected true negatives. DDIT4 and IKBKB showed a significant increase in enrichment after miR-22 transfection (Fig. 6.3). The expected true negatives B2M and E2F1 did not show any increase in enrichment in response to miR-22 transfection (Fig. 6.3). However, TP53INP1, which was a confirmed target, showed only a modest increase in enrichment after miR-22 transfection (Fig. 6.3). The standard deviations for the technical replicates was high in the case of TP53INP1 and hence the difference between mock and miR-22 transfection was not deemed significant. This data suggests that the assay can be expected to be highly specific with a low false positive rate. However, further optimizations will be required to increase sensitivity. One possibility is to increase the amount of miRNA transfected and/or increase the duration of the transfection. We used 100 nM final concentration of the miRNA and cells were harvested 16 hours post-transfection. Increasing the concentration to 200 nM and/or harvesting the cells 24 or even 48 hours post-transfection may recover weaker targets. Experiments are already underway to optimize these new conditions.

MATERIALS AND METHODS

Ago2 immunoprecipitation

HeLa cells were maintained in DMEM (Dulbecco's Modified Eagle's Medium) supplemented with 10% FBS at 37C under 5% CO2 and harvested at 70% confluence.

Cells were lysed by adding 0.5 ml of lysis buffer (150 mM KCL, 25 mM Tris, pH 8.0, 0.5 M EDTA, 0.5% NP-40) directly to the cells and incubating at 4C for 30 minutes. Lysate was scraped off and spun down at 14,000 rpm for 15 minutes at 4C. 50 ul of the cleared lysate was used as input while the rest was pre-cleared with 50 ul of protein G bead slurry (25 ul bead volume) for 3 hours at 4C. Pre-cleared lysate was incubated with 10 ug of Ago2 antibody overnight at 4C. After the overnight incubation, 50 ul of protein G bead slurry (25 ul bead volume) was added to the lysate and incubated at 4C for 4 hours. Beads were washed with lysis buffer with or without 0.1% SDS for the indicated number of times. For the western blot assay elution was performed by adding 100 ul of TE buffer (Tris-EDTA) containing 1% SDS directly to the beads and incubating at 65C for 15 minutes. 15 ul of the eluate was loaded on 10% gel and immunoblotting was performed as described in Chapter 4 using the anti-Ago2 antibody. For the quantitative RT-PCR assay, RNA was eluted by adding 1ml TRIZOL reagent directly to the beads. Further RNA purification with TRIZOL was according to the manufacturer's instructions.

Ago2 immunoprecipitations to detect miR-22 targets

HeLa cells were grown as described above in 10 cm^2 culture dishes (Nunc). Cells were transfected at a confluence of 50-70% with miR-22 mature duplexes at a final concentration of 100 nM using Lipofectamine 2000 according to manufacturers instructions. Cells mock transfected with lipofectamine alone served as control. Cells were harvested 16 hours post-transfection for Ago2 IP where the IP was performed as described above.

Quantitative Reverse-Transcription PCR (RT-PCR)

Quantitative RT-PCR was performed by the ΔΔCt method as describe in Chapter 4. Ct values for the gene of interest were normalized to GAPDH.

Chapter 7: Summary and Future Directions

The ChIP-chip technology enables the identification of TF binding sites across the genome in a truly high-throughput manner. The main drawback of this technology however, is the requirement for microarrays. Whole genome microarrays for complex organisms like human and mouse are still expensive and the entire process labor intensive. For other complex organisms like rat and chimp, whole genome microarrays are not even available. Additionally, microarrays suffer from low sensitivity and a higher background due to potential cross-hybridization. Sequencing based technologies circumvent many of these pitfalls. Sequencing techniques can be used to identify binding sites in any sequenced organism. There is no potential for cross hybridization and the dynamic range of the detectable signal is much higher. Additionally, there no constraints arising from annotations since if the reference genome gets updated, all that needs to be done is to realign the sequence data [164]. With these advantages in mind, we first developed STAGE or Sequence Tag Analysis of Genomic Enrichment, which involved deriving tags or 21 bp DNA sequences from ChIP-enriched DNA. We developed algorithms to analyze the sequencing data and successfully applied this technique as a proof-of-principle to map E2F4 binding sites in the human genome. The first application of this technique was based on the expensive and laborious Sanger sequencing. Later, the advent of 454 sequencing technology enabled us to make STAGE more high throughput, less labor intensive and therefore more competitive with contemporary microarrays. We enhanced STAGE further by incorporating the tetratag technique that increased throughput an additional 100%. However, the depth of sequencing possible with 454 technology at the given expense was not enough to cover all binding sites for any TF.

Further improvements in sequencing technologies including Solexa sequencing and ABI SOLID finally established sequencing as the clear winner over microarrays for genome wide identification of TF binding. This “new kid on the block” has been termed ChIP-Seq and involves ChIP followed by direct sequencing of the ChIP-enriched DNA by ultra high-throughput sequencing that generates several million short reads from the ends of DNA fragments [165]. We applied the ChIP-Seq technique to identify binding sites of the oncofactor E2F4 in the human genome. However, the generation of such vast amounts of data posed algorithmic challenges and opportunities. A higher sampling of each binding locus now made it possible to define binding sites more precisely as compared to STAGE, theoretically at a single base resolution. In order to take advantage of this sampling depth, we developed a novel algorithm that could detect binding sites at the level of a single base pair dependent upon the distribution of reads around any given base pair. Our analysis detected around 16,000 E2F4 binding sites across the entire genome with high confidence. Further analysis on the detected peaks revealed E2F4 binding in intergenic regions, possibly to enhancers or yet to be discovered genes, that were not observed before. We discovered novel E2F4 binding motifs and showed that E2F4 can also regulate microRNAs.

An important limitation in TF binding site identification is the availability of ChIP-grade antibodies. Unlike yeast, where tagged ORF libraries are available for the entire genome, tagging mammalian TFs is seriously lagging behind. Comprehensive identification of whole genome transcriptional networks would benefit greatly if such libraries are made available to researchers.

For a TF to bind its regulatory element on the genome, it must be granted access to its specific binding site of the genome, which in turn is regulated by nucleosome positioning and remodeling. Hence, it is important to understand how and where

nucleosomes are positioned and how nucleosome remodeling affects gene regulation. We mapped nucleosome positions in normally growing and heat shocked yeast using ultra high throughput sequencing. The purpose of this study was not only to identify all nucleosome positions at high resolution but also to observe the correlation between nucleosome remodeling and the transcriptional re-programming that sets in upon exposing yeast cells to heat shock. We found that the TSS and stop codons of yeast genes are marked by well-positioned nucleosomes. Most nuclesomes retain their positions between normally growing cells and heat shocked cells with nucleosome positioning extending well into the coding regions. Nucleosomes exhibit strong positioning in TATA-less promoters and transcriptionally inactive. Interestingly, the first nucleosome is unaffected by the transcription rate. Gene-specific remodeling is restricted to the eviction, appearance or repositioning of one or two nucleosomes. Activation of genes is usually accompanied by the eviction of nucleosomes around the promoter region while repression is associated with appearance of nucleosomes. However, we observed that eviction of nucleosome near the TSS of activated genes can be accompanied by the appearance of nucleosomes distally. Finally, stimulus-dependent nucleosome remodeling increases or decreases accessibility of functional binding sites of specific TFs that are mediate the response of the cell to the stimulus.

Although important, identifying nucleosome positions and mapping nucleosome re-positioning events across the genome represents just the first step in understanding epigenetic regulation in eukaryotic cells. Identifying the various histone modifications and variants, their turnover and how they are regulated at genes that are activated or repressed is crucial to understand how chromatin remodeling shapes gene expression patterns under different conditions. The same techniques and analysis methods described in chapter 3 can be applied to map sites of histone modifications genome-wide. Instead

digesting chromatin with MNase, ChIP can be used to isolate genomic loci associated with specific histone modifications and sequenced using ultra high throughput technologies. Performing such assays under conditions where gene expression is perturbed on a global scale would enable us to better correlate histone remodeling to expression changes.

Yet another important aspect of gene regulation operates at the post-transcriptional level. Small non-coding microRNAs suppress translation or degrade mRNAs by mRNA decay pathways such P-bodies or NMD. MiRNAs can regulate TFs and in turn can be regulated by TFs. We characterized a TF-miRNA network involving the oncofactor Myc and the miRNA miR-22 that suppresses the interferon pathway as primary fibroblasts enter a stage of rapid proliferation. We found that Myc activates miR-22 during the transition from quiescence to proliferation. By using expression microarrays, we identified downstream targets of miR-22 in primary fibroblasts and HeLa cells and observed that the miR-22 targets pro-apoptotic genes. Surprisingly, we found that miR-22 also inhibits the interferon response that is activated in quiescent fibroblasts. We further characterized miR-22 mediated interferon suppression in HeLa cells by using the viral mimetic poly I:C and found that miR-22 prevents activation of the interferon-beta promoter by inhibiting NF-kappaB nuclear localization. We developed a novel miRNA target prediction strategy based on microarray expression data and identified IKBKB as a likely miR-22 target. We show that IKBKB protein expression is suppressed by miR-22 thus providing a possible mechanistic explanation as to how miR-22 prevents NF-kappaB activation. Additionally, we show that miR-22 participates in feedback loops involving Myc and its repressor MXD4. Finally, we also show that miR-22 inhibits p53 activated cell cycle arrest and pro-apoptotic genes thereby revealing a

novel cross talk between the pro-proliferative Myc and anti-proliferative p53 regulatory networks.

The experimental methods described in chapter 4 can be applied to identify downstream targets of other miRNAs that are differentially regulated during the quiescence to proliferation transition. Identifying miRNA targets by assaying changes in mRNA expression in response to miRNA perturbation is a powerful method but has some significant drawbacks. The assay may identify indirect targets as false positives and miss targets that are suppressed only at the translational level as false negatives. Hence, a more direct approach is required to identify miRNA targets. On such approach is based on Ago2 immunoprecipitations (IPs). In this method, transcripts that show a change in Ago2 occupancy in response to miRNA perturbation are identified as targets. We have made significant progress in this direction as well. By using an antibody directed against the endogenous Ago2 protein, we have eliminated the need to use tagged Ago2 proteins thereby eliminating the need for transfections that are expensive and can vary in efficiency. We have optimized the Ago2 IP such that for the expected true positives we get significant enrichments but the expected true negatives are not enriched. We have applied this methodology to identify direct targets of the miRNA miR-22 and for a small subset of assayed genes, we see that our method is highly specific though it still suffers from lower sensitivity than desired. Further optimizations will no doubt be required to enhance the sensitivity of the method.

In order to gain a holistic view of gene regulatory mechanism underlying set expression patterns in a cell, it is not enough to focus on just the transcriptional or the epigenetic or the post-transcriptional aspect. It is necessary to take into account all the players that participate in the regulation game and with current advances in technology, this is now possible. However, breakthroughs in science commonly result from advances

in technology and hence, more efforts should be diverted to develop novel methods to analyze the various mechanisms a cell utilizes to establish it's regulatory network.

Appendix A Human E2F4 Targets Detected by STAGE

Genes marked with an asterisk are regulated by bi-directional promoters i.e. the same promoters directs transcription of two genes on the opposite strand.

No.	Gene symbol	Score	E2F4 site
1	MTUS1	1971	
2	ULBP3	1961	
3	SNRPD2*	1933	
	QPCTL*	1923	
4	PXK	1015	
5	FLJ22353	993	
6	GAJ	993	Yes
7	ACR	992	
8	RAD54L	992	
9	AAMP	982	
10	ABHD2	982	
11	BLVRB*	982	
	SPTBN4*	982	
12	DC2	982	
13	FLJ13912	982	
14	FLJ25416	982	
15	FLJ32000	982	Yes
16	FLJ90834	982	
17	MPV17	982	Yes
18	PRCP	982	
19	PSMA4	982	
20	RNF29	982	
21	TOPK	982	Yes
22	DRF1	974	
23	LMO7	971	
24	SLC3A2	971	Yes
25	SOAT2	971	
26	ARHGAP11A	965	Yes
27	ABC1	961	
28	BTRC	961	
29	GAL3ST1	961	
30	CSTF3	961	
31	CTAG3*	961	
	RIOK1*	961	
32	DNALI1	961	
33	EPHA3	961	
34	FIBL-6	961	Yes
35	FLJ20712	961	
36	HIST2H2AC	961	
37	HOZA3	961	
38	JPH2	961	
39	MAP3K7	961	Yes
40	METAP2	961	Yes
41	PDGFA	961	
42	RPL23A	961	Yes
43	SNIP1	961	Yes
44	CCRL2	926	
45	C20orf141	913	

Appendix B Optimal Window Size for Scanning the Genome

Window size = 300 bp

k	STAT1	Random	%FDR
1	84038	90921	108
2	3971	1235	31
3	436	11	2.5
4	73	0.15	0.2
5	6	0	0

Window size = 500 bp

k	STAT1	Random	%FDR
1	82296	90267	110
2	5373	1885	35
3	734	26	3.5
4	166	0.3	0.2
5	33	0	0

Window size = 1000 bp

k	STAT1	Random	%FDR
1	79461	88701	111
2	7829	3433	44
3	1222	79	2.5
4	297	2	0.6
5	81	0.1	0.12
6	27	0	0

Window size = 2000 bp

k	STAT1	Random	%FDR
1	75202	85833	114
2	11418	6291	55
3	2002	255	1.5
4	567	8.4	0.6
5	170	0.3	0.2
6	64	0.05	0.08

K is the number of single hit tags found within a given window. The false discovery rate (FDR) was calculated as (Random/STAT1)*100.

Appendix C STAT1 Target Genes

RefSeq annotated genes that had a STAT1 binding site within 20 kb upstream or downstream of their TSS.

Genes	Description	Position of STAT1 binding site	
Immune response			
C1S	complement component 1, s subcomponent	UPS	111
SECTM1	secreted and transmembrane 1	UPS	5365
IL18BP	interleukin 18 binding protein	EXN	527
STAT3	signal transducer and activator of transcription 3	UPS	307
BST2	bone marrow stromal cell antigen 2	UPS	50
IFI35	interferon-induced protein 35	EXN	35
HLA-E	major histocompatibility complex, class I, E	UPS	168
IRF1	interferon regulatory factor 1	UPS	5091, 6761
CD7	CD7 antigen (p41)	DS	19119
IFI16	interferon, gamma-inducible protein 16	UPS	9910
IL4R	interleukin 4 receptor	INT	13152
STAT1	signal transducer and activator of transcription 1	UPS	611
PLSCR1	phospholipid scramblase 1	INT	1135
Lipid metabolism			
SMPD1	sphingomyelin phosphodiesterase 1	UPS	376
SULT1A1	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1	DS	5222
LRP8	low density lipoprotein receptor-related protein 8	DS	1685
SLC27A4	solute carrier family 27 (fatty acid transporter), member 4	UPS	769, 17881
APOL6	apolipoprotein L, 6	INT	247
Cell adhesion			
ITGB3	integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61)	UPS	13708
ICAM1	intercellular adhesion molecule 1 (CD54), human rhinovirus receptor	INT	392
NID2	nidogen 2 (osteonidogen)	UPS	2976
FNBP4	formin binding protein 4	UPS	875
COL5A1	collagen, type V, alpha 1	UPS	15031
ENG	endoglin (Osler-Rendu-Weber syndrome 1)	INT	7660
Cell growth, differentiation and death			
NRG1	neuregulin 1	UPS	15639
TNFAIP2	tumor necrosis factor, alpha-induced protein 2	UPS	8019
PHLDA2	pleckstrin homology-like domain, family A, member 2	UPS	11153
DAPK3	death-associated protein kinase 3	UPS	14381
DAPK3	death-associated protein kinase 3	INT	777
TNFRSF1A	tumor necrosis factor receptor superfamily, member 1A	INT	5733
CDC7	CDC7 cell division cycle 7 (S. cerevisiae)	UPS	196
CDK6	cyclin-dependent kinase 6	DS	657
BRCA2	breast cancer 2, early onset	INT	465
RAP1A	RAP1A, member of RAS oncogene family	UPS	527

TREX1	three prime repair exonuclease 1	DS	669
MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2	DS	7741
Protein metabolism			
TARS	threonyl-tRNA synthetase	UPS	233
DUSP14	dual specificity phosphatase 14	UPS	6197
EEF2	eukaryotic translation elongation factor 2	DS	1254, 16412
SURF6	surfeit 6	UPS	86
MATN1	matrilin 1, cartilage matrix protein	DS	4665
EEF2K	eukaryotic elongation factor-2 kinase	INT	976
PSMB3	proteasome (prosome, macropain) subunit, beta type, 3	INT	93
GFM1	G elongation factor, mitochondrial 1	INT	194
USP48	ubiquitin specific protease 48	UPS	831
RPL35	ribosomal protein L35	UPS	1730
DPP9	dipeptidylpeptidase 9	DS	10682
SLC9A3R1	solute carrier family 9 (sodium/hydrogen exchanger)	INT	9291
PARP14	poly (ADP-ribose) polymerase family, member 14	UPS	19283
PTP4A2	protein tyrosine phosphatase type IVA, member 2	UPS	1310
ADAM17	a disintegrin and metalloproteinase domain 17 (tumor necrosis factor, alpha, converting enzyme)	INT	342
VRK3	vaccinia related kinase 3	UPS	123
HSPB1	heat shock 27kDa protein 1	UPS	7760
PTK6	PTK6 protein tyrosine kinase 6	UPS	651
TRIO	triple functional domain (PTPRF interacting)	INT	13363
CARS	cysteinyl-tRNA synthetase	INT	7483
DTX3L	deltex 3-like (Drosophila)	UPS	195
HS6ST1	heparan sulfate 6-O-sulfotransferase 1	INT	13493
Signal transduction			
SNX27	sorting nexin family member 27	UPS	5560
IPO8	importin 8	UPS	367
RAB36	RAB36, member RAS oncogene family	DS	7340
GPR37L1	G-protein coupled receptor 37 like 1	UPS	15711
CXXC5	CXXC finger 5	UPS	12166
PITPNC1	phosphatidylinositol transfer protein, cytoplasmic 1	INT	12593
RHOF	ras homolog gene family, member F (in filopodia)	UPS	1131
ITPK1	inositol 1,3,4-triphosphate 5/6 kinase	UPS	5305
GUCA1B	guanylate cyclase activator 1B (retina)	INT	3969
RASSF5	Ras association (RalGDS/AF-6) domain family 5	INT	6210, 7336
Transport			
ABCC11	ATP-binding cassette, sub-family C (CFTR/MRP), member 11	EXN	321
SCNN1A	sodium channel, nonvoltage-gated 1 alpha	DS	4824
VMD2L3	vitelliform macular dystrophy 2-like 3	EXN	171
SLC25A1	solute carrier family 25 (mitochondrial carrier; citrate transporter), member 1	UPS	922
KCNJ12	potassium inwardly-rectifying channel, subfamily J, member 12	UPS	4203

SLC22A2	solute carrier family 22 (organic cation transporter), member 2	DS	19540
C1QTNF6	C1q and tumor necrosis factor related protein 6	UPS	671, 9571
CLIC2	chloride intracellular channel 2	UPS	366
SLC25A25	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 25	INT	256
COPG	coatomer protein complex, subunit gamma	UPS	3780
ZNF406	zinc finger protein 406	INT	10934, 11357
VPS18	vacuolar protein sorting protein 18	UPS	2668
HPX	hemopexin	DS	1813
SLC35A2	solute carrier family 35 (UDP-galactose transporter), member A2	UPS	195
MTCH2	mitochondrial carrier homolog 2 (C. elegans)	UPS	66

Other biological processes

C1orf191	chromosome 1 open reading frame 191	UPS	8524
BATF2	hypothetical protein BC012330	EXN	118
RND1	Rho family GTPase 1	DS	12542, 13750
EIF1	putative translation initiation factor	UPS	9819
FLJ32926	hypothetical protein FLJ32926	DS	13806
FLJ25660	hypothetical protein FLJ25660	DS	14356
C2orf18	chromosome 2 open reading frame 18	UPS	6304
LHFPL5	lipoma HMGIC fusion partner-like 5	INT	6227
PRPF4	PRP4 pre-mRNA processing factor 4 homolog (yeast)	UPS	306
C16orf47	FLJ26184 protein	DS	2438, 19985
PAF1	hypothetical protein F23149_1	UPS	11736
ZNF473	zinc finger protein 473	UPS	283
CLPS	colipase, pancreatic	UPS	14195
LONPL	peroxisomal lon protease	UPS	12349
ZNF114	hypothetical protein MGC17986	DS	18873
RPL7A	ribosomal protein L7a	UPS	11938
PUS3	pseudouridylate synthase 3	INT	85
LOC441108	hypothetical protein	DS	16654
VAT1	vesicle amine transport protein 1 homolog (T californica)	DS	15536
C9orf88	chromosome 9 open reading frame 88	INT	12964
IXL	intersex-like (Drosophila)	DS	11453
APBB1	amyloid beta (A4) precursor protein-binding, family B, member 1	UPS	19736
LOC399900	hypothetical protein	DS	5254
NAGLU	N-acetylglucosaminidase, alpha- (Sanfilippo disease IIIB)	DS	5397
BBS4	Bardet-Biedl syndrome 4	UPS	15058
MGC14327	hypothetical protein MGC14327	DS	7245
PTPN21	protein tyrosine phosphatase, non-receptor type 21	UPS	5222
SEMA3F	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3F	DS	8681
COQ4	coenzyme Q4 homolog (yeast)	DS	17256
COQ4	coenzyme Q4 homolog (yeast)	EXN	144
SERPINB6	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 6	UPS	16784
DSCR4	Down syndrome critical region gene 4	UPS	18568
SRMS	src-related kinase lacking C-terminal regulatory tyrosine and N-	DS	9499

	terminal myristylation sites		
PLEKHG2	pleckstrin homology domain containing, family G (with RhoGef domain) member 2	UPS	10334
ASB6	ankyrin repeat and SOCS box-containing 6	DS	15858
APOA1BP	apolipoprotein A-I binding protein	DS	9445
TMEM55B	chromosome 14 open reading frame 9	UPS	307
NQO2	NAD(P)H dehydrogenase, quinone 2	UPS	11356
MGC13168	vitelliform macular dystrophy 2-like 3	DS	10227
ICAM4	intercellular adhesion molecule 4, Landsteiner-Wiener blood group	UPS	15479
C1orf147	chromosome 1 open reading frame 147	UPS	16027, 17153
MCM2	MCM2 minichromosome maintenance deficient 2, mitotin (S. cerevisiae)	DS	19019
PODXL2	podocalyxin-like 2	UPS	11788
FLJ37078	hypothetical protein	DS	9788
TRUB2	TruB pseudouridine (psi) synthase homolog 2 (E. coli)	UPS	261, 17373
DKFZp686I15217	hypothetical protein DKFZp686I15217	UPS	58
ZNF302	zinc finger protein 302	UPS	292
C17orf41	hypothetical protein FLJ12735	EXN	317
SSTR3	somatostatin receptor 3	DS	14452
C6orf65	chromosome 6 open reading frame 65	UPS	16498
SLBP	stem-loop (histone) binding protein	UPS	15299
C1orf111	hypothetical protein LOC284680	UPS	4553
C1orf90	hypothetical protein MGC10820	UPS	15
DENND1A	KIAA1608	UPS	3696
SHC1	SHC (Src homology 2 domain containing) transforming protein 1	UPS	9158
C9orf32	AD-003 protein	INT	152
UBASH3A	ubiquitin associated and SH3 domain containing, A	UPS	7729, 8564
NUP210	nucleoporin 210kDa	UPS	581
NPFFR1	G protein-coupled receptor 147	UPS	1091
MCL1	myeloid cell leukemia sequence 1 (BCL2-related)	DS	17076
LOC93343	hypothetical protein BC011840	UPS	14477
PELI3	pellino 3 alpha	UPS	12477
TMEM104	hypothetical protein FLJ20255	UPS	18599
HSD17B1	hydroxysteroid (17-beta) dehydrogenase 1	UPS	10636
CRLF3	cytokine receptor-like factor 3	UPS	7648
ANKRD13D	hypothetical protein LOC338692	UPS	1179
C1orf138	chromosome 1 open reading frame 138	UPS	469
ZNF335	zinc finger protein 335	UPS	11289
FLJ45248	FLJ45248 protein	EXN	961
RARRES2	retinoic acid receptor responder (tazarotene induced) 2	DS	18775
ZC3H3	zinc finger CCCH type domain containing 3	UPS	5825
SNX1	sorting nexin 1	UPS	2129
PLXDC1	plexin domain containing 1	DS	12343
DSCR8	Down syndrome critical region gene 8	DS	18456
KIAA0040	KIAA0040 gene product	UPS	697
LOC284751	hypothetical protein	INT	236
C7orf29	chromosome 7 open reading frame 29	UPS	7006

TMPRSS3	transmembrane protease, serine 3	DS	746
TMPRSS3	transmembrane protease, serine 3	UPS	89
TMEM61	hypothetical protein LOC199964	UPS	7273
THAP2	hypothetical protein DKFZp564I0422	EXN	145
MTHFR	5,10-methylenetetrahydrofolate reductase (NADPH)	DS	15179
RARG	retinoic acid receptor, gamma	UPS	19431
AEBP2	AE binding protein 2	UPS	7648
LCK	lymphocyte-specific protein tyrosine kinase	UPS	4113
GPATC4	G patch domain containing 4	INT	268
TMEM50A	small membrane protein 1	DS	17864
RND2	ras homolog gene family, member N	UPS	18398
CKS1B	CDC28 protein kinase regulatory subunit 1B	DS	8785
PPAP2C	phosphatidic acid phosphatase type 2C	INT	951
ELF3	E74-like factor 3 (ets domain transcription factor, epithelial-specific)	EXN	2183
SULT1A2	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 2	UPS	7036
PCGF2	ring finger protein 110	UPS	4536
ZNF367	zinc finger protein 367	EXN	313
PSCA	prostate stem cell antigen	INT	614
MFSD5	hypothetical protein MGC11308	UPS	468
PSRC1	p53-regulated DDA3	UPS	2058
C11orf10	chromosome 11 open reading frame 10	EXN	10
LOC153222	adult retina protein	UPS	144
NDOR1	NADPH dependent diflavin oxidoreductase 1	UPS	7337
CRI1	CREBBP/EP300 inhibitor 1	UPS	6502
FOXL2	forkhead box L2	DS	5389
GOLGA	similar to Golgi autoantigen, golgin subfamily A member 6 (Golgin linked to PML) (Golgin-like protein)	DS	16438
ZNF181	zinc finger protein 181	UPS	377
LOC389289	hypothetical protein	UPS	175
RPL27	ribosomal protein L27	DS	8414
SSNA1	Sjogren's syndrome nuclear autoantigen 1	DS	9713
JRK	jerky homolog (mouse)	UPS	14561
KRTHB1	keratin, hair, basic, 1	DS	10462, 10906
ESPL1	extra spindle poles like 1 (S. cerevisiae)	UPS	17139
SPINK4	serine protease inhibitor, Kazal type 4	UPS	5807
C9orf50	hypothetical protein LOC375759	UPS	5531
ARL2	ADP-ribosylation factor-like 2	UPS	17269
HAND1	heart and neural crest derivatives expressed 1	DS	6787
HM13	histocompatibility (minor) 13	INT	345
TACC3	transforming, acidic coiled-coil containing protein 3	DS	6064
TMEM62	hypothetical protein FLJ23375	UPS	10223
P8	p8 protein (candidate of metastasis 1)	DS	6914
MRPL4	mitochondrial ribosomal protein L4	DS	19531
MRPL11	mitochondrial ribosomal protein L11	UPS	15616
CLCN6	chloride channel 6	UPS	15357
FAM96A	hypothetical protein FLJ22875	EXN	160
ZFP36	zinc finger protein 36, C3H type, homolog (mouse)	UPS	4071

PREX1	PREX1 protein	UPS	4123
C20orf195	hypothetical protein MGC5356	UPS	15014
SSH3	slingshot homolog 3 (Drosophila)	UPS	15361
APOC2	apolipoprotein C-II	DS	8883
PIM2	pim-2 oncogene	DS	7156
ICAM5	intercellular adhesion molecule 5, telencephalin	UPS	18484
C9orf106	chromosome 9 open reading frame 106	DS	12981
CDC26	cell division cycle 26	EXN	202
LENEP	lens epithelial protein	UPS	10063
UBR1	ubiquitin protein ligase E3 component n-recognin 1	UPS	17212
ATP5B	ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide	DS	9386
TMEM129	hypothetical protein BC009331	UPS	6245
LOC284912	hypothetical gene supported by BC001801	UPS	13488
EDEM2	chromosome 20 open reading frame 31	INT	281
SURF5	surfeit 5	DS	11842
C5orf13	chromosome 5 open reading frame 13	DS	9040
C20orf149	chromosome 20 open reading frame 149	DS	17226
HYLS1	hypothetical protein FLJ32915	DS	19512
DDX23	DEAD (Asp-Glu-Ala-Asp) box polypeptide 23	EXN	54
DDX23	DEAD (Asp-Glu-Ala-Asp) box polypeptide 23	UPS	1154
C9orf75	chromosome 9 open reading frame 75	INT	1712
BAZ2A	bromodomain adjacent to zinc finger domain, 2A	UPS	303
C10orf47	chromosome 10 open reading frame 47	UPS	12703
OSGEP	O-sialoglycoprotein endopeptidase	UPS	6745
APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1	DS	6654
LRRC61	hypothetical protein MGC3036	UPS	673
WDR34	WD repeat domain 34	EXN	33
APOC4	apolipoprotein C-IV	DS	12631
GSDMDC1	gasdermin domain containing 1	UPS	11063
NOB1	nin one binding protein	UPS	35
FEN1	flap structure-specific endonuclease 1	UPS	74
PQBP1	polyglutamine binding protein 1	DS	13590
KRT7	keratin 7	DS	11172
KRT7	keratin 7	UPS	946
C14orf79	chromosome 14 open reading frame 79	UPS	8799
NAT9	embryo brain specific protein	DS	18396
RHOV	ras homolog gene family, member V	UPS	17520
TIMM17B	translocase of inner mitochondrial membrane 17 homolog B (yeast)	UPS	13702
RTDR1	rhabdoid tumor deletion region gene 1	UPS	10611
PARP9	poly (ADP-ribose) polymerase family, member 9	INT	354
HAPLN2	hyaluronan and proteoglycan link protein 2	UPS	18083
DDX25	DEAD (Asp-Glu-Ala-Asp) box polypeptide 25	UPS	1382
LTBR	lymphotoxin beta receptor (TNFR superfamily, member 3)	UPS	13465
NP	nucleoside phosphorylase	UPS	7621
KRTAP17-1	keratin associated protein 17-1	UPS	13707
WWP2	WW domain containing E3 ubiquitin protein ligase 2	UPS	7409

ARPC5L	actin related protein 2/3 complex, subunit 5-like	UPS	5513
FLAD1	FAD synthetase	INT	182
ADAMTSL4	thrombospondin repeat containing 1	DS	13163
DAB2IP	DAB2 interacting protein	UPS	5701
PSRC2	hypothetical protein MGC23401	UPS	179
CLPTM1	cleft lip and palate associated transmembrane protein 1	UPS	512
LY6K	lymphocyte antigen 6 complex, locus K	UPS	19042
FCHO2	FCH domain only 2	UPS	72
FTH1	ferritin, heavy polypeptide 1	UPS	5387
ANAPC2	anaphase promoting complex subunit 2	UPS	9788
FLJ11806	nuclear protein UKp68	UPS	2968
CCNL1	cyclin L1	UPS	14267
KIAA1618	KIAA1618	UPS	6047
STOM	stomatin	INT	7363
MRVI1	murine retrovirus integration site 1 homolog	INT	17906

UPS: upstream, DS: downstream, INT: intron, EXN: exon.

Appendix D Primers Used for Quantitative PCR Analysis

ChIP-PCR

Locus	Primer sequence
IRF1-D	Forward: CTAGAACCCACCAACCTCCA
	Reverse: TGCCTCGAACTCACCCTACT
IRF1-P	Forward: CTGAAGCTGGCTGGAAAATC
	Reverse: AGCACTGGAGCAATTCCTTG
SLC25A25	Forward: GGCAGCATTTAGGGAACTTG
	Reverse: CAGGCACAGACAGAGCATGT
STAT3	Forward: ACGCGGGAATCAGCTAGTTA
	Reverse: TTTTTGTTGTCCCCAGAACC
chr22-34786430	Forward: GGATTTTCACCATCGGACTG
	Reverse: TCTCCTCCCTTCTCCCTGAT
TNFRSF1A	Forward: AGGGGAGAGGGAACTAGCAG
	Reverse: CCTTCTGCCTTTCTGCTGAC
DAPK3	Forward: AACAATTGACCGAGGTTTGG
	Reverse: GCCCAGCTCTTGGATGTTTA
C1S	Forward: GAGGACGCTGTCCTTGTTTC
	Reverse: GGCTGGGAGACCATGACTTA
APOL6	Forward: CCTCCCTTACAGCCATTTCA
	Reverse: AGTGGAGGGACAAATGCAAC
ADAM17	Forward: ATCCAGCCACCCTACTCCTT
	Reverse: GCTCCCTAGCTTTGTGTTCG
GAPDH	Forward: AAAAGCGGGGAGAAAGTAGG
	Reverse: GTCTTGAGGCCTGAGCTACG

Quantitative reverse transcription PCR

Gene	Primers (Forward/Reverse)
IFNB1	Forward: AGTCTCATTCCAGCCAGTGC
	Reverse: AGCTGCAGCAGTTCCAGAAG
OAS1	Forward: CAAGCTCAAGAGCCTCATCC
	Reverse: TCCCAAGCATAGACCGTCAG
IFITM1	Forward: ATGTCGTCTGGTCCCTGTTC
	Reverse: CCATCTTCCTGTCCCTAGACC
STAT1	Forward: CCTGCTCCAGGAATTTTGAG
	Reverse: GCTGCTCCTTTGGTTGAATC
GAPDH	Forward: CTGGGCTACACTGAGCACCAG
	Reverse: CCAGCGTCAAAGGTGGAG

Appendix E Nucleosome Overlaps

Table 1 Nucleosome overlaps between our study and previous studies. Percentages were calculated with respect to the lower of the two numbers considered in the overlap. Nucleosomes were considered to overlap if they were within 50 bp of each other.

Previous work	Score cut-off	Normally growing cells	After heat shock
Yuan et. al.	0.25	75.6%	79.3%
	0.5	51%	55%
Segal et. al.	0.25	73.5%	78.8%
	0.5	51.5%	60.6%
Lee et.al.	0.25	79.3%	78.2%
	0.5	82%	79.7%

Table 2 Overlap between nucleosome positions mapped in our study before and after heat shock

Score cut-off	≤ 20 bp	≤ 25 bp	≤ 30 bp	≤ 50 bp	≤ 100 bp	> 100 bp
0.1	52%	59%	64%	77%	93.3%	6.7%
0.2	56%	62%	67%	79%	93.8%	6.2%
0.25	54%	60%	65%	76%	90.4%	9.6%
0.3	55%	61%	66%	77%	90.1%	9.9%
0.4	55%	61%	65%	75%	86%	14%

Appendix F TF Target Enrichment

Enrichment or depletion of transcription factor targets in nucleosome profile clusters in Fig. 4.11

ChIP-chip targets

		Enriched	Depleted
ESA1 (Robert et. al.)			
	Group 3	P = 0.008	P = 8×10^{-4}
	Group 1		
GCN5 (Robert et. al.)			
	Group 3	P = 0.01	
	Group 4		P = 0.007
FHL1 (Harbison et. al.)			
	Group 3	P = 0.004	
	Group 4	P = 0.0015	
	Group 1		P = 1.5×10^{-4}
SFP1 (Harbison et. al.)			
	Group 3	P = 3.1×10^{-5}	
	Group 1		P = 0.005
RAP1 (Lieb et. al.)			
	Group 3	P = 1.2×10^{-4}	
	Group 1		P = 1.4×10^{-4}

Yeast functional regulatory network

		Enriched	Depleted
HSF1			
	Group 3	$P = 1.6 \times 10^{-4}$	
	Group 1		$P = 7.5 \times 10^{-3}$
CST6			
	Group 3	P = 0.002	
SPT10			
	Group 3	$P = 4.8 \times 10^{-4}$	
	Group 1		P = 0.009
SFP1			
	Group 3	$P = 6 \times 10^{-4}$	
	Group 1		P = 0.01
RAP1			
	Group 3	$P = 1.4 \times 10^{-4}$	
	Group 1		P = 0.015
	Group 2		P = 0.018

Appendix G Scoring Putative miR-22 Targets

Gene symbol	Score	Num of seeds
MAP3K3	0.999995505569568	15
ATP2A2	0.99999333158012	15
CCNL2	0.999973427522098	14
GPD1	0.999886772435224	12
OPHN1	0.999758499139842	10
TRAF7	0.999617515032484	9
BTRC	0.999339542245348	8
ADAR	0.997660029707889	7
IKBKB	0.997651129608894	7
MCC	0.997389839179348	7
FKBP5	0.996623619543265	6
RIPK1	0.996335768236977	7
VISA	0.996230293599273	7
MYD88	0.991944637082921	6
MLX	0.991378114241732	5
TBL1X	0.991018269632856	6
EMID2	0.987412247557907	5
IRAK1	0.985630912232194	5
NICN1	0.982965467866151	5
MAP3K1	0.98154643354612	4
CCL22	0.980959903337217	4
CLDN12	0.979094775119869	4
TNFRSF11A	0.978896907561604	4
TIRAP	0.978431541024315	4
CAPN5	0.977983189489246	5
ENPP7	0.973860888039642	4
MAP3K14	0.971873363149146	4
TBKBP1	0.97169260840058	4
TNFRSF10B	0.971574185531853	4
TYRO3	0.971243848654206	5
PLEKHF1	0.969960730614003	5
CAMK2D	0.965456120596103	4
STARD8	0.964977798153065	4
NGB	0.959586536385999	4
ST13	0.945301548744847	3
SYNGR2	0.942507671595737	3
RAI1	0.934439418232847	3
TNFRSF1B	0.928502873825659	4
CUL1	0.922901725868198	3
TRAF2	0.91919770671306	3
GNA15	0.916010617527479	3
MAP3K2	0.91251150638087	3
PPP2R5C	0.907443958405756	3
GDA	0.903934488431557	3
FADD	0.896484525722531	3
ARHGEF11	0.88640716738081	2
TRAF6	0.874719404508762	2
IKBKG	0.867133199590082	3
TRAF5	0.863327416426504	2
TULP3	0.858474383654667	3
PLK2	0.853572511793026	2
CD40	0.848291426145035	2
RELB	0.839503561934977	2
TNFRSF25	0.832197051909666	2
MAP4K2	0.828009092075309	2
RASAL2	0.826337626444088	2

SNX17	0.82362604466056	2
ACVR1	0.82362604466056	2
WBSCR16	0.82362604466056	2
LTA	0.816278067393294	2
CLDN5	0.814931516564797	2
CENTA2	0.802005952352699	2
SKP1	0.749756971344825	2
RAN	0.746305589975086	2
TMEM9B	0.738708958706129	2
TSPAN33	0.656119593661784	2
WDTC1	0.656119593661784	2
TNFRSF1A	0.650022360870953	2
ATPBD3	0.645749613601238	1
EEF1A1	0.639093484419263	1
CRK	0.628286491387127	1
TLR2	0.628286491387127	1
B3GALT4	0.628286491387127	1
OR2L13	0.628286491387127	1
FBXO18	0.628286491387127	1
NUTF2	0.628286491387127	1
MAP1LC3A	0.628286491387127	1
CYBB	0.614192139737993	1
IL1B	0.614192139737993	1
NFKB1	0.614192139737993	1
TNF	0.614192139737993	1
PLEKHB1	0.606074342701724	1
KLHL12	0.606074342701724	1
LILRB2	0.606074342701724	1
CD79B	0.563392470791866	1
GPR108	0.563392470791866	1
PHYH	0.563392470791866	1
PAK4	0.563392470791866	1
FGD3	0.535469107551487	1
HPS5	0.523799405014875	1
TRIM22	0.523799405014875	1
CTNNBL1	0.523799405014875	1
TNFSF10	0.504597079502434	1
PYCARD	0.504597079502434	1
SNX16	0.504597079502434	1
TBK1	0.502120640904808	1
SLC22A7	0.502120640904808	1
SLC19A2	0.502120640904808	1
CDK10	0.502120640904808	1
PEX1	0.501525320317267	1
HSBP1	0.501525320317267	1
SFRS7	0.501525320317267	1
B2M	0.49738219895288	1
CHRNA5	0.49738219895288	1
TSPAN13	0.49738219895288	1
MBD2	0.297063369397218	1
MAP3K7	0.212380952380952	1
NET1	0.212380952380952	1

References

1. Gregory TR: **Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma**. *Biol Rev Camb Philos Soc* 2001, **76**(1):65-101.
2. Hodgkin J: **What does a worm want with 20,000 genes?** *Genome Biol* 2001, **2**(11):COMMENT2008.
3. Pennisi E: **Genetics. Working the (gene count) numbers: finally, a firm answer?** *Science* 2007, **316**(5828):1113.
4. Levine M, Tjian R: **Transcription regulation and animal diversity**. *Nature* 2003, **424**(6945):147-151.
5. Messina DN, Glasscock J, Gish W, Lovett M: **An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression**. *Genome Res* 2004, **14**(10B):2041-2047.
6. Thomas MC, Chiang CM: **The general transcription machinery and general cofactors**. *Crit Rev Biochem Mol Biol* 2006, **41**(3):105-178.
7. Lee TI, Young RA: **Transcription of eukaryotic protein-coding genes**. *Annu Rev Genet* 2000, **34**:77-137.
8. Bondarenko VA, Liu YV, Jiang YI, Studitsky VM: **Communication over a large distance: enhancers and insulators**. *Biochem Cell Biol* 2003, **81**(3):241-251.
9. Tan K, Tegner J, Ravasi T: **Integrated approaches to uncovering transcription regulatory networks in mammalian cells**. *Genomics* 2008, **91**(3):219-231.
10. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays**. *Nat Genet* 1999, **21**(1 Suppl):33-37.
11. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF**. *Nature* 2001, **409**(6819):533-538.
12. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E *et al*: **Genome-wide location and function of DNA binding proteins**. *Science* 2000, **290**(5500):2306-2309.
13. Hu Z, Killion PJ, Iyer VR: **Genetic reconstruction of a functional transcriptional regulatory network**. *Nat Genet* 2007, **39**(5):683-687.
14. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**(2):166-176.
15. Beer MA, Tavazoie S: **Predicting gene expression from sequence**. *Cell* 2004, **117**(2):185-198.
16. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J *et al*: **Transcriptional regulatory code of a eukaryotic genome**. *Nature* 2004, **431**(7004):99-104.
17. Jiang C, Pugh BF: **Nucleosome positioning and gene regulation: advances through genomics**. *Nat Rev Genet* 2009, **10**(3):161-172.
18. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ: **Crystal structure of the nucleosome core particle at 2.8 A resolution**. *Nature* 1997, **389**(6648):251-260.

19. Kornberg RD, Lorch Y: **Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome**. *Cell* 1999, **98**(3):285-294.
20. Kouzarides T: **Chromatin modifications and their function**. *Cell* 2007, **128**(4):693-705.
21. Workman JL: **Nucleosome displacement in transcription**. *Genes Dev* 2006, **20**(15):2009-2017.
22. Sarma K, Reinberg D: **Histone variants meet their match**. *Nat Rev Mol Cell Biol* 2005, **6**(2):139-149.
23. Li B, Carey M, Workman JL: **The role of chromatin during transcription**. *Cell* 2007, **128**(4):707-719.
24. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF: **Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome**. *Nature* 2007, **446**(7135):572-576.
25. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J: **A genomic code for nucleosome positioning**. *Nature* 2006, **442**(7104):772-778.
26. Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C: **A high-resolution atlas of nucleosome occupancy in yeast**. *Nat Genet* 2007, **39**(10):1235-1244.
27. Mellor J: **The dynamics of chromatin remodeling at promoters**. *Mol Cell* 2005, **19**(2):147-157.
28. Kobor MS, Venkatasubrahmanyam S, Meneghini MD, Gin JW, Jennings JL, Link AJ, Madhani HD, Rine J: **A protein complex containing the conserved Swi2/Snf2-related ATPase Swr1p deposits histone variant H2A.Z into euchromatin**. *PLoS Biol* 2004, **2**(5):E131.
29. Audic Y, Hartley RS: **Post-transcriptional regulation in cancer**. *Biol Cell* 2004, **96**(7):479-498.
30. Bartel DP: **MicroRNAs: target recognition and regulatory functions**. *Cell* 2009, **136**(2):215-233.
31. Friedman RC, Farh KK, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome Res* 2009, **19**(1):92-105.
32. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14**. *Cell* 1993, **75**(5):843-854.
33. Ambros V: **A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans**. *Cell* 1989, **57**(1):49-57.
34. Ambros V, Horvitz HR: **Heterochronic mutants of the nematode Caenorhabditis elegans**. *Science* 1984, **226**(4673):409-416.
35. Wightman B, Burglin TR, Gatto J, Arasu P, Ruvkun G: **Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development**. *Genes Dev* 1991, **5**(10):1813-1824.
36. Wightman B, Ha I, Ruvkun G: **Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans**. *Cell* 1993, **75**(5):855-862.

37. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G: **The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans**. *Nature* 2000, **403**(6772):901-906.
38. He L, Hannon GJ: **MicroRNAs: small RNAs with a big role in gene regulation**. *Nat Rev Genet* 2004, **5**(7):522-531.
39. Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Muller P *et al*: **Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA**. *Nature* 2000, **408**(6808):86-89.
40. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs**. *Science* 2001, **294**(5543):853-858.
41. Lau NC, Lim LP, Weinstein EG, Bartel DP: **An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans**. *Science* 2001, **294**(5543):858-862.
42. Kim VN: **MicroRNA biogenesis: coordinated cropping and dicing**. *Nat Rev Mol Cell Biol* 2005, **6**(5):376-385.
43. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S *et al*: **The nuclear RNase III Drosha initiates microRNA processing**. *Nature* 2003, **425**(6956):415-419.
44. Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U: **Nuclear export of microRNA precursors**. *Science* 2004, **303**(5654):95-98.
45. Rana TM: **Illuminating the silence: understanding the structure and function of small RNAs**. *Nat Rev Mol Cell Biol* 2007, **8**(1):23-36.
46. Brodersen P, Voinnet O: **Revisiting the principles of microRNA target recognition and mode of action**. *Nat Rev Mol Cell Biol* 2009, **10**(2):141-148.
47. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.
48. Esquela-Kerscher A, Slack FJ: **Oncomirs - microRNAs with a role in cancer**. *Nat Rev Cancer* 2006, **6**(4):259-269.
49. Brodersen P, Voinnet O: **Revisiting the principles of microRNA target recognition and mode of action**. *Nat Rev Mol Cell Biol* 2009, **10**(2):141-148.
50. O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT: **c-Myc-regulated microRNAs modulate E2F1 expression**. *Nature* 2005, **435**(7043):839-843.
51. Fontana L, Pelosi E, Greco P, Racanicchi S, Testa U, Liuzzi F, Croce CM, Brunetti E, Grignani F, Peschle C: **MicroRNAs 17-5p-20a-106a control monocytopoiesis through AML1 targeting and M-CSF receptor upregulation**. *Nat Cell Biol* 2007, **9**(7):775-787.
52. Chang TC, Wentzel EA, Kent OA, Ramachandran K, Mullendore M, Lee KH, Feldmann G, Yamakuchi M, Ferlito M, Lowenstein CJ *et al*: **Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis**. *Mol Cell* 2007, **26**(5):745-752.
53. He L, He X, Lim LP, de Stanchina E, Xuan Z, Liang Y, Xue W, Zender L, Magnus J, Ridzon D *et al*: **A microRNA component of the p53 tumour suppressor network**. *Nature* 2007, **447**(7148):1130-1134.

54. Chen JF, Mandel EM, Thomson JM, Wu Q, Callis TE, Hammond SM, Conlon FL, Wang DZ: **The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation**. *Nat Genet* 2006, **38**(2):228-233.
55. Dews M, Homayouni A, Yu D, Murphy D, Sevignani C, Wentzel E, Furth EE, Lee WM, Enders GH, Mendell JT *et al*: **Augmentation of tumor angiogenesis by a Myc-activated microRNA cluster**. *Nat Genet* 2006, **38**(9):1060-1065.
56. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.
57. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B: **A high-resolution map of active promoters in the human genome**. *Nature* 2005, **436**(7052):876-880.
58. Kim J, Bhinge AA, Morgan XC, Iyer VR: **Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment**. *Nat Methods* 2005, **2**(1):47-53.
59. Bhinge AA, Kim J, Euskirchen GM, Snyder M, Iyer VR: **Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE)**. *Genome Res* 2007, **17**(6):910-916.
60. Attwooll C, Lazzerini Denchi E, Helin K: **The E2F family: specific functions and overlapping interests**. *EMBO J* 2004, **23**(24):4709-4716.
61. Gaubatz S, Lees JA, Lindeman GJ, Livingston DM: **E2F4 is exported from the nucleus in a CRM1-dependent manner**. *Mol Cell Biol* 2001, **21**(4):1384-1392.
62. Balciunaite E, Spektor A, Lents NH, Cam H, Te Riele H, Scime A, Rudnicki MA, Young R, Dynlacht BD: **Pocket protein complexes are recruited to distinct targets in quiescent and proliferating cells**. *Mol Cell Biol* 2005, **25**(18):8166-8178.
63. Schwemmle S, Pfeifer GP: **Genomic structure and mutation screening of the E2F4 gene in human tumors**. *Int J Cancer* 2000, **86**(5):672-677.
64. Xu X, Bieda M, Jin VX, Rabinovich A, Oberley MJ, Green R, Farnham PJ: **A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members**. *Genome Res* 2007, **17**(11):1550-1561.
65. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.
66. Platanias LC: **Mechanisms of type-I- and type-II-interferon-mediated signalling**. *Nat Rev Immunol* 2005, **5**(5):375-386.
67. Ramana CV, Chatterjee-Kishore M, Nguyen H, Stark GR: **Complex roles of Stat1 in regulating gene expression**. *Oncogene* 2000, **19**(21):2619-2627.
68. Hartman SE, Bertone P, Nath AK, Royce TE, Gerstein M, Weissman S, Snyder M: **Global changes in STAT target selection and transcription regulation upon interferon treatments**. *Genes Dev* 2005, **19**(24):2953-2968.
69. Eferl R, Wagner EF: **AP-1: a double-edged sword in tumorigenesis**. *Nat Rev Cancer* 2003, **3**(11):859-868.

70. Adhikary S, Eilers M: **Transcriptional regulation and transformation by Myc proteins**. *Nat Rev Mol Cell Biol* 2005, **6**(8):635-645.
71. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M *et al*: **Distribution of NF-kappaB-binding sites across human chromosome 22**. *Proc Natl Acad Sci U S A* 2003, **100**(21):12247-12252.
72. Kawai T, Akira S, Reed JC: **ZIP kinase triggers apoptosis from nuclear PML oncogenic domains**. *Mol Cell Biol* 2003, **23**(17):6174-6186.
73. Liu Z, Lu H, Jiang Z, Pastuszyn A, Hu CA: **Apolipoprotein l6, a novel proapoptotic Bcl-2 homology 3-only protein, induces mitochondria-mediated apoptosis in cancer cells**. *Mol Cancer Res* 2005, **3**(1):21-31.
74. Stephanou A, Brar BK, Knight RA, Latchman DS: **Opposing actions of STAT-1 and STAT-3 on the Bcl-2 and Bcl-x promoters**. *Cell Death Differ* 2000, **7**(3):329-330.
75. Stephanou A, Latchman DS: **Opposing actions of STAT-1 and STAT-3**. *Growth Factors* 2005, **23**(3):177-182.
76. Wesemann DR, Benveniste EN: **STAT-1 alpha and IFN-gamma as modulators of TNF-alpha signaling in macrophages: regulation and functional implications of the TNF receptor 1:STAT-1 alpha complex**. *J Immunol* 2003, **171**(10):5313-5319.
77. Hook EB, Regal RR: **Capture-recapture methods in epidemiology: methods and limitations**. *Epidemiol Rev* 1995, **17**(2):243-264.
78. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD: **E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints**. *Genes Dev* 2002, **16**(2):245-256.
79. Tabach Y, Brosh R, Buganim Y, Reiner A, Zuk O, Yitzhaky A, Koudritsky M, Rotter V, Domany E: **Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site**. *PLoS One* 2007, **2**(8):e807.
80. Zheng N, Fraenkel E, Pabo CO, Pavletich NP: **Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP**. *Genes Dev* 1999, **13**(6):666-674.
81. Eden E, Lipson D, Yogev S, Yakhini Z: **Discovering motifs in ranked lists of DNA sequences**. *PLoS Comput Biol* 2007, **3**(3):e39.
82. Yang J, Song K, Krebs TL, Jackson MW, Danielpour D: **Rb/E2F4 and Smad2/3 link survivin to TGF-beta-induced apoptosis and tumor progression**. *Oncogene* 2008, **27**(40):5326-5338.
83. Zwicker J, Lucibello FC, Wolfraim LA, Gross C, Truss M, Engeland K, Muller R: **Cell cycle regulation of the cyclin A, cdc25C and cdc2 genes is based on a common mechanism of transcriptional repression**. *EMBO J* 1995, **14**(18):4514-4522.
84. Caretti G, Salsi V, Vecchi C, Imbriano C, Mantovani R: **Dynamic recruitment of NF-Y and histone acetyltransferases on cell-cycle promoters**. *J Biol Chem* 2003, **278**(33):30435-30440.

85. Mendell JT: **miRiad roles for the miR-17-92 cluster in development and disease**. *Cell* 2008, **133**(2):217-222.
86. Sampson VB, Rong NH, Han J, Yang Q, Aris V, Soteropoulos P, Petrelli NJ, Dunn SP, Krueger LJ: **MicroRNA let-7a down-regulates MYC and reverts MYC-induced growth in Burkitt lymphoma cells**. *Cancer Res* 2007, **67**(20):9762-9770.
87. Chen CR, Kang Y, Siegel PM, Massague J: **E2F4/5 and p107 as Smad cofactors linking the TGFbeta receptor to c-myc repression**. *Cell* 2002, **110**(1):19-32.
88. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.
89. Dimova DK, Dyson NJ: **The E2F transcriptional network: old acquaintances with new faces**. *Oncogene* 2005, **24**(17):2810-2826.
90. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: **Genome-scale identification of nucleosome positions in S. cerevisiae**. *Science* 2005, **309**(5734):626-630.
91. Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z: **Nucleosome positioning signals in genomic DNA**. *Genome Res* 2007, **17**(8):1170-1177.
92. Rando OJ, Ahmad K: **Rules and regulation in the primary structure of chromatin**. *Curr Opin Cell Biol* 2007, **19**(3):250-256.
93. Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR: **Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation**. *PLoS Biol* 2008, **6**(3):e65.
94. Fascher KD, Schmitz J, Horz W: **Structural and functional requirements for the chromatin transition at the PHO5 promoter in Saccharomyces cerevisiae upon PHO5 activation**. *J Mol Biol* 1993, **231**(3):658-667.
95. Hahn JS, Hu Z, Thiele DJ, Iyer VR: **Genome-wide analysis of the biology of stress responses through heat shock transcription factor**. *Mol Cell Biol* 2004, **24**(12):5249-5256.
96. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD: **Evidence for nucleosome depletion at active regulatory regions genome-wide**. *Nat Genet* 2004, **36**(8):900-905.
97. Gorner W, Durchschlag E, Martinez-Pastor MT, Estruch F, Ammerer G, Hamilton B, Ruis H, Schuller C: **Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity**. *Genes Dev* 1998, **12**(4):586-597.
98. Marion RM, Regev A, Segal E, Barash Y, Koller D, Friedman N, O'Shea EK: **Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression**. *Proc Natl Acad Sci U S A* 2004, **101**(40):14315-14322.
99. Lieb JD, Liu X, Botstein D, Brown PO: **Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association**. *Nat Genet* 2001, **28**(4):327-334.
100. Rudra D, Zhao Y, Warner JR: **Central role of Ifh1p-Fhl1p interaction in the synthesis of yeast ribosomal proteins**. *EMBO J* 2005, **24**(3):533-542.

101. Reid JL, Iyer VR, Brown PO, Struhl K: **Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase**. *Mol Cell* 2000, **6**(6):1297-1307.
102. Miyake T, Loch CM, Li R: **Identification of a multifunctional domain in autonomously replicating sequence-binding factor 1 required for transcriptional activation, DNA replication, and gene silencing**. *Mol Cell Biol* 2002, **22**(2):505-516.
103. Kent NA, Eibert SM, Mellor J: **Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast**. *J Biol Chem* 2004, **279**(26):27116-27123.
104. Wade JT, Hall DB, Struhl K: **The transcription factor Ifh1 is a key regulator of yeast ribosomal protein genes**. *Nature* 2004, **432**(7020):1054-1058.
105. Schawalder SB, Kabani M, Howald I, Choudhury U, Werner M, Shore D: **Growth-regulated recruitment of the essential yeast ribosomal protein gene activator Ifh1**. *Nature* 2004, **432**(7020):1058-1061.
106. Zanton SJ, Pugh BF: **Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock**. *Genes Dev* 2006, **20**(16):2250-2265.
107. Guillemette B, Bataille AR, Gevry N, Adam M, Blanchette M, Robert F, Gaudreau L: **Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning**. *PLoS Biol* 2005, **3**(12):e384.
108. Whitehouse I, Rando OJ, Delrow J, Tsukiyama T: **Chromatin remodelling at promoters suppresses antisense transcription**. *Nature* 2007, **450**(7172):1031-1035.
109. Garcia-Martinez J, Aranda A, Perez-Ortin JE: **Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms**. *Mol Cell* 2004, **15**(2):303-313.
110. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, Trent JM, Staudt LM, Hudson J, Jr., Boguski MS *et al*: **The transcriptional program in the response of human fibroblasts to serum**. *Science* 1999, **283**(5398):83-87.
111. Gu J, Iyer VR: **PI3K signaling and miRNA expression during the response of quiescent human fibroblasts to distinct proliferative stimuli**. *Genome Biol* 2006, **7**(5):R42.
112. Liu H, Adler AS, Segal E, Chang HY: **A Transcriptional Program Mediating Entry into Cellular Quiescence**. *PLoS Genet* 2007, **3**(6):e91.
113. Ho A, Dowdy SF: **Regulation of G(1) cell-cycle progression by oncogenes and tumor suppressor genes**. *Curr Opin Genet Dev* 2002, **12**(1):47-52.
114. Dvorak HF: **Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing**. *N Engl J Med* 1986, **315**(26):1650-1659.
115. Bissell MJ, Radisky D: **Putting tumours in context**. *Nat Rev Cancer* 2001, **1**(1):46-54.

116. Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, Montgomery K, Chi JT, van de Rijn M, Botstein D, Brown PO: **Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds**. *PLoS Biol* 2004, **2**(2):E7.
117. Blackwood EM, Eisenman RN: **Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc**. *Science* 1991, **251**(4998):1211-1217.
118. Blackwood EM, Luscher B, Kretzner L, Eisenman RN: **The Myc:Max protein complex and cell growth regulation**. *Cold Spring Harb Symp Quant Biol* 1991, **56**:109-117.
119. Boxer LM, Dang CV: **Translocations involving c-myc and c-myc function**. *Oncogene* 2001, **20**(40):5595-5610.
120. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of mammalian microRNA targets**. *Cell* 2003, **115**(7):787-798.
121. Ziegelbauer JM, Sullivan CS, Ganem D: **Tandem array-based expression screens identify host mRNA targets of virus-encoded microRNAs**. *Nat Genet* 2009, **41**(1):130-134.
122. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP: **The impact of microRNAs on protein output**. *Nature* 2008, **455**(7209):64-71.
123. Kim HK, Lee YS, Sivaprasad U, Malhotra A, Dutta A: **Muscle-specific microRNA miR-206 promotes muscle differentiation**. *J Cell Biol* 2006, **174**(5):677-687.
124. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM: **Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs**. *Nature* 2005, **433**(7027):769-773.
125. Coller HA, Sang L, Roberts JM: **A new description of cellular quiescence**. *PLoS Biol* 2006, **4**(3):e83.
126. Reimer T, Brcic M, Schweizer M, Jungi TW: **poly(I:C) and LPS induce distinct IRF3 and NF-kappaB signaling during type-I IFN and TNF responses in human macrophages**. *J Leukoc Biol* 2008, **83**(5):1249-1257.
127. Haller O, Kochs G, Weber F: **The interferon response circuit: induction and suppression by pathogenic viruses**. *Virology* 2006, **344**(1):119-130.
128. Decker T, Muller M, Stockinger S: **The yin and yang of type I interferon activity in bacterial infection**. *Nat Rev Immunol* 2005, **5**(9):675-687.
129. Liu H, Deng X, Shyu YJ, Li JJ, Taparowsky EJ, Hu CD: **Mutual regulation of c-Jun and ATF2 by transcriptional activation and subcellular localization**. *EMBO J* 2006, **25**(5):1058-1069.
130. Falvo JV, Parekh BS, Lin CH, Fraenkel E, Maniatis T: **Assembly of a functional beta interferon enhanceosome is dependent on ATF-2-c-jun heterodimer orientation**. *Mol Cell Biol* 2000, **20**(13):4814-4825.
131. Hayden MS, Ghosh S: **Signaling to NF-kappaB**. *Genes Dev* 2004, **18**(18):2195-2224.
132. Perkins ND: **Integrating cell-signalling pathways with NF-kappaB and IKK function**. *Nat Rev Mol Cell Biol* 2007, **8**(1):49-62.

133. Okamura S, Arakawa H, Tanaka T, Nakanishi H, Ng CC, Taya Y, Monden M, Nakamura Y: **p53DINP1, a p53-inducible gene, regulates p53-dependent apoptosis**. *Mol Cell* 2001, **8**(1):85-94.
134. Ellisen LW, Ramsayer KD, Johannessen CM, Yang A, Beppu H, Minda K, Oliner JD, McKeon F, Haber DA: **REDD1, a developmentally regulated transcriptional target of p63 and p53, links p63 to regulation of reactive oxygen species**. *Mol Cell* 2002, **10**(5):995-1005.
135. Marcotte R, Chen JM, Huard S, Wang E: **c-Myc creates an activation loop by transcriptionally repressing its own functional inhibitor, hMad4, in young fibroblasts, a loop lost in replicatively senescent fibroblasts**. *J Cell Biochem* 2005, **96**(5):1071-1085.
136. Kime L, Wright SC: **Mad4 is regulated by a transcriptional repressor complex that contains Miz-1 and c-Myc**. *Biochem J* 2003, **370**(Pt 1):291-298.
137. Martinez NJ, Ow MC, Barrasa MI, Hammell M, Sequerra R, Doucette-Stamm L, Roth FP, Ambros VR, Walhout AJ: **A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity**. *Genes Dev* 2008, **22**(18):2535-2549.
138. Hobert O: **Gene regulation by transcription factors and microRNAs**. *Science* 2008, **319**(5871):1785-1786.
139. Killion P: **Fungus to fibroblast: A functional genomic exploration of eukaryotic transcriptional regulation**. Austin: University of Texas at Austin; 2007.
140. Sonkoly E, Wei T, Janson PC, Saaf A, Lundeberg L, Tengvall-Linder M, Norstedt G, Alenius H, Homey B, Scheynius A *et al*: **MicroRNAs: novel regulators involved in the pathogenesis of Psoriasis?** *PLoS ONE* 2007, **2**(7):e610.
141. Sullivan CS, Grundhoff AT, Tevethia S, Pipas JM, Ganem D: **SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells**. *Nature* 2005, **435**(7042):682-686.
142. Cullen BR: **Viral and cellular messenger RNA targets of viral microRNAs**. *Nature* 2009, **457**(7228):421-425.
143. Randall G, Panis M, Cooper JD, Tellinghuisen TL, Sukhodolets KE, Pfeffer S, Landthaler M, Landgraf P, Kan S, Lindenbach BD *et al*: **Cellular cofactors affecting hepatitis C virus infection and replication**. *Proc Natl Acad Sci U S A* 2007, **104**(31):12884-12889.
144. Jiang J, Gusev Y, Aderca I, Mettler TA, Nagorney DM, Brackett DJ, Roberts LR, Schmittgen TD: **Association of MicroRNA expression in hepatocellular carcinomas with hepatitis infection, cirrhosis, and patient survival**. *Clin Cancer Res* 2008, **14**(2):419-427.
145. Ibarra I, Erlich Y, Muthuswamy SK, Sachidanandam R, Hannon GJ: **A role for microRNAs in maintenance of mouse mammary epithelial progenitor cells**. *Genes Dev* 2007, **21**(24):3238-3243.
146. Chang TC, Yu D, Lee YS, Wentzel EA, Arking DE, West KM, Dang CV, Thomas-Tikhonenko A, Mendell JT: **Widespread microRNA repression by Myc contributes to tumorigenesis**. *Nat Genet* 2008, **40**(1):43-50.

147. Iliopoulos D, Malizos KN, Oikonomou P, Tsezou A: **Integrative microRNA and proteomic approaches identify novel osteoarthritis genes and their collaborative metabolic and inflammatory networks**. *PLoS ONE* 2008, **3**(11):e3740.
148. Halsey TA, Yang L, Walker JR, Hogenesch JB, Thomas RS: **A functional map of NFkappaB signaling identifies novel modulators and multiple system controls**. *Genome Biol* 2007, **8**(6):R104.
149. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets**. *Cell* 2005, **120**(1):15-20.
150. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M *et al*: **Combinatorial microRNA target predictions**. *Nat Genet* 2005, **37**(5):495-500.
151. Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard D, Gagneur J, Ghidelli S *et al*: **A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway**. *Nat Cell Biol* 2004, **6**(2):97-105.
152. Ejeskar K, Krona C, Caren H, Zaibak F, Li L, Martinsson T, Ioannou PA: **Introduction of in vitro transcribed ENO1 mRNA into neuroblastoma cells induces cell death**. *BMC Cancer* 2005, **5**:161.
153. Michalak EM, Villunger A, Adams JM, Strasser A: **In several cell types tumour suppressor p53 induces apoptosis largely via Puma but Noxa can contribute**. *Cell Death Differ* 2008, **15**(6):1019-1029.
154. Li B, Cong F, Tan CP, Wang SX, Goff SP: **Aph2, a protein with a zf-DHHC motif, interacts with c-Abl and has pro-apoptotic activity**. *J Biol Chem* 2002, **277**(32):28870-28876.
155. Ho JS, Ma W, Mao DY, Benchimol S: **p53-Dependent transcriptional repression of c-myc is required for G1 cell cycle arrest**. *Mol Cell Biol* 2005, **25**(17):7423-7431.
156. Ragimov N, Krauskopf A, Navot N, Rotter V, Oren M, Aloni Y: **Wild-type but not mutant p53 can repress transcription initiation in vitro by interfering with the binding of basal transcription factors to the TATA motif**. *Oncogene* 1993, **8**(5):1183-1193.
157. Levy N, Yonish-Rouach E, Oren M, Kimchi A: **Complementation by wild-type p53 of interleukin-6 effects on M1 cells: induction of cell cycle exit and cooperativity with c-myc suppression**. *Mol Cell Biol* 1993, **13**(12):7942-7952.
158. Feo S, Arcuri D, Piddini E, Passantino R, Giallongo A: **ENO1 gene product binds to the c-myc promoter and acts as a transcriptional repressor: relationship with Myc promoter-binding protein 1 (MBP-1)**. *FEBS Lett* 2000, **473**(1):47-52.
159. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA targets**. *PLoS Biol* 2004, **2**(11):e363.
160. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, Lim B, Rigoutsos I: **A pattern-based method for the identification of MicroRNA**

binding sites and their corresponding heteroduplexes. *Cell* 2006, **126**(6):1203-1217.
161. Karginov FV, Conaco C, Xuan Z, Schmidt BH, Parker JS, Mandel G, Hannon GJ: **A biochemical approach to identifying microRNA targets**. *Proc Natl Acad Sci U S A* 2007, **104**(49):19291-19296.
162. Hendrickson DG, Hogan DJ, Herschlag D, Ferrell JE, Brown PO: **Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance**. *PLoS One* 2008, **3**(5):e2126.
163. Chi SW, Zang JB, Mele A, Darnell RB: **Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps**. *Nature* 2009.
164. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**(1):57-63.
165. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions**. *Science* 2007, **316**(5830):1497-1502.

Vita

Akshay Anant Bhinge attended Mulund High School till class Xth and Kelkar College till class XIIth after which he joined the Grant Medical College and Sir JJ group of hospitals for his Bachelor of Medicine and Bachelor of Surgery (M.B.B.S.) degree. He received the M.B.B.S. degree in 2000 and the Master of Technology in Biomedical Engineering degree from the Indian Institute of Technology, Bombay in 2002. Later, he worked as a research assistant for one year at the Indian Institute of Science, Bangalore. In Fall 2003, he entered graduate school at the University of Texas at Austin as a Ph.D. student in the Institute of Cell and Molecular Biology.

Selected publications:

1. "Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation." *Shivaswamy S, *Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. PLoS Biol. 2008 Mar 18;6(3):e65. (*Equal contribution)

2. "Mapping the chromosomal targets of STAT1 by sequence tag analysis of genomic enrichment (STAGE)." *Bhinge AA, *Kim J, Euskirchen GM, Snyder M, Iyer VR. Genome Res. 2007 Jun;17(6):910-6. (*Equal contribution)

3. "Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment." Kim J, Bhinge AA, Morgan XC, Iyer VR. Nat Methods 2005 Jan; 2, 47-53.

4. "Accurate detection of protein:ligand binding sites using molecular dynamics simulations." Bhinge A, Chakrabarti P, Uthanumallian K, Bajaj K, Chakraborty K, Varadarajan R. Structure 2004 Nov;12(11):1989-99.

Permanent address: 206 W 38th St, Apt 221, Austin, TX 78705, U.S.A.

This dissertation was typed by the author.