DISSECTING GENETIC DETERMINANTS OF TRANSCRIPTION FACTOR ACTIVITY

EUNJEE LEE

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

COLUMBIA UNIVERSITY

2013

ABSTRACT

Dissecting Genetic Determinants of Transcription Factor Activity

Eunjee Lee

Understanding how phenotype relates to genotype, in terms of the myriad molecular processes that govern the behavior of cells and organisms, has been one of the central goals of biology for a long time. Transcription factors (TFs) play a mediating role connecting genotype with gene expression, which provides high-dimensional information about end phenotype. In particular, gene expression levels depend on their *cis*-regulatory sequence bound by TFs and condition-specific regulatory activity of TFs determined by its modulators through interaction with cofactors or signaling molecules. This thesis consists of two parts that related to the overall goal of dissecting upstream modulators of transcription factor activity. The first study is to dissect genetic determinants of transcription factor activity in a segregating population. We exploit prior knowledge about the *in vitro* DNA-binding specificity of a TF in order to map the loci ('aQTLs') whose inheritance modulates its protein-level regulatory activity. The second study is to identify regulatory mechanisms underlying tumorigenesis in mice by using genotyping and gene expression data across a set of 97 splenic tumors induced by retroviral insertional mutagenesis. We identify several instances of sequence-specific TFs whose activities are specifically affected by insertions mutations. Our results underscore the value of explicitly modeling TF activity and a strategy for finding upstream modulators of TF activity.

Contents

List of Tables

List of Figures

Acknowledgements

I would first like to thank my advisor, Dr. Harmen Bussemaker, for being a great scientist and mentor. Harmen has taught and showed me how to do science. I believe that his rigorous scientific training will guide me through my career. Not only has he shown me how to do science but also he helped me develop skills in scientific communications, both in presentation and writing. Moreover, through patience, generosity and caring, he showed me how to become a great mentor. I also thank the members of my thesis committee: Ron Prywes, Matthew Rockman, Brent Stockwell and Saeed Tavazoie. All four members have provided helpful advice throughout my thesis. I am grateful for the friends in graduate school for guiding me academically and supporting me personally. I am fortunate to be friends with the great people in the Bussemaker lab and I will miss spending time with them.

I would like to give the greatest thanks to God and my family for their love and support. Without my parent's encouragement, patience and prayers I would not have come to this point. I would like to give special thanks to my parents-in-law as well. Finally, I want to thank my husband, Joong Youn Cho, for being together throughout the years. I could not have done it without his endless love. Also, I give special thanks to my one-year old son, Joshua, for just being there.

Chapter 1

Introduction

Genetic variation perturbs phenotype. Understanding how phenotype relates to genotype, in terms of the myriad molecular processes that govern the behavior of cells and organisms, is one of the central goals of biology for a long time. The genetic information in DNA sequence, along with epigenetic information, controls the gene expression and phenotypes. On the road from genotype to phenotype, many layers of regulations, including chromatin state [1], transcriptional rate [2], splicing [3], mRNA localization [4], mRNA stability [5], translational rate [6], and protein stability [7], are programmed to act at the right time. Therefore, it is required to obtain additional functional information to elucidate precise regulatory networks underlying phenotypes.

Recent technology development greatly facilitated understanding the molecular process in a cell. Genome-scale technologies made possible by the emergence of microarray technology since the early 1990s [8] have allowed unbiased, genome-wide views of genetic networks in different ways. cDNA microarrays have led to analyze gene expression patterns globally in various conditions and genetic backgrounds for genome-wide modeling of transcriptional networks [9, 10]. Additionally, microarray techniques were used to determine the large-scale genotype [11, 12]. These days, as many as 500,000 SNPs can be profiled at the same time for thousands of individuals. High-throughput SNP genotyping makes it possible to perform genome-wide association

(GWA) studies for identifying disease loci. Furthermore, microarrays also have been adopted to determine the genomic binding sites of transcription factors in genome scale [13]. A more recent high-throughput sequencing technology has been used to sequence cDNA in order to get information about RNA contents at a high resolution (i.e. RNA-seq) [14], and combined with chromatin immunoprecipitation (ChIP) to identify the binding sites of DNA-associated proteins (i.e. ChIP-seq) [15].

TFs play central roles in the regulation of gene expression that further determines phenotypes (**Figure 1.1**). They bind specific DNA sequences and recruit other proteins such as chromatin remodeling factors, RNA polymerase, and histone-modifying factors to the genome sequence. Furthermore, they regulate gene expression in a condition-specific manner through interaction with cofactors or signaling molecules.

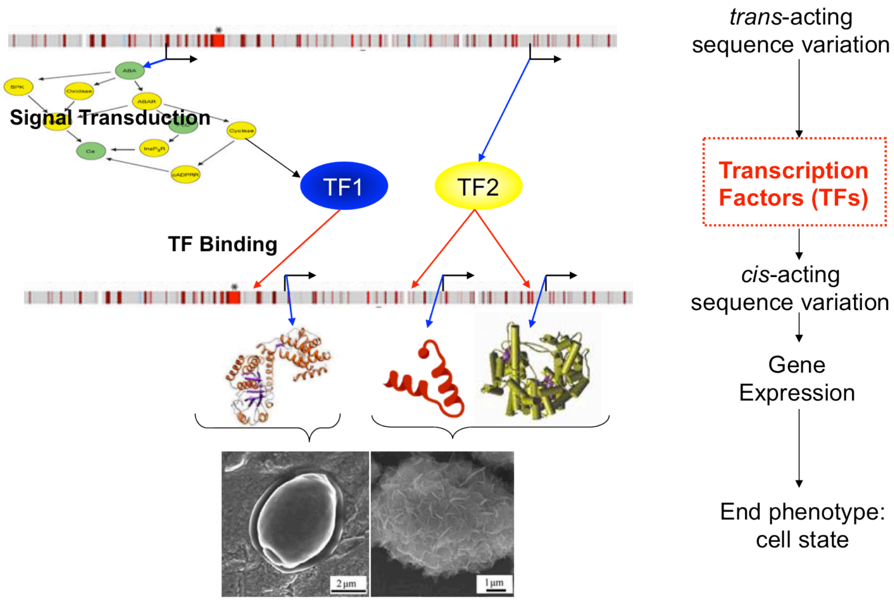

Figure 1.1: TFs as a crucial layer of regulation. The genomic information in DNA regulates gene expression level, and further end phenotypes through condition-specific regulatory activity of transcription factors

This thesis consists of two parts that both relate to the overall goal of dissecting genetic determinants of transcription factor activity. In the first part, we aim to identify genetic modulators of TFs activity. In particular, we propose a method that exploits prior knowledge about the in vitro DNA-binding specificity of a TF in order to map the loci ('aQTLs') whose inheritance modulates its protein-level regulatory activity. Genome-wide regression of differential mRNA expression on predicted promoter affinity is used to estimate segregant-specific TF activity, which is subsequently mapped as a quantitative phenotype. Furthermore, we have extended this approach to other biological contexts, including post-transcriptional regulatory networks and the promiscuous binding of TFs to high-occupancy target (HOT) regions.

In the second part, we aim to identify regulatory mechanisms underlying tumorigenesis in mice using a set of 97 splenic tumors induced by retroviral insertional mutagenesis. We present locus expression signature analysis (LESA), a novel approach that defines and exploits the gene expression signature associated with each insertion. To identify regulatory mechanisms of tumorigenesis, we hypothesized that gene expression is affected by the insertional mutations through one of two mechanisms: (i) regulation by sequence-specific transcription factors (TFs) or (ii) changes in chromosomal domain organization leading to changes in gene expression. We also investigated the relationship between drug response and my locus-specific expression signatures.

This thesis is organized into six chapters. Chapter 2 provides background to cover main concepts required to understand later chapters; background on genetical genomics approach, retroviral insertional mutagenesis screens, and key aspects of transcription factors as a crucial layer of regulation. Chapter 3 describes our aQTL approach, a method that exploits prior knowledge about the in vitro DNA-binding specificity of a TF in order to map the loci (‘aQTLs') whose inheritance modulates its protein-level regulatory activity. Chapter 4 describes the extension of the aQTL approach to other biological contexts, including the promiscuous binding of TFs to high-occupancy target (HOT) regions and post-transcriptional regulatory networks. Chapter 5 describes locus expression signature analysis (LESA), a novel approach that defines and exploits the gene expression signature associated with each insertion using gene expression levels of tumors induced by retroviral insertional mutagenesis screens in mouse. Chapter 6 summarizes our findings and provides some possible future directions for this work.

Chapter 2

Background and Literature Review

This chapter is intended to cover the main concepts required to understand later chapters, and to examine the relevant literature in the field. First, we give an introduction to 'genetical genomics' approaches [16] that use parallel high-throughput gene expression and genotype data to identify expression quantitative trait loci (eQTLs). We continue with background on retroviral insertional mutagenesis screens. Finally, we review the key aspects of transcription factors, which are crucial players in the regulation of gene expression.

2.1 Natural variations to perturbs expression levels

How phenotype relates to genotype in terms of the molecular processes that govern the behavior of the cell is one of the central questions in biology. After the rediscovery of Mendel's paper in 1900, the genetic variation in populations became a subject of scientific inquiry for a long time. Much effort has been devoted to identifying causative DNA variants and elucidating the mechanisms underlying diverse phenotype. Gene expression levels represent an intermediate molecular phenotype of great utility. They provide high-dimensional information about the cellular state. Gene expression also provides a universal sub-phenotype for complex and heterogeneous organismal phenotypes. Furthermore, the emergence of microarray technology in the early 1990s has greatly facilitated quantitative measurements of mRNA abundance, and made it possible to analyze gene expression patterns globally in order to understand the molecular pathways underlying the organismal phenotypes.

Dissecting natural variations in gene expression improves our understanding of the molecular processes underlying phenotypes. In this section, we give an overview of the recent strategies used to find the natural variations that perturb gene expression, which are called expression quantitative trait loci (eQTLs), and describe the features of the two types of regulatory sequence variation underlying differences in gene expression as well as methods for identifying master regulators responsible for the large number of genes.

2.1.1 Parallel use of high-throughput genotype and gene expression data

Building on a long history in small scale studies to understand genetic variation in gene expression [17, 18], Jansen and Nap proposed genetic mapping of genome-wide gene expression, which is called 'genetical genomics' [16] (**Figure 2.1**). The use of parallel high-throughput genotyping and expression profiling on segregating populations has enabled researchers to ask quantitative questions regarding the genetics of genome-wide expression in a variety of organisms [16, 19-22]. Genetical genomics has provided substantial additional insight into the functional landscape of gene regulation, improved our understanding of transcriptional regulation and regulatory variation, and provided a new approach for connecting DNA sequence variation with phenotype variation.

For any genetic study it is required to demonstrate that the trait in question is influenced by inherited factors. Although the exact heritability estimates depends on factors such as sample size, tissue type, statistical mode, amount of genetic diversity and environmental variability [23], genetical genomics studies have revealed that hundreds to thousands of steady-state mRNA abundance for individual genes are highly heritable. Therefore, mRNA abundance can be treated as a quantitative trait in classical genetics methods (**Figure 2.1**), and the expression quantitative trait loci (eQTLs), whose allelic variation influences the expression level of individual genes, have successfully been mapped in a number of model organisms from yeast to mouse [19-22] using linkage analysis or association mapping.

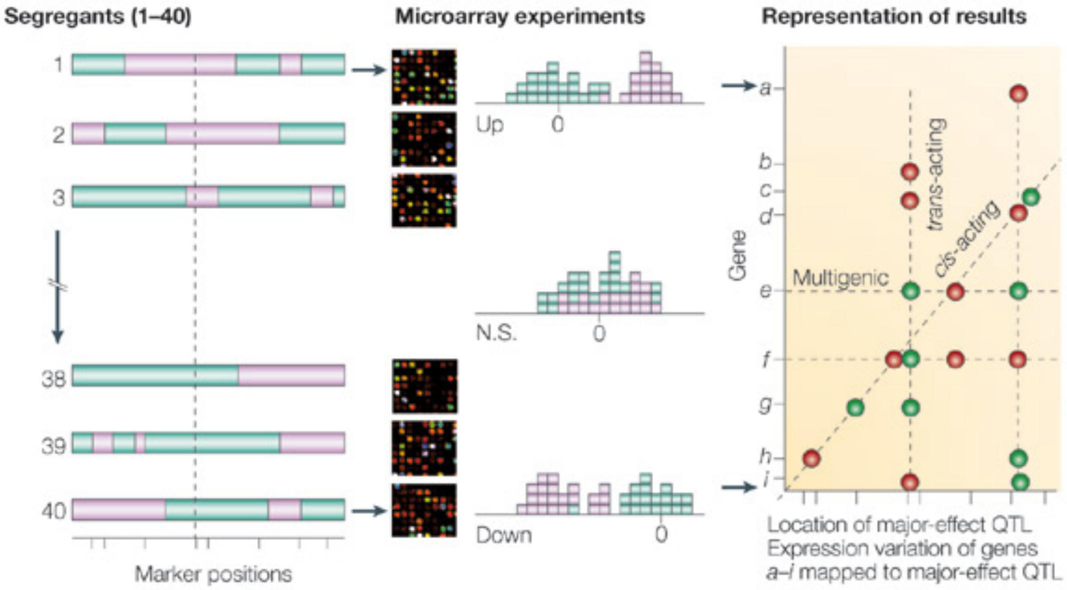

Figure 2.1: Genetical genomics approach. 40 segregants obtained by crossing the parental strains are genotyped for a set of markers that cover the genome. Two different parental alleles and genome blocks are indicated by green and blue, respectively. The expression profiling of each segregant also is obtained. The expression level of each gene is treated as a quantitative trait and analyzed across the 40 segregants with to map expression quantitative trait loci (eQTLs). This method detects the significant genes whose expression levels are explained by the genetic variation between the two parental strains through either upregulation (Up) or downregulation (Down) in gene expression (N.S., not significant change). Several gene transcripts were found that map to two or more loci. In diagram, horizontal line represents locations of eQTL and vertical line represents location of genes affected by eQTL. eQTL can be cis-acting (dotted line at an angle) or trans-acting (dotted vertical lines). Figure taken from [24]

mRNA abundance differs from other phenotypes in many ways, in regards of understanding of its genetics. First, the genome-wide gene expression profiles provide a large and unbiased set of traits that can be assayed simultaneously. The study of these traits can provide a detailed landscape of transcriptional regulation, but statistical methods for analysis of eQTL mapping are required to adjust for multiple tests due to large number of trait-marker combinations. Efforts have generally involved controlling false discovery rate (FDR), and approaches to control FDR have relied on calculations of q-values as described in Storey and Tibshirani [25] or on permutations [26]. Each transcript whose abundance treated as a trait has a corresponding encoding gene with a known position in the genome. This special feature provides mechanisms underlying DNA variations on gene expression, and will be discussed in the next section.

2.1.2 Mapping eQTLs: local versus distal QTLs

eQTL studies have extended our understanding of the contribution of *cis*- versus *trans*-acting variation to the expression of any given gene. According to the relative genomic locations of transcript and its QTL, an expression QTL can be explicitly classified as 'local' (near the genomic location of the gene encoding the transcript) or 'distal' (elsewhere in the genome) [27]. The mode of effect of regulatory variants is further investigated to understand the underlying molecular nature of eQTLs (**Figure 2.1**).

Local eQTL linkages frequently act as cis-*acting* polymorphisms [28]. The polymorphisms in the cis-regulatory region alter transcription factor binding sites, and consequently may alter transcription levels. They also affect mRNA abundance in a post-transcriptional way, such as through changes in mRNA stability or processing. Many eQTL studies have reported that the strongest eQTLs are most frequently classified as local eQTLs [29-31], suggesting that the transcription factor binding involved in RNA polymerase II recruitment or post-transcriptional regulations are key components of transcriptional regulation. On the other hand, *trans*-acting local variation is also likely to be partly responsible for the local eQTL linkages. The polymorphisms in the coding region might change the mRNA abundance by triggering feedback loops or an auto-regulatory mechanism [27]. One study experimentally confirmed such a hypothesis that the local regulatory variation acts in *trans* through a feedback loop [28].

Distal regulatory variation typically acts in *trans*. The most obvious hypothesis is that *trans*-acting polymorphisms should be enriched near transcription factor genes, or these *trans*-eQTL genes can mediate the effect of a transcription factor instead of directly encoding a transcription factor. Even though distal regulation can occur with many degrees of indirectness, consistent with the observation that distal acting elements exert a weaker influence than local QTLs [29-31], many researches have explored causal relationships underlying *trans*-eQTLs. It is because trans-acting polymorphisms at distal loci can influence the expression of large numbers of genes in countless ways by changing the state and/or connectivity of the gene regulatory network of the cell [32]. It is therefore expected that such polymorphisms account for much of the genetic variance of gene expression.

2.1.3 Methods for finding master regulators

In addition to sequence variations that affect single transcript phenotypes in *cis* or in *trans*, previous eQTL studies have shown that genomic regions containing transcription regulators influence expression levels of hundreds of genes, and seem to be responsible for most differences in gene expression [19, 21, 32, 33]. This existence of ‘master regulators’ makes it possible to enrich our understanding of regulatory networks, and many researchers have tried to identify such regulators.

Perhaps the simplest method for finding master regulators is to identify eQTL “hotspots” that influence the expression of a disproportionate number of genes [19]. A number of such hotspots have been identified in yeast and other organisms [27]. The genes that link to a particular hotspot are often enriched for specific biological functions, and tend to be controlled through the same regulatory sub-network [19, 34]. A different approach has been to map *trans*-acting loci for sets of co-expressed genes identified using hierarchical clustering [32] or more sophisticated module inference algorithms [35]. However, methods based on co-expression are most useful when a relatively small number of cell state parameters are perturbed and the expression of large subsets of genes changes in a coherent way. One expects them to be less naturally suitable for analyzing natural gene expression variation, where the segregation of alleles in a genetic cross causes a very large number of cell state parameters to be independently perturbed. Indeed, with some exceptions, the number of genes in genetic co-expression modules is very small [32, 35]. Principal Component Analysis (PCA) [36] of the matrix of genes by segregants, and extensions of PCA that incorporate qualitative information about regulatory network topology [37-39], have also been applied to map trans-acting loci. While these methods all improve upon single-gene based approaches, the lion’s share of the heritable variation in gene expression remains to be accounted for.

2.2 Mutagenesis using retrovirus to generate tumors

Cancer arises as a result of the accumulation of genetic and epigenetic changes that deregulate a specific aspect of normal cell function. High-throughput technologies for mapping genetic and chromosomal aberrations have revealed complex changes in genomes of individual tumors [40-43]. However, It is difficult to determine whether such mutations are causal "driver" mutation or an incidental "passenger" mutation. Genetic screens can greatly facilitate to identify causal genes involved in tumorigenesis in model organisms such as the mouse. Retroviral insertional mutagenesis screens in mice are efficient tools for identification of oncogenic mutations in an *in vivo* setting [44, 45]. Recent advances in sequencing technology and availability of the mouse genome sequence have had a large effect on the potential of insertional mutagenesis screens. In this chapter, we discuss the detailed features of retroviral insertional mutagenesis screens, including the mechanisms of mutation and the previous oncogenes discovered from this screening method, and identification of common insertions sites.

2.2.1 Oncogenic retroviruses as a tool for genetic screening

Oncogenic retroviruses can cause cancer in various species. They generally can be divided into two classes according to their mechanisms: acute and slow transforming viruses. These two classes have two distinct molecular mechanisms of retroviral oncogenesis. The Abelson Murine leukaemia Virus (AML) [46] and avian myoblastosis virus (AMV) [47] are acute transforming retroviruses and express viral oncogene v-Abl and v-Myb, respectively. This type of retrovirus usually generates tumors within 2-3 weeks after infection through expression of virally encoded oncogenic versions of normal cellular genes in host cells.

Slow transforming retroviruses do not carry viral oncogenes, but can induce tumors through mutation of the cellular genes caused by integration of their proviruses into the host genome. This type of retrovirus usually induces tumors within 3-12 months. Elements in the proviral genome that regulate the viral transcript act in *cis* on cellular gene transcripts. Such oncogenic mutations by these elements may either cause alteration of a gene product or influence the expression levels of one or more genes surrounding the insertion depending on where the provirus integrate into. Genetic screens using slow transforming retrovirus have efficiently been used to discovery cancer genes in various model organisms like cats (FeLV), birds (ALV and REV) and mice (murine leukaemia virus (MuLV)).

2.2.2 Proviral insertions drive tumorigenesis

To initiate murine leukemia virus (MuLV)-mediated insertional mutagenesis, newborn mice are infected with MuLV by injection with virus-producing cells (**Figure 2.2**). The virus will infect host cells, and the insertion of proviruses in the genome of the host cell can mutate cellular genes in multiple ways. The affected cell will expand in a clonal outgrowth given that the mutations caused by provirus insertions have a selective advantage. Cells acquiring multiple mutations induced by repeated infections can generate the development of a tumor. The proviral insertion site can be easily identified by amplification of the genome sequences flanking the retroviral insertion, and mapping the resulting sequences on to the genome.

The ability to induce multiple mutations in the same cell makes slow retroviruses very suitable for genetic screens for oncogenic mutations [44, 45]. It is not straightforward to acquire multiple mutations from retroviral infection. This is because after incorporation of the provirus into the genome, cells will start producing viral envelope proteins and cell surface receptors become occupied with these proteins, thus inhibiting reinfection of the cell. Mutant viruses encoding non-functional envelope proteins created by recombination of the viral sequences with endogenous viral sequences can utilize different receptors, and re-infect an infected cell [44, 45].

The proviral insertions in the genome of the host cell can mutate cellular genes in different ways [44]. It results in enhanced transcripts levels, viral-host fusion transcripts or truncation of gene transcripts. In particular, mutation of cellular genes by proviral insertions is mediated by proviral elements that drive and regulate retroviral transcription. These elements are present in the Long Terminal Repeats (LTRs) at the end of the provirus and contain an enhancer, promoter region, the start and termination sites of transcription. The promoter sequence in the LTR such as a TATA box and GC-rich sequences can recruit the basal transcription machinery [48]. On the other hand, the enhancer region contains binding sites for various transcription factors [49-51], which is required for cell-type specificity of retrovirus.

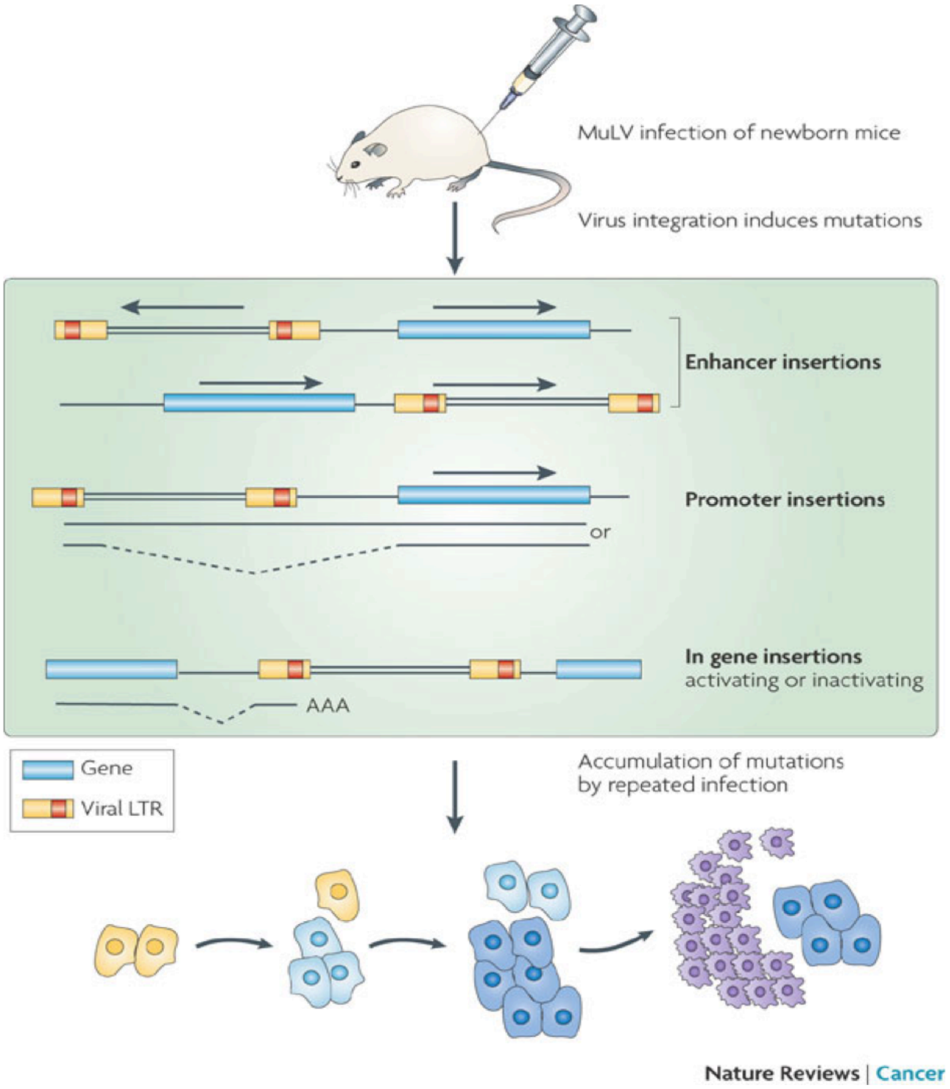
MuLV infection of newborn mice
Virus integration induces mutations
Enhancer insertions
Promoter insertions
or
In gene insertions
activating or inactivating
AAA
Gene
Viral LTR
Accumulation of mutations by repeated infection
Nature Reviews | Cancer

Figure 2.2: Retroviral insertional mutagenesis screens. The infection of mice by MuLV induces mutation in multiple ways by integration of the retrovirus into the host genome. According to the location of insertions, truncating, inactivation, overexpression or misexpression of cellular genes can be induced. Cells can be re-infected by the virus and accumulate multiple mutations that induce tumors. Figure taken from [45].

2.2.3 Identification of oncogenes and tumor suppressor genes

Many of the screens published to date identify hundreds of insertions [52-60]. Recent advances in sequencing technology and availability of the mouse genome sequence had a large effect on the potential of insertional mutagenesis screens. One recent screen yielded an average of 20 insertions per tumor, in total, 10,806 independent insertions for 510 tumors [61]. To date, more than 10,000 genomic regions where retroviral insertions are found in close proximity in multiple tumors (referred as Common Insertion Sites or CISs) have been identified in Retroviral Tagged Cancer Gene Database (RTCGD) (http://variation.osu.edu/rtcgd) [62].

Many regions that are targeted in multiple independent tumors show a significant overlap with oncogenes and tumor suppressor genes, validating the use of this screening strategy. For example, insertion by retrovirus near the c-Myc proto-oncogene is frequently found to induce tumors like erythroleukaemias and T-cell lymphomas [63]. The previous studies also detected that the proviral insertions in the 3'UTR region of the Pim1 and Nmyc gene induce tumors by removal of regulatory or destabilizing motifs in mRNAs [64, 65]. Furthermore, tumor suppressor genes such as neurofibromatosis 1 (Nf1) have been identified by retroviral tagging [59]. The large-scale mutagenesis also detected frequent insertion near tumor suppressor genes such as Ikaros, Zfpn1a3, Nf1l, Ovca2 and Wwox [61].

2.2.4 Method for finding common insertion sites

A common insertion site (CIS) is defined as a region in the genome that has been hit by viral insertions in multiple independent tumors significantly more frequently than expected by chance. Not all insertions are causal to tumor development. Non-oncogenic insertions can be detected in the tumors when this insertion co-occurs with oncogenic insertions early in the tumor development phase. However, it is unlikely to detect this non-oncogenic insertion in the other independent tumor. The insertions found at the same locus for multiple independent tumors are likely to be a result of selective expansion of tumor cells. Therefore, finding CISs greatly reduce the probability of detecting non-oncogenic insertion.

To identify CISs, researchers have calculated random distribution of insertions with fixed windows in the genome assuming a Poisson distribution [56] or using Monte Carlo simulation [59]. However, as the number of insertions in a study increases, these fixed window size can increase the probability to detect false CISs: the small window size can reduce the number of false detections, but biologically this is undesirable because some retroviruses can affect gene expression over distances much larger than 30 kb [66]. On the other hand, the large window size can increase the detection of false CISs. Therefore, more sophisticated statistical methods are required.

One study proposed a statistical framework based on Gaussian kernel convolution (GKC), which estimates a smoothed density distribution of inserts over the entire genome [67]. Depending on choice of kernel size and p-value, the total number of statistically significant CISs varies. The smaller window size separates CISs that influence the same gene, on the other hand, the increasing window size may result in merging of independent CISs. This framework can consider CISs and evaluate their significance level with varied window width, and connect to biological relevance in the behavior of CISs. They also provide the background insertion distribution required for analyzing preferential insertions near transcription start sites.

It is not straightforward to determine which genetic lesions near a CIS are playing a causal role in oncogenesis. The effect of insertions on the nearby targets is dependent on the relative position and orientation of the target transcript as well as the orientation of the viral integration. To exploit this information, the previous study has employed a rule-based mapping (RBM) procedure [68]. RBM distinguishes insertions by their occurrence within a transcript, or outside a transcript. Based on the orientation and locus of an insertion, RBM assigns one or more target transcripts to an insertion. RBM is based on the assumption that the influence of insertions on gene expression of nearby genes is dependent on the distance of the insertion to the gene. The unique list of transcripts that follows from the procedure is used to generate binary profiles that, for each tumor, indicate if a transcript is a putative target. It was observed that the proximal transcripts frequently results the same binary profile. These were therefore combined into a single profile [68].

2.3 TFs as a crucial layer of gene regulation

The genome-wide pattern of steady-state mRNA levels constitute an intermediate molecular phenotype, and their condition-specific regulation is mainly mediated by transcription factors (TFs). The transcription factors act by sequence-specific binding to regulatory DNA elements in the vicinity of their target genes. Additionally, the condition-specific regulation by TFs is explained by their regulatory activity and involvement of other proteins such as cofactors or signaling molecules. In this section, we discuss the key aspects of transcription factors as a crucial layer of gene expression regulation.

2.3.1 Predicting the sequence specificity of transcription factors

Transcription factors contain a DNA-binding domain (DBD) that determines sequence-specific binding to regulatory DNA elements. Microarray techniques used to facilitate the measurement of the sequence specific binding of TFs along the genome. Chromatin immunoprecipitation followed by microarray hybridization (ChIP-chip) and DNA adenine methyltransferase identification (DamID) have allowed genomic mapping of transcription factors in various conditions [13, 69, 70]. Nucleosome occupancy and covalent modifications to histones have also been assayed [70-72]. Recently, high-throughput sequencing, ChIP-seq, can profile high-resolution mapping of protein-DNA interactions [15]. DIP-chip and protein-binding microarrays (PBMs) have been used to determine the *in vitro* sequence specificity of DNA-binding proteins [73, 74].

It is important to model the interaction between protein and DNA sequence to have accurate quantitative information about the DNA binding specificity of TFs. Berg and Von Hippel [75] developed a theoretical framework to parameterize protein-DNA sequence specificity using position weight matrix (PWM), assuming additivity of the binding energy for each base pair. However, the position-specific scoring matrix (PSSM), whose entries can be related to binding free energies and defined from PWMs, does not represent the accurate quantitative information about sequence specificity from a biophysical view because binding energies are inferred up to an unknown scaling factor.

High-throughput technology has allowed inference of more accurate TF binding specificity using biophysical models [76-79]. In particular, Foat *et al.* [78] represents DNA binding specificity of transcription factors (TFs) in the form of position-specific affinity matrices, whose affinities at each position are directly related to the free energy of binding. Therefore, these matrices can be used to predict the accurate quantitative affinity with which each TF binds to the promoter region of each gene [80].

2.3.2 Construction of regulatory network connectivity

To understand the comprehensive regulation of gene expression by transcription factors, the regulatory connectivity, that is, which genes are the targets of TFs, should be solved. Many studies have predicted regulatory connectivity based on a genome-wide expression levels. The regulatory connectivity has been first analyzed by clustering of expression profiles across multiple conditions [81]. The resulting disjoint sets of genes are considered co-regulated and their promoter sequences are investigated to find binding of regulators [82-84].

Another class of approaches explicitly parameterizes the condition-specific activity of transcription factors using prior information about the physical interaction between TFs and DNA sequence. To construct regulatory network connectivity, some studies have relied the mRNA expression levels of the gene encoding the TF as a surrogate for the condition-specific regulatory activity of the TF [85]. However, the regulatory activity of a TF is usually not well predicted by its mRNA expression level [86]. Most TFs are modulated at the post-translational level through non-covalent modification by signaling proteins, changes in the subcellular localization of the TF protein, and the availability of co-factors. Other methods estimated post-translational activity of each TF using cis-regulatory sequence and the expression level of target genes. The condition specific TF activity is determined as expression changes in terms of counts of regulatory motifs in its promoter regions using multivariate linear regression [87] or elaborated feature selection method [88]. With availability of large-scale data, a more recent approach that is motivated by a biophysical description of gene expression regulation used prior information about sequence specificity of transcription factors to estimate TF activity [89].

Chapter 3

Identifying the genetic determinants of transcription factor activity: aQTL

This chapter has been adapted from an article by Eunjee Lee and Harmen J. Bussemaker that was published in the journal Molecular Systems Biology *[90]*.

3.1 Abstract

Analysis of parallel genotyping and expression profiling data has shown that mRNA expression levels are highly heritable. Currently, only a tiny fraction of this genetic variance can be mechanistically accounted for. The influence of *trans*-acting polymorphisms on gene expression traits is often mediated by transcription factors (TFs). We present a method that exploits prior knowledge about the *in vitro* DNA-binding specificity of a TF in order to map the loci ('aQTLs') whose inheritance modulates its protein-level regulatory activity. Genome-wide regression of differential mRNA expression on predicted promoter affinity is used to estimate segregant-specific TF activity, which is subsequently mapped as a quantitative phenotype. In budding yeast, our method identifies six times as many locus-TF associations and more than twice as many *trans*-acting loci as all existing methods combined. Application to mouse data from an F2 intercross identified an aQTL on chromosome VII modulating the activity of Zscan4 in liver cells. Our method has greatly improved statistical power over existing

methods, is mechanism based, strictly causal, computationally efficient, and generally applicable.

3.2 Introduction

We here present a transcription-factor-centric and sequence-based method for the dissection of genetic expression variation. A key feature of our approach is the use of quantitative prior information about the DNA-binding specificity of transcription factors (TFs) in the form of position-specific affinity matrices [80]. These matrices are used to predict the affinity with which each TF binds to the promoter region of each gene. We use a linear regression model motivated by a biophysical description of gene expression regulation [80, 87] to explain the genome-wide transcriptional response to the genetic perturbations in each segregant in terms of changes in ‘hidden' TF activity. Treating the latter as a quantitative trait allows us to map the activity quantitative trait loci (‘aQTLs') whose allelic status modulates the regulatory activity of specific TFs.

As we will demonstrate below, our method has a greatly improved statistical power to detect regulatory mechanisms underlying the heritability of genome-wide mRNA expression. Specifically, it identified six times as many locus-TF associations from a genetic cross between two haploid yeast strains as all existing methods combined. This includes novel *trans*-acting polymorphisms in the TF-encoding gene *STB5*,*RFX1*, and *HAP4*. We also identified 20 previously unknown *trans*-acting loci. Furthermore, for many of the 13 known eQTL hotspots in yeast, our method implicated several TFs that were not previously known to mediate the effect of inheritance of these loci on gene expression levels. We validated our ability to predict locus-TF associations in yeast using gene expression profiles for allele replacement strains. Finally, application to mouse data identified an aQTL modulating the activity of a specific TF in liver cells, demonstrating that our method also works in higher eukaryotes.

3.3 Results

We applied our method in two different organisms: budding yeast and mouse. For yeast, the data set we used [91] covers 108 haploid segregants from a cross between two haploid strains of *Saccharomyces cerevisiae*—a lab strain (BY) and a wild isolate from a vineyard (RM). It includes two-color DNA microarray measurements for each gene of the mRNA abundance in each individual segregant relative to a pooled reference consisting of equals amounts of mRNA from both parental strains, and genotyping information at 2957 genomic marker locations. The mouse data set consisted of gene expression levels in the liver cell lines of an F2 intercross population between C57BL/6J and DBA/2J (BXD) consisting of 111 animals [21], and the genotyping at 139 microsatellite markers uniformly distributed over the mouse genome [92].

3.3.1 Inferring segregant-specific TF activities

Figure 3.1 provides an overview of our computational procedure. As inputs, it requires: (i) the nucleotide sequence of the *cis*-regulatory region associated with each gene; (ii) a weight matrix for each TF, used to predict the strength with which the TF binds to each *cis*-regulatory region; (iii) a matrix containing continuous values, whose rows correspond to genes and whose columns contain the genome-wide mRNA expression profile of a particular segregant; and (iv) a genotype matrix containing binary values, whose rows correspond to genetic markers, and whose columns specify from which parent each marker was inherited in a particular segregant. As *cis*-regulatory sequence, we used 600 bp upstream of each open reading frame. We previously demonstrated that when the binding specificity of a TF is known, quantitative changes in its regulatory activity can be inferred by performing genome-wide linear regression of differential mRNA expression on the predicted *in vitro* binding affinity of *cis*-regulatory regions [93]. The biophysical foundation that underlies this regression approach requires the binding specificity of each TF to be represented as a position-specific affinity matrix (PSAM)[94]. We used an existing compendium of position weight matrices (PWMs) for yeast TFs [95], converting each PWM to an approximate PSAM by assuming base frequencies to be proportional to relative binding affinities at each position within the binding site [80]. Each PSAM was then used to estimate the segregant-specific promoter affinity for all genes (**Figure 3.1A**). With only a few exceptions, these promoter affinity profiles are not correlated between TFs. This allowed us to estimate the segregant-specific regulatory activity of most TFs in an independent manner. For each segregant, genome-wide linear regression of differential mRNA expression on segregant-specific promoter affinity for each TF was performed (**Figure 3.1B**). The coefficients from this fit represent protein-level TF activities, which we treat as a quantitative phenotype. Whenever the distribution of TF activity depends on the inheritance at a particular genomic position, this indicates the presence of an aQTL (**Figure 3.1C**). Details are provided in the Materials and methods section.

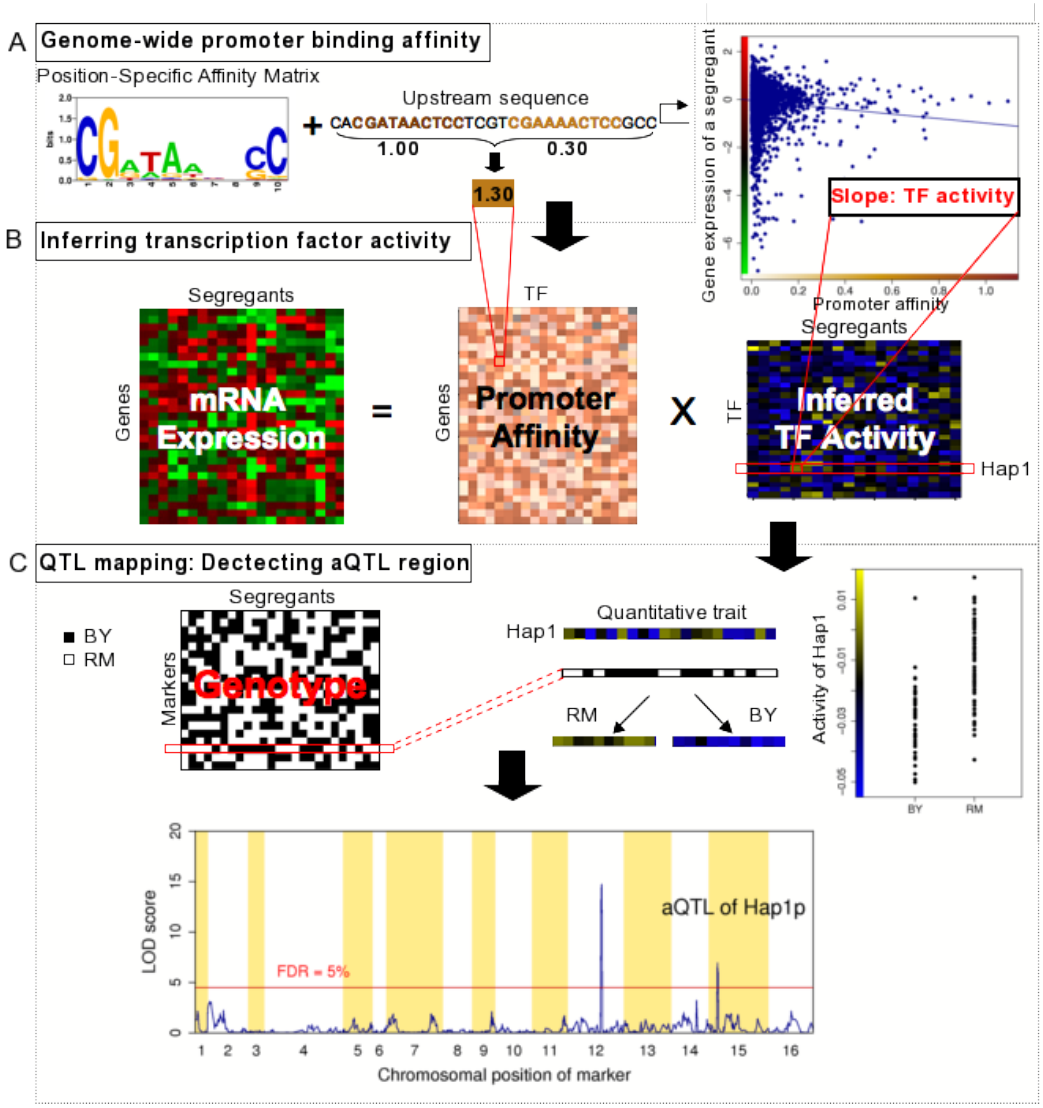

Figure 3.1: Overview of aQTL approach. (**A**) We estimated genome-wide promoter binding affinity using 600bp upstream sequence of each open reading frame and position-specific affinity matrix (PSAM). (**B**) Genome-wide promoter binding affinity was then used to infer transcription factor activity. For each segregant, the coefficients from genomewide linear regression of differential mRNA expression on promoter affinity for each TF represent protein-level TF activities. (**C**) We treat protein level TF activities as a quantitative phenotype in QTL mapping to detect aQTL region. Whenever the distribution of TF activity depends on the genotype of the specific markers at a particular genomic position, this shows high LOD score at these markers and indicates the presence of an aQTL.

3.3.2 TF activity is a highly heritable quantitative trait

To establish that the TF activities inferred by our regression procedure are meaningful, we calculated their heritability h^2 (see Materials and methods). Encouragingly, we found that the activity of 102 of the 123 TFs tested is heritable at a false discovery rate (FDR) of 5% corresponding to h^2>80.4%. In general, the heritability of the inferred TF activity is higher than that of the mRNA expression level of the gene encoding the TF (**Figure 3.3**). **Figure 3.2** shows differences in TF activity between the BY and RM parental strains as estimated by applying the regression procedure of Figure 3.1 to the average differential mRNA expression profile between BY and RM [91]. Hap1p is the factor whose regulatory activity is the most strongly modulated between the BY and RM strains. Indeed, it is known that a Ty1 insertion in the *HAP1* coding region occurs in BY and other derivatives of the lab strain S288C [96] and that this insertion is absent in RM [19]. Overall, 46 TFs are more active in RM, whereas 56 are more active in BY, at a 5% FDR. Merely comparing the two parental strains, however, does not reveal which loci are responsible for the differences in TF activity. Only genetic mapping to quantitative trait loci can provide that information.

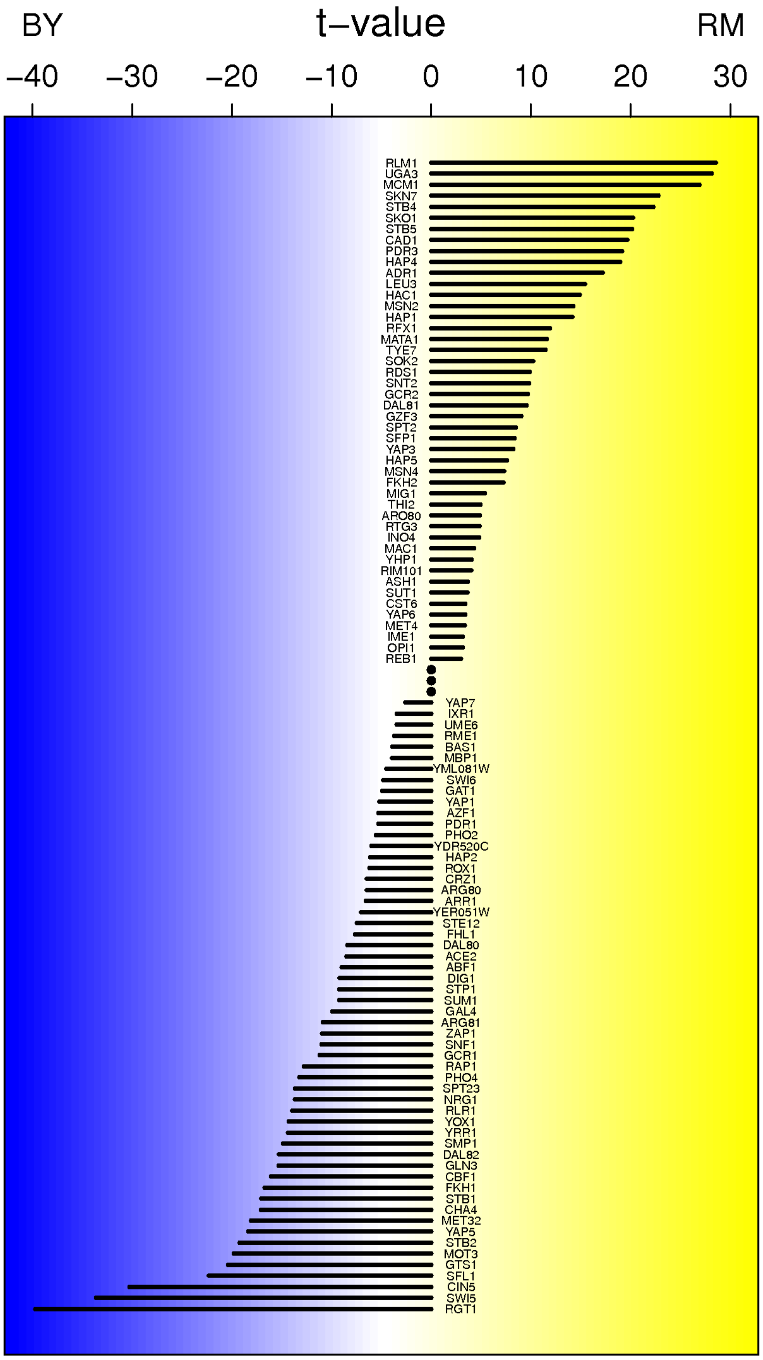

Figure 3.2: Differences in TF activities between the BY and RM parental strains. Shown are the t-values corresponding to the regression coefficients in a multivariate linear model that predicts genome-wide differential mRNA expression from predicted binding affinity of upstream promoter regions.

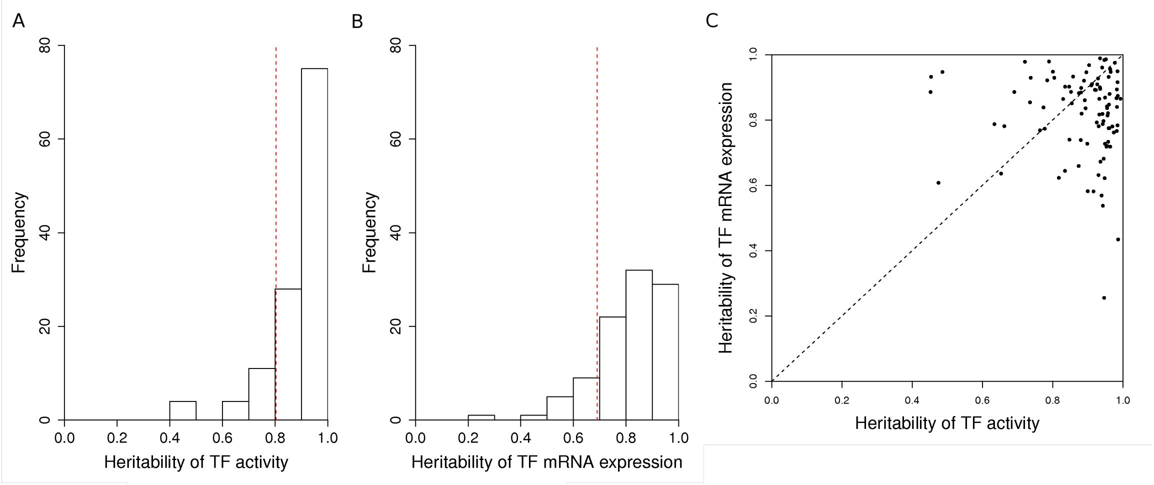

Figure 3.3: Comparing the heritability of activity and mRNA expression level of TFs. (**A**) Histogram of heritability of TF activity. (**B**) Histogram of heritability of TF mRNA expression. (**C**) Scatter plot of heritability of TF activity versus TF mRNA expression

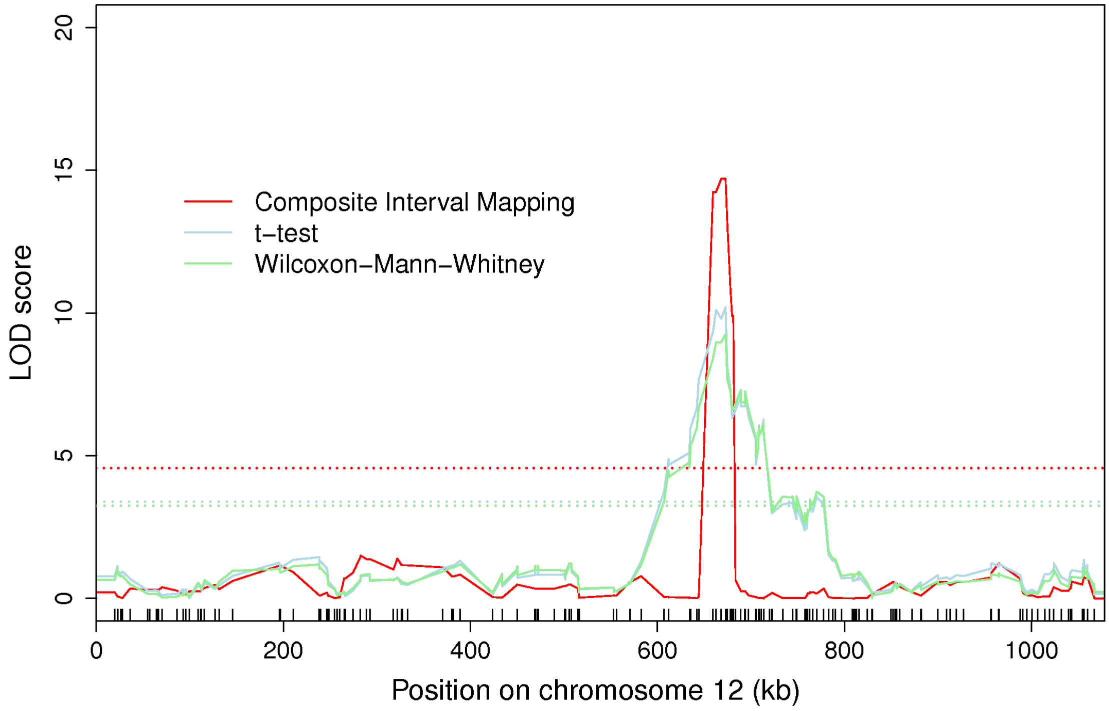

Figure 3.4: Comparison of three different methods for mapping aQTLs. Shown are LOD score profiles across the *HAP1* locus on chromosome XII obtained using (i) Composite Interval Mapping and two single markers methods, (ii) the parametric Welch's t-test, and (iii) the non-parametric Wilcoxon-Mann-Whitney test.

3.3.3 Identifying aQTLs: genomic loci that modulate TF activity

The regression procedure of **Figure 3.1** takes into account prior information about the connectivity of the transcriptional network of the cell in a way that allows us to directly treat TF activity as a quantitative trait. To identify aQTLs for each TF, we used composite interval mapping (CIM) [97], which accounts for linkage between neighboring markers and has significantly better spatial resolution than single-marker methods (**Figure 3.4**). This is important, as even the aQTL regions detected using CIM typically encompass 20–30 genes, and our goal is to uncover *trans*-acting causal mutations in individual genes or even nucleotides. **Figure 3.5** provides an overview of the TF-locus associations identified using our method. To control for multiple testing, we use a log-odds (LOD) score threshold (red line in **Figure 3.1C**) corresponding to a 5% FDR (see Materials and methods). We identified a single aQTL for 55 and multiple aQTLs for 22 of the 123 TFs analyzed. Together, the mapped aQTLs cover several dozen distinct genomic loci. Note that all aQTLs are by definition *trans*-acting from the point of view of the mRNA expression level of individual genes, as the trait analyzed is the ‘hidden' regulatory activity of each TF.

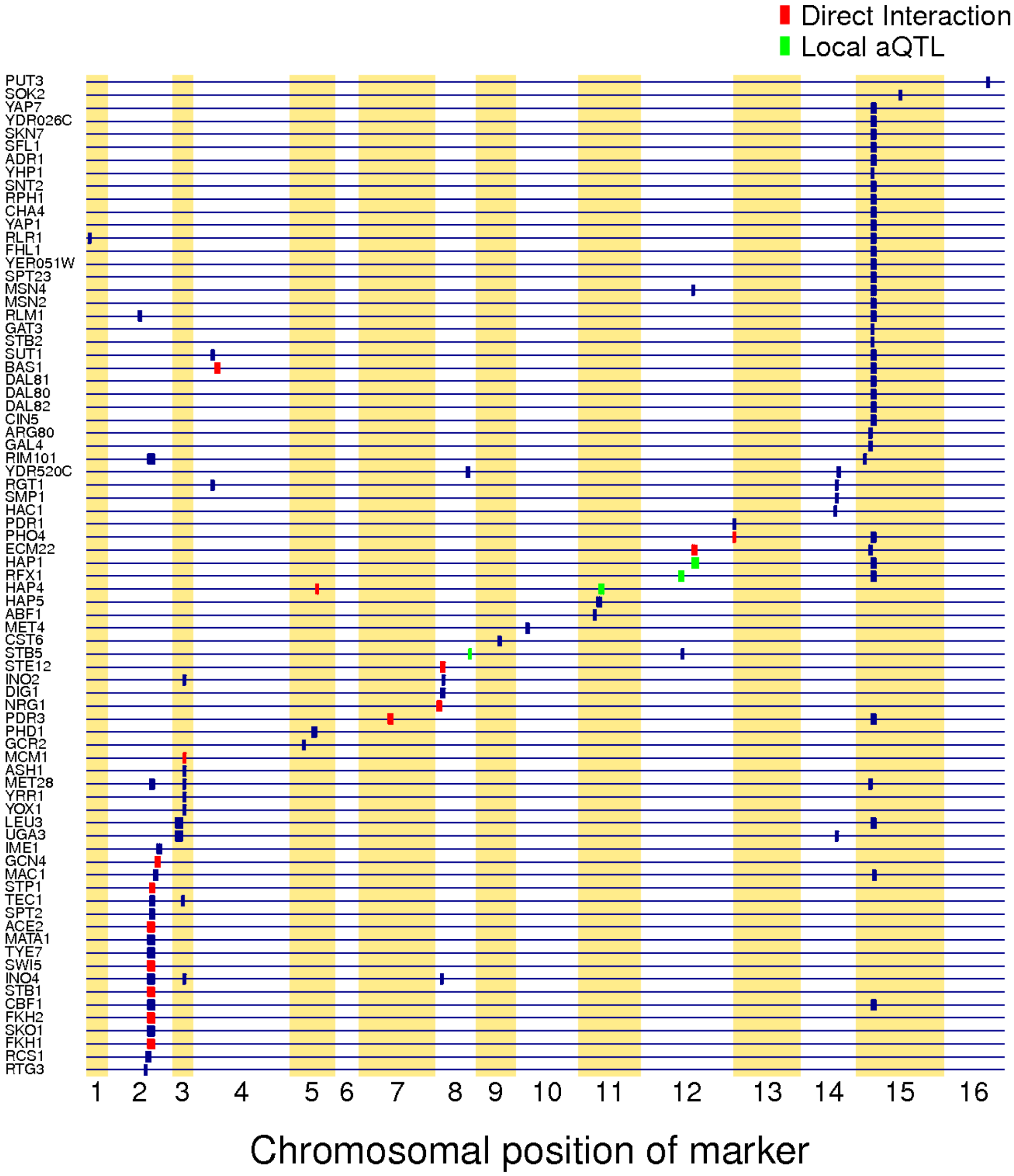

Figure 3.5: Overview of the trans-acting genetic modulators of TF activity mapped using our method. All transcription factors that have at least one significant aQTL region at a 5% FDR are shown. Transcription factors are sorted according to the chromosomal position of their maximum LOD score. Putative causal gene assignments are indicated in green (local aQTL: TF encoded by gene in aQTL) or red (protein-protein interaction identified between TF and gene in aQTL).

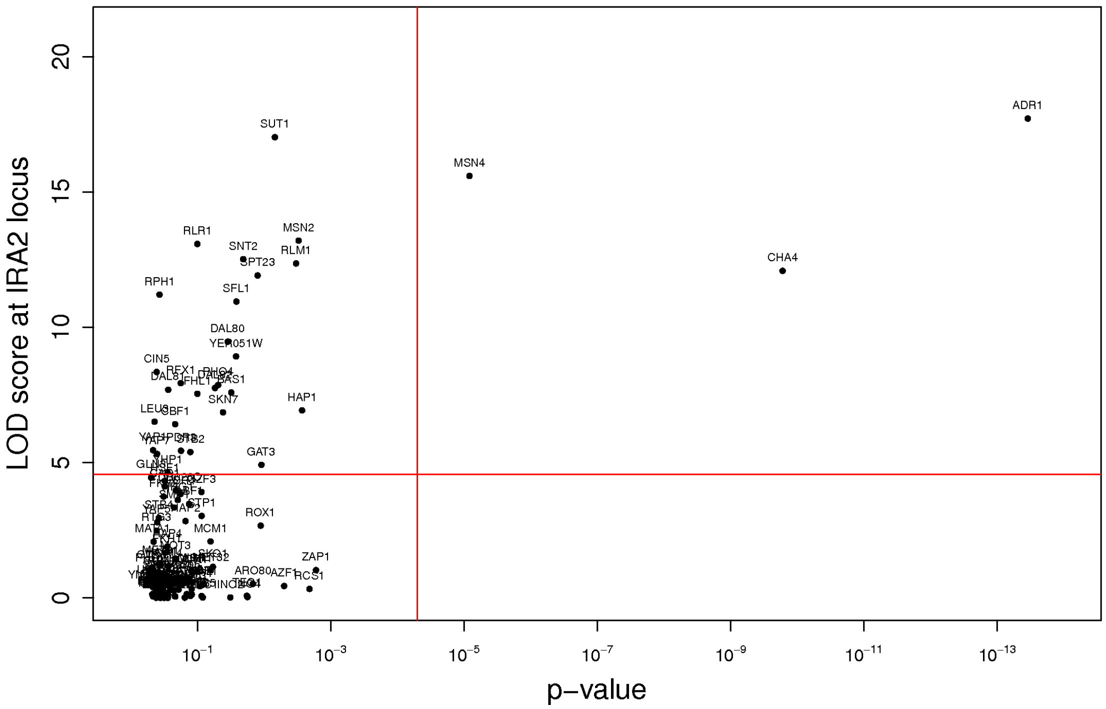

Figure 3.6: The effect of IRA2 allele replacement on TF activities. Comparing the average effect of *IRA2* allele replacement in a BY or RM background on the activities of TFs (x-axis) with the LOD score for linkage to the *IRA2* locus obtained using our aQTL method (y-axis). The red lines denote the LOD score and p-value thresholds corresponding to a 5% FDR

3.3.4 Validation of aQTL-TF linkage using a IRA2 allele swap

To the extent that aQTLs act independently, the regulatory consequences of allelic variation at a particular locus should be independent of the genetic background in which it occurs. To validate our method, we therefore analyzed gene expression profiles of allele replacement strains from a previous study [91]. According to our analysis, chromosome 15 contains an aQTL that influences the activity of several dozen distinct TFs (**Figure 3.5**). Among the 19 genes in this region is *IRA2*, which encodes a GTPase-activating protein that negatively regulates Ras proteins and thereby controls intracellular cAMP levels [98]. The coding region of IRA2 is highly polymorphic [91]. We analyzed the gene expression profile of a BY strain carrying the RM allele of the IRA2 coding region, and vice versa, and found that the activity of Adr1p, Cha4p, and Msn4p was significantly affected by the allele replacement (**Figure 3.6**; *P*-value 3.3×10^{-16}, 1.1×10^{-10}, and 1.6×10^{-5}, respectively, see Materials and methods). Each of these TFs was indeed predicted by our method to link to the IRA2 locus. Consistently, cAMP-dependent protein kinase is known to influence Adr1p activity [99] and regulate subcellular localization of Msn4p, which influences its activity [100]. Altogether, there are 30 TFs with an aQTL region containing the IRA2 gene. They do not all need to be influenced by the polymorphism(s) in its coding region; additional causal polymorphisms in nearby genes, modulating other subsets of the 30 TFs, may well exist. It is therefore not surprising that the activity of only 3 out of 30 TFs was significantly affected by the *IRA2* allele replacement. On the other hand, we do not expect any TF whose activity does*not* link to the IRA2 locus to be affected by the allele replacement. Indeed, as can be seen from **Figure 3.6**, our method achieved 100% specificity in this regard: none of the 93 TFs whose aQTL(s) do not contain IRA2 showed a change in regulatory activity.

3.3.5 Novel trans-acting polymorphisms in transcription factor genes

Of the aQTL linkages we detected, only four—those of Hap1p, Stb5p, Rfx1p, and Hap4p—are local (**Figure 3.5**, green boxes). The probability that a locus showing aQTL linkage encompasses the gene encoding the TF itself by chance is typically <1% (it equals the ratio of the number of genes in the aQTL and the total number of genes).

Therefore, whenever such local linkage happens, it is highly likely that the causal polymorphism resides in the coding region or regulatory region of the TF gene. The aQTL profile for Hap1p is shown in **Figure 3.1C,** and the polymorphism in *HAP1* that gives rise to it was already discussed above.

Stb5p is a C2H2 zinc finger protein that serves as an activator of multidrug resistance genes [101]. A significant difference in Stb5p activity exists between the BY and RM strains (**Figure 3.7A**), and this activity is highly heritable (h^2=95%). We detected highly significant local linkage (LOD score=10.84; Q-value=2.69×10^{-8}) between Stb5p activity and the allelic status of the *STB5* locus (**Figure 3.7B**). Alignment of the BY and RM protein sequences for Stb5p revealed five amino-acid mutations, all of which occur outside the DNA-binding domain. We found no nucleotide differences in the 5′ and 3′ untranslated regions or <1 kb upstream of the transcription start site of *STB5*. Consistently, the mRNA expression level of the *STB5* gene is not significantly correlated with the activity of Stb5p (r=0.18; P-value>0.05). Furthermore, CIM analysis of the mRNA expression level of the *STB5* gene did not reveal any local eQTL linkage (**Figure 3.7C**). The power of our aQTL approach is further underscored by the fact that no eQTL hotspot has been detected at the *STB5* locus [19]. It will be interesting to further dissect the post-translational mechanism(s) by which the sequence differences between the BY and RM alleles of Stb5p cause a difference in its regulatory activity.

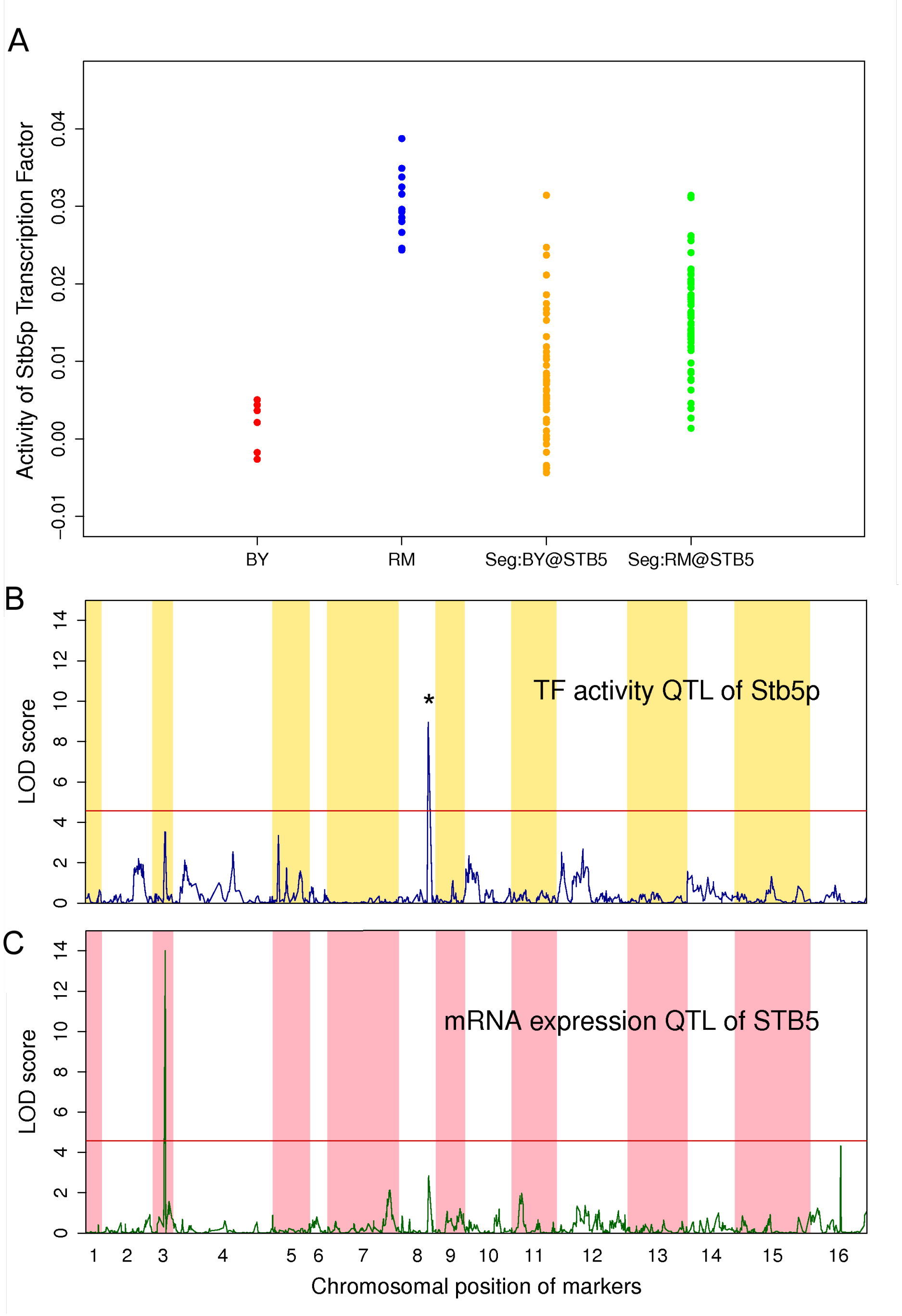

Figure 3.7: local-aQTL for STB5. (**A**) Inferred activity of Stb5p in parental strains and segregants. The first and second columns show the activity of Stb5 in 6 replicates of a BY-reference comparison and 6 replicates of a RM-reference comparison. The third and fourth columns show the activity of Stb5p for segregants that inherited the BY and RM at the STB5 locus, respectively. (**B**) LOD score profile for the activity of Stb5p. *Locus including STB5 gene. (**C**) LOD score profiles for the expression level of *STB5* gene.

Rfx1p is a major transcriptional repressor of the DNA damage response. The RM allele of the *RFX1* gene contains a premature stop codon. Consistently, genes whose promoter is predicted to be bound by Rfx1p tend to be more highly expressed in the BY strain than in the RM strain (**Figure 3.2**).

The last local aQTL we discovered was for Hap4p, a subunit of the heme-activated, glucose-repressed Hap2p/3p/4p/5p CCAAT binding complex. Consistently, the mRNA expression level of the *HAP4* gene is highly correlated with the activity of Hap4p (*r*=0.79).

3.3.6 CDC28 antagonistically modulates Fkh1 and fkh2

Chromosome II contains an ‘aQTL hotspot' whose allelic status influences the activity of no fewer than 15 distinct TFs (**Figure 3.5**), including Fkh1p and Fkh2p. The locus contains the *CDC28* gene, which encodes a cyclin-dependent kinase. Phosphorylation by Cdc28p is known to regulate the activity of Fkh2 by promoting interaction with a coactivator [102]. On the basis of the aQTL mapping to the *CDC28* locus in combination with high-throughput evidence of their physical interaction [103] with Cdc28p, we predict that Fkh1p is also post-translationally modulated by Cdc28p. The sign of the aQTL linkage to the *CDC28* locus for Fhk2p is the opposite of that for Fkh1p (**Figure 3.8A**): whereas the transcriptional targets of Fkp1p are more highly expressed in segregants carrying the BY allele at the *CDC28* locus, the opposite is true for the targets of Fkh2p (**Figure 3.8B**). The same pattern holds for the inferred difference in TF activity between the two parental strains (**Figure 3.2**). The antagonism between Fkh1p and Fkh2p is consistent with previously observed differences in function between the two factors

[104, 105]. These two TFs have similar sequence specificity, and consequently their total promoter affinity profiles are correlated across genes (*r*=0.72). Nevertheless, we were able to detect the opposite influence of the *CDC28* polymorphism on their activity because our method uses multivariate regression, which forces TFs with correlated promoter affinity profiles to compete for the same differential mRNA expression signal. When we analyze each TF separately using a univariate model, the CIM regression coefficients for Fkh1p and Fkh2p (incorrectly) have the same sign. This example underscores the importance of our affinity-based quantification of the matrix of regulatory connectivities between TFs and their target genes.

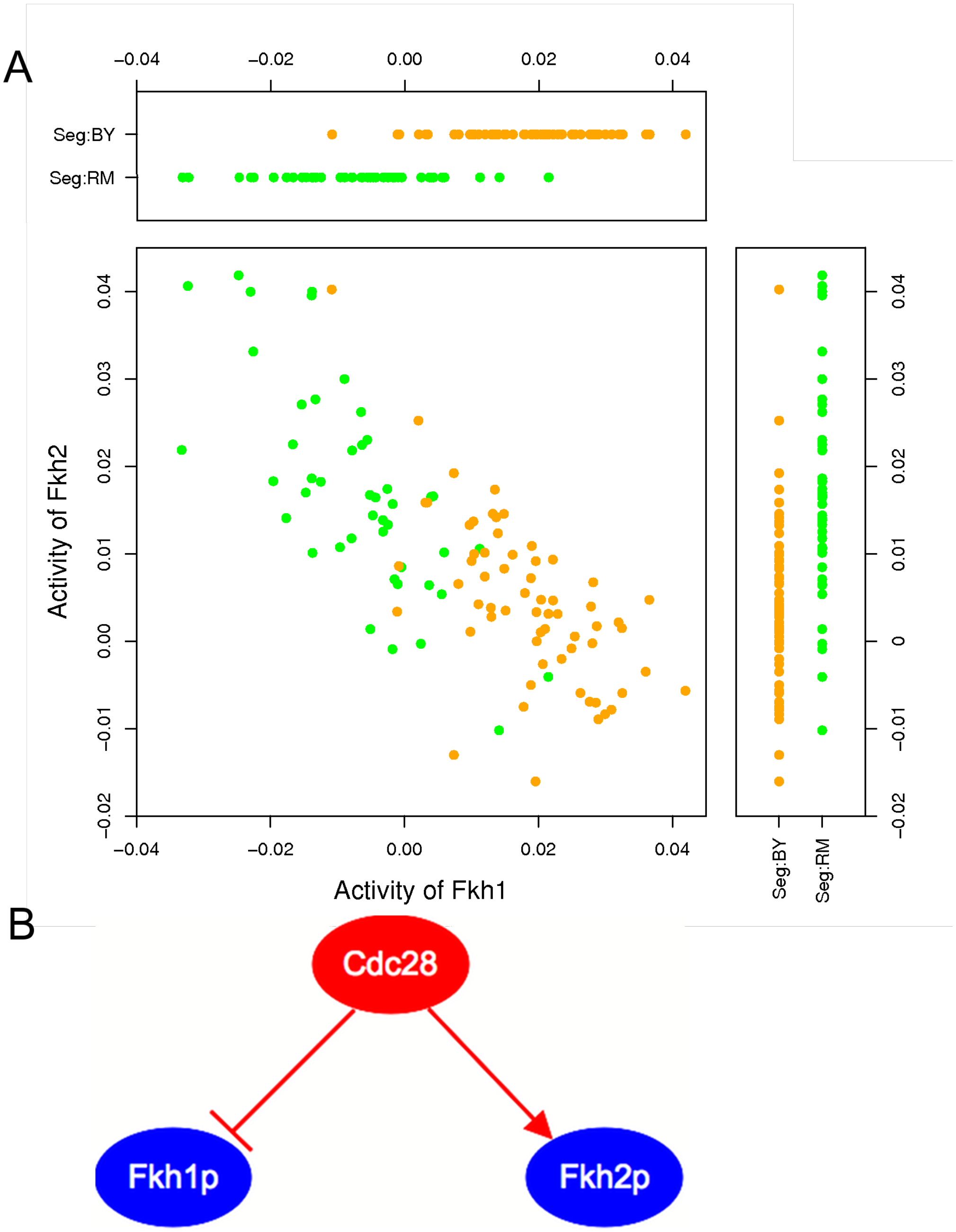

Figure 3.8: CDC28 as a modulator of Fkh1p and Fkh2p. (**A**) Activity of Fkh1p and Fkh2p across all segregants. The activity of Fkh1p is negatively correlated with that of Fkh2p. The yellow dots correspond to segregants carrying the BY allele at the *CDC28* locus, the green dots to those carrying the RM allele. (**B**) Schematic diagram illustrating the antagonistic modulation of Fkh1p and Fkh2p by Cdc28p. While the transcriptional targets of Fkh1p are more highly expressed in segregants carrying the BY allele at the *CDC28* locus, the opposite is true for the targets of Fkh1p.

3.3.7 An aQTL on chromosome VII controlling Zscan4 activity in mouse liver cells

To determine whether our method could map aQTLs for mammalian TFs, we applied it to parallel genotyping and liver cell expression data for an F2 mouse population [21]. Weight matrices derived from protein-binding microarray (PBM) data for 104 mouse TFs were used [106]. The model we used to analyze the yeast segregants contains '*cis*' coefficients, which explicitly model changes in expression because of allelic variation in promoter sequence, in addition to the '*trans*' coefficient that model the changes in TF activity. However, we found that a simpler '*trans*-only' model performed equally well in terms of mapping aQTLs when applied to the yeast segregant data. This gave us confidence to use a '*trans*-only' model in mouse, where the density of markers is too low to assign gene-specific promoter sequences. We identified an aQTL for Zscan4, a TF containing four zinc finger domains and a SCAN domain, which is also known as the leucine-rich region [107] (**Figure 3.9**). Using a multivariate linear model to analyze the homozygous C57BL/6J (BB), homozygous DBA/2J (DD), and heterozygous (BD) genotype at the aQTL locus (**Figure 3.9A**), we found the behavior of the aQTL to be additive and show no significant dominant effect (see Materials and methods). A highly significant linkage (LOD score=10.8) with Zscan4 activity occurs between 43 and 66 cM on mouse chromosome 7 (**Figure 3.9B**). This region contains >500 genes, which makes it difficult to predict the causal polymorphism. Limited information is available about protein–protein interaction (PPI) for mouse, and we could not detect any direct interaction between genes within this region and Zscan4p. However, our result demonstrates that TF activity can also be inferred and mapped in mammalian cells using our method, and provides a starting point for further dissection of *trans*-acting regulatory variation mediated by Zscan4p.

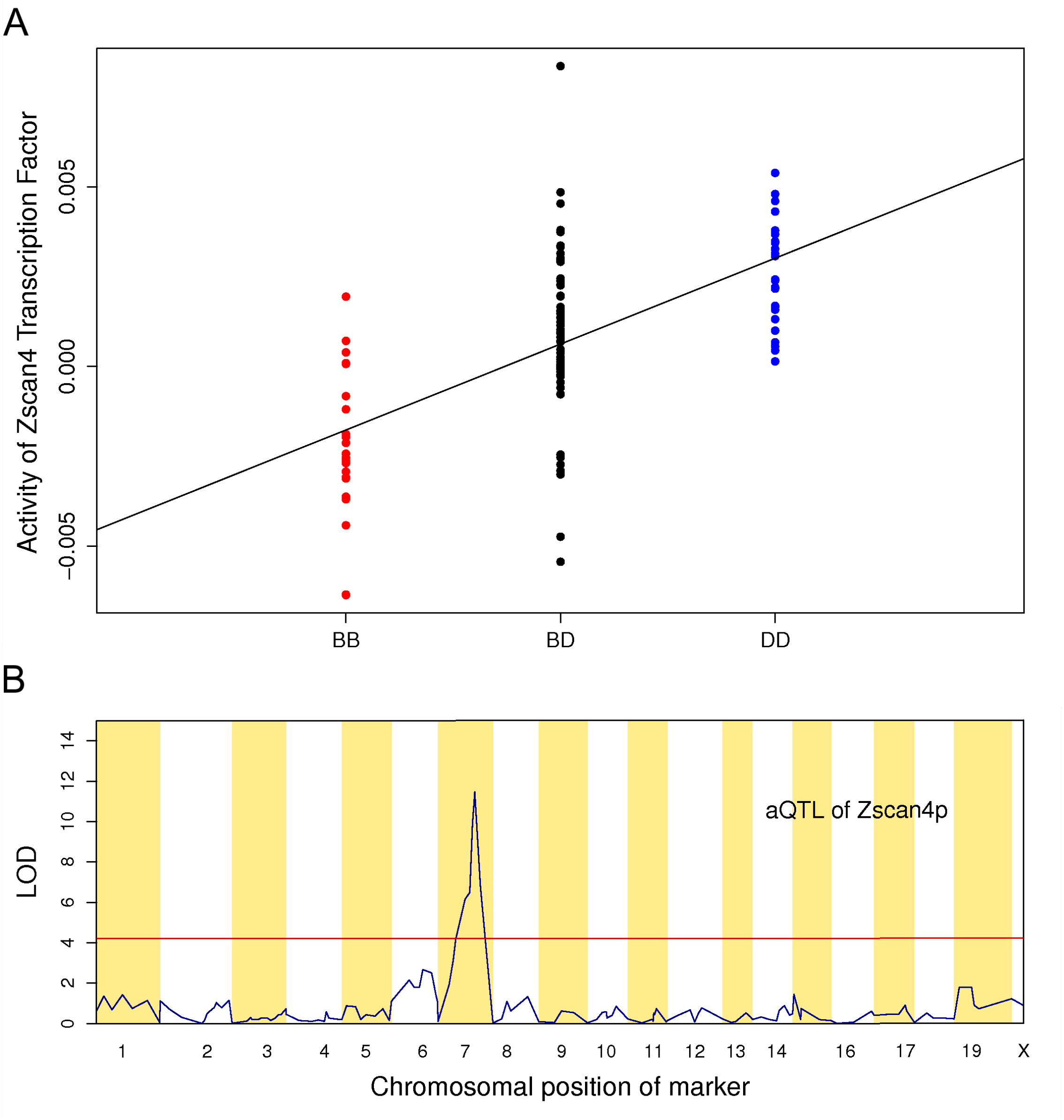

Figure 3.9: aQTL of Zscan4p. (**A**) Inferred activity of Zscan4p across all F2 mouse population. Each column show the activity of Zscan4 in homozygous C57BL/6J (BB), heterozygous (BD), and homozygous DBA/2J (DD) mice at aQTL positions, respectively. (**B**) LOD score profile for Zscan4p.

3.4 Discussion

We have presented a transcription-factor-centric method for identifying *trans*-acting genetic modulators of gene expression using parallel genotyping and mRNA expression phenotyping data. Our approach is based on the idea of treating the genotype-specific regulatory activity of each TF as a quantitative trait. It exploits prior information about the network of interactions between TFs and their target genes to infer genotype-specific TF activities from genome-wide measurements of mRNA expression. Our method has greatly increased statistical power to detect locus-TF associations. It is sensitive even to a relatively subtle influence of genotype-specific TF activity on mRNA expression because it is based on a statistical analysis across both genes and segregants. The fact that TF activity is not a gene-specific phenotype allows us to make the rather crude assumption that the strength of the regulatory connectivity between TF and target gene is proportional to *in vitro* promoter affinity. In reality, many of the predicted binding sites in promoter regions are not functional, due to complex interactions with nucleosomes and other chromatin-associated factors. It is remarkable that our method works in spite of this complexity.

Application of our aQTL method to a data set for 108 haploid segregants from a cross between two yeast strains [91] demonstrated a dramatic increase in statistical power to uncover the regulatory mechanisms underlying genetic variation in gene expression levels. We identified a total of 103 locus-TF associations, a more than six-fold improvement over the 17 locus-TF associations identified by several existing methods [19, 32, 34, 35, 91]. The total number of distinct genomic loci identified as an aQTL for one or more TFs equals 31, which includes 11 of the 13 previously identified eQTL hotspots [91]. Thus, our method identifies 20 novel *trans*-acting polymorphisms: almost double the number of known such loci in yeast. For many of the eQTL hotspots, it also implicated several TFs not previously known to mediate the influence of these loci on genome-wide mRNA expression.

Our regression procedure fully accounts for post-translational regulation of TF activity at the protein level, as we do not use the mRNA expression level of either the gene encoding the TF or one its upstream modulators as a surrogate for regulatory activity. Indeed, the correlation between the protein-level regulatory activity of a TF and its expression at the mRNA level across a large number of experimental conditions in yeast was recently found to often be quite poor [86]. The present study confirms this observation: Only one third of TFs analyzed show a significant (<5% FDR) correlation between mRNA expression and activity. Moreover, only 12 of the 103 TF-locus associations could be confirmed when mRNA expression level was used as a proxy inferred protein-level TF activity.

We also applied our aQTL method to the earlier yeast segregant data set of [26]. This confirmed the dramatic increase in statistical power afforded by our approach. We detected a total of 79 locus-TF associations, which again is a more than six-fold improvement over the 14 locus-TF associations detected from these data by several existing methods [34, 35, 37, 38] combined. Furthermore, 28 of these 79 locus-TF associations were also detected using the data of [91]. This degree of reproducibility strongly validates our method: given that the number of possible such associations equals the number of TFs (123) times the number of markers (~3000) divided by the average number of genes per locus (~20), we would expect this overlap to be ~0.4 by random chance. There is also no reason to expect complete overlap, as the data sets were similar but not identical. Indeed, although 13 eQTL hotspots have been identified in each respective data set, only 8 of these are the same [34, 91].

Our findings are consistent with previous observations [32] that most *trans*-acting variation in yeast does not map to TF genes, but to upstream modulators of TF activity. Indeed, of the total of 103 TF-locus associations shown in **Figure 3.5** only four are local. We confirmed that *HAP1* is directly affected by a sequence polymorphism, and discovered novel *trans*-acting polymorphisms in the TF-encoding gene*STB5*, *RFX1*, and *HAP4*. Unexpectedly, our analysis revealed loci on chromosomes II and XV that are informative for a large number of TFs (‘aQTL hotspots'). We stress that this cannot be accounted for in terms of correlated profiles of promoter affinity across genes, as we found these to be largely independent between TFs. Rather, this phenomenon seems to point to one-to-many relationships between signal transduction pathways and TFs. For instance, our method predicts that genetic variation at the locus on chromosome II encoding the cyclin-dependent kinase CDC28 changes the activity of multiple cell cycle associated TFs (Ace2p, Fkh1p, Fkh2p, and Swi5p). At the same time, distinct polymorphisms at the same aQTL could be responsible for modulating different subsets of linked TFs. Evidence for this is our observation that allele replacement at the IRA2 locus on chromosome XV only affected a small subset of the TFs whose activity is linked to this aQTL (cf. **Figure 3.6**).

In an effort to uncover further specific molecular mechanisms underlying the aQTL linkages summarized in **Figure 3.5** we supplemented our genetic analysis with knowledge about physical and genetic PPIs; see Materials and methods for details. The information provided by PPI and aQTL is highly complementary. On the one hand, aQTL linkage can only implicate relatively large genomic regions, not individual genes, as genetic modulators of TF activity. On the other hand, although PPI data can connect a TF to a putative modulator of its activity, it would be questionable to conclude that the interaction corresponds to a functional regulatory network connection without the strict causality and directionality associated with aQTL linkage. In all cases, the probability that a gene within the aQTL region encodes one of the direct interactors of the TF by chance is <3% (see Materials and methods). Therefore, most of these genes (aQTG) are expected to encode direct or indirect modulators of the TF's activity. We were able to implicate a non-coding polymorphism in the *CDC28* gene as a plausible genetic factor underlying the major eQTL hotspot on chromosome II (in addition to the experimentally validated*trans*-acting polymorphism in the *AMN1* gene in the same region [32]) and make a strong prediction that the functionally distinct cell cycle regulators Fkh1p and Fkh2p are modulated by the cyclin-dependent kinase Cdc28p in an antagonistic manner.

Extensive transgressive segregation has been previously identified for the expression levels of individual genes [26]. However, when we tested for the same phenomenon at the level of TF activity (see Materials and methods), we were only able to detect transgressive segregation for Ecm22p and Tec1p; in both cases, the effects of two aQTLs for same TF cancel each other in both parental strains, and no differential activity between RM and BY could be observed (**Figure 3.2**). Presumably, much of the transgressive segregation at the level of individual genes is due to the fact that positive and negative contributions from different TFs can cancel each other. Our multivariate modeling of each individual gene's expression level in terms of the activity of multiple TFs accounts for such compensation explicitly, and hence the transgression is much less prevalent for aQTLs than for eQTLs.

In our approach, ‘phenotype space' is reduced from that of all genes to that of all TFs. Rather than mapping the measured mRNA expression level of individual genes to eQTLs, we map the inferred activity of each TF to ‘aQTLs.' This enhances statistical power in two distinct ways. First, it improves the signal-to-noise ratio for the quantitative trait itself, as the activity of each TF is estimated from the mRNA expression levels of its many targets. Second, the severity of the multiple-testing problem associated with QTL mapping because of the large number of marker/trait combinations is greatly reduced. Running in only seconds on a single processor, our algorithm is also computationally efficient.

It is important to emphasize that in our method the molecular identity of a TF is only defined through the PSAM that parameterizes its DNA-binding specificity. The sequence-to-affinity model for each TF needs to be specific enough to allow differentiation from all other TFs. We found that in the case of the budding yeast *S. cerevisiae* this condition generally holds. Given the rapid pace at which *in vitro* DNA-binding data is currently being generated for mammalian TFs [106], together with the demonstrated ability of regression-based models to infer TF activity in human cells [108], we expect application of our method also to be feasible in higher eukaryotes.

Taken together, our results underscore the value of explicitly treating TF activity as a quantitative trait from a systems biology perspective as a promising strategy for increasing the statistical power of genome-wide linkage and association studies. More generally, our method is applicable whenever a matrix of connection strengths between regulators and targets, independent of the phenotype matrix, is available as prior information. There are several directions in which this approach can be extended. First, the use of more sophisticated methods for causal gene identification [37, 109, 110] is likely to uncover additional molecular mechanisms. It will also be interesting to analyze to what extent the connectivity between the TF and their genetic modulators depends on the nutrient condition in which the yeast cells are grown [91]. Furthermore, aQTLs provide a novel vantage point for analyzing locus–locus interactions. Finally, it should be interesting to analyze to what extent genetic variation in steady-state gene expression levels because of post-transcriptional regulation of mRNA stability [94, 110] is amenable to dissection using the method introduced in this paper.

3.5 Materials and methods

3.5.1 Gene expression and genotyping data

We analyzed genome-wide mRNA expression data from a study performed by [91], which used two-color cDNA arrays. The data (GEO accession number GSE9376) cover a genetic cross between two haploid yeast strains—a laboratory strain (BY4716) and a natural isolate (RM11-1a). The data set includes six biological replicates of the BY parental strain, six replicates of the RM parental strain, and one replicate for each of 108 haploid segregants grown in two different conditions, with glucose and ethanol as the carbon source, respectively. For the present study, we only used data for the glucose condition. The study used a reference design in which all hybridizations were performed using equal amounts of mRNA from both parents (BY and RM) grown in both conditions as a reference. Log_2-ratios, averaged over a dye swap, were used for all further analysis.

For comparison, we also analyzed genome-wide mRNA expression data for yeast segregants from a cross between BY and RM strains (GEO accession number GSE1990) from an earlier study performed by [26]. Following these authors, we excluded ORFs rejected by [111]. The data set covers 6 biological replicates of the BY parental strain, 12 replicates of the RM parental strain, and 1 replicate for each of 112 haploid segregants. The study used the BY material as a reference. Log_2-ratios, averaged over the dye swap, were used for all further analysis. In addition, we averaged log-ratios for 13 ORFs that were spotted twice. Finally, we normalized each array by subtracting the mean log-ratio. For each of the segregants whose expression levels were determined, 2957 markers were genotyped by [26], who kindly made this data available to us.

We also analyzed previously published F2 mouse genome-wide expression data [21, 92](GEO accession GSE2008). The data set contains genome-wide oligonucleotide microarrays profiled using liver tissue from 111 F2 mice, which were constructed from two standard inbred strains, C57BL/6J and DBA/2J. The F2 mice fed an atherogenic diet for 4 months beginning at 12 months of age. This study used a common pool created from equal portions of RNA from each of the samples as a reference. Following the previous study, expression changes between each sample and a reference were quantified as expression $\log_{10}$-ratios between normalized, background-corrected intensity values for the two channels. The F2 intercross mice were genotyped at 139 microsatellite markers uniformly distributed over the mouse genome.

3.5.2 Genome sequence of BY and RM strains

We obtained RM11-1a sequence data from the Broad Institute (http://www.broad.mit.edu) and BY4716 sequence data from the *Saccharomyces Genome Database* (SGD; http://www.yeastgenome.org).

3.5.3 Defining genotype-specific promoter sequences

To define genotype-specific promoter sequences, we first identified pairs of genes orthologous between BY and RM. We aligned coding sequences of RM genes to the BY strains using BLAST in Bioperl [112], and chose the best BLAST hits to identify the orthologous genes. Then, we obtained 600 bp upstream sequences of each orthologous pair to define BY and RM-specific promoter sequence. For segregants, we determined whether the promoter sequence of a particular gene was inherited from BY or RM strains. To this end, we first identified all genetic markers located within the 600 bp upstream of each open reading frame. If no genetic marker within 600 bp could be found, we selected the marker closest to the upstream region. The genotype of the selected markers was used to assign either the BY or RM promoter sequence to the gene. If multiple markers with inconsistent genotypes were selected, we discarded the gene.

3.5.4 Inferring segregant-specific TF activities

We downloaded a collection of 124 position weight matrices (PWMs) from a study by MacIsaac *et al*. [95] (we excluded Hap3, as it has the exact same PWM as Hap5). Next, we used the `convert2psam` utility from the MatrixREDUCE v2.0 software package (see `http://bussemakerlab.org`) to convert each PWM to a position-specific affinity matrix (PSAM) [78, 113, 114]. Pseudo-counts equal to one were added to the PWM at each position, and the resulting base counts were divided by that of the most frequent base at each position to get an estimate for the relative affinity associated with each point mutation away from the optimal binding sequence [115]. The resulting PSAM collection was used to infer genotype-specific changes in TF activity.

The occupancy $N_{\phi g}$ of the upstream region U_g of gene g by transcription factor ϕ depends on the nuclear concentration $[\phi]$ of the TF and on the landscape of binding affinity across U_g. Both these quantities are genotype-specific. At non-saturating concentrations of the TF, the occupancy in genotype G can be approximated by the product of concentration and affinity [78]:

$$N_{g\phi}(G) \approx [\phi](G) K_{\phi g}(G)$$

The total promoter affinity $K_{\phi g}(G)$ depends on the segregant-specific upstream sequence $U_g(G)$, and is given by:

$$K_{\phi g} = \sum_{i \in U_g} K_{g\phi i} = \sum_{i \in U_g} \prod_{j=1}^{L_\phi} w_{\phi j b_{i+j-1}(U_g)}$$

Here, $K_{g\phi i}$ represents the binding affinity (relative to the optimal DNA sequence) between transcription factor ϕ and the DNA in a window of length L_ϕ starting at position i within U_g. Assuming independence between nucleotide positions, we approximate $K_{g\phi i}$ by a product of position-specific relative affinities $w_{\phi jb}$. Finally, $b_i(U_g)$ denotes the base identity at nucleotide position i within U_g.

We assume that when steady-state mRNA abundances are being compared between genotype G and reference genotype G_{ref}, the expression log$_2$-ratio for gene g, to linear approximation, is proportional to the difference in promoter occupancy:

$$\begin{aligned}
\log_2\left([mRNA_g](G)\right) - \log_2\left([mRNA_g](G_{\text{ref}})\right) &\propto N_{\phi g}(G) - N_{\phi g}(G_{\text{ref}}) \\
&\approx [\phi](G)K_{\phi g}(G) - [\phi](G_{\text{ref}})K_{\phi g}(G_{\text{ref}}) \\
&= \left([\phi](G) - [\phi](G_{\text{ref}})\right)K_{\phi g}(G) + [\phi](G_{\text{ref}})\left(K_{\phi g}(G) - K_{\phi g}(G_{\text{ref}})\right)
\end{aligned}$$

All total promoter affinities are known, so we can use the differential mRNA abundances to estimate coefficients $\beta^{\text{cis}} \equiv [\phi](G_{\text{ref}})$ and $\beta^{\text{trans}} \equiv [\phi](G) - [\phi](G_{\text{ref}})$. This motivated us to fit the following multivariate linear model to each segregant:

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

Here y_{gs} represents mRNA expression log-ratios for gene g in segregant s. For the segregant data of Smith *et al.* [116], whose used a pool of equals amounts of parental strains as their reference sample, $\left\langle K_{\phi g}\right\rangle_{\text{ref}}$ equals the average of BY and RM promoter affinities, while for that of Brem *et al.* [117], who used the BY strain as their reference,

$\left\langle K_{\phi g} \right\rangle_{\text{ref}}$ equals the BY promoter affinity. The intercept β_{0s} absorbs any normalization differences that may occur. The genomewide affinity profiles for several PSAMs are highly correlated (e.g., Msn2 and Msn4, Ino2 and Ino4). To avoid any problems resulting from such multicollinearity, we used ridge regression, which minimizes the residual sum of squares subject to a penalty proportional to the L_2-norm of the coefficients, and gives a slightly biased but more precise estimator of coefficients than ordinary least squares [118]. We also fit the above model in "trans-only" mode ($\beta^{\text{cis}} \equiv 0$).

To infer segregant-specific TF activities in mouse, we downloaded PWMs defined by Badis *et al.* [106] who used protein binding microarray (PBM) technology to determine the *in vitro* DNA binding specificities of 104 different mouse TFs. We estimated PSAM and total promoter affinity from PWMs using 1000bp upstream sequence of C57BL/6J strain by the same procedure explained above. We obtained C57BL/6J mouse genome sequence from UCSC Genome Browser (`http://genome.ucsc.edu/`).

3.5.5 Heritability

We calculated the heritability of the activity of each TF as follows:

$$h^2 = (\sigma_s^2 - \sigma_p^2) / \sigma_s^2$$

Here σ_s^2 and σ_p^2 are the variance of the linear regression coefficient from the ridge regression across the segregants, and the pooled variance of the parental strains, respectively. To determine the statistical significance of the heritability, we performed ridge regression after independent random permutation of expression log-ratios (parents and segregants combined) for each gene (1,000 samples) and used the resulting empirical null distribution to compute a false discovery rate (FDR).

3.5.6 aQTL mapping in yeast

To detect significant genetic contributions to TF activity by specific loci, we performed a split of the segregants by each specific marker and tested for a difference between the two distributions of ridge regression coefficients using Welch's *t*-test and the non-parametric Wilcoxon–Mann–Whitney test. We also used CIM, which uses multivariate regression on multiple markers for increased precision of QTL mapping [97], as implemented in the R/qtl package [119]. Statistical significance was determined by performing independent random permutation of expression log-ratios (segregants only) for each gene. The FDR corresponding to a given LOD score threshold was computed as the ratio of the number of linkages above threshold averaged over 20 randomized data sets, and the number of transcripts with detected linkage. We also estimated the FDR using the standard Benjamini-Hochberg procedure [120]. For the CIM method, a 5% FDR based on the empirical permutation test corresponded to a LOD score >4.49.

3.5.7 aQTL mapping in mouse

In the case of mouse data analysis, aQTL mapping was conducted using a linear model. First, we constructed explanatory variables for the additive and dominance terms for each marker from the estimated genotype probabilities and used them in the regression analysis. Linkages were identified by comparing the likelihood, maximized as a function of the regression coefficients, for the following multivariate linear model

$$\beta_{\phi s}^{\mathrm{trans}} = \beta_0 + \beta_{\phi}^{\mathrm{add}} X_{ms}^{\mathrm{add}} + \beta_{\phi}^{\mathrm{dom}} X_{ms}^{\mathrm{dom}}$$

to the likelihood for the null model $\beta_{\phi s}^{\mathrm{trans}} = \beta_0$. Here, the dependent variable $\beta_{\phi \mathrm{s}}^{\mathrm{trans}}$ represents the TF activity as estimated using the affinity-based model defined above in "trans-only" mode ($\beta^{\mathrm{cis}} \equiv 0$). The independent variables X_{ms}^{add} (taking values 0, 1, and 2, for (diploid) genotypes BB, BD, and DD, respectively) and $X_{\mathrm{ms}}^{\mathrm{dom}}$ (taking values 0, 1, and 1, for the same respective genotypes) represent additive and dominant terms for each marker respectively. The LOD score was defined as the $\log_{10}$ of the likelihood ratio between the two models. The FDR was computed using the same procedure described above; an FDR <5% based on empirical permutation test corresponded to a LOD score >4.21.

3.5.8 Protein-protein interaction data

To identify putative causal genes from the aQTL regions of each specific TF, we used three different types of protein-protein interaction data: (i) physical and genetic interactions in the BioGRID database [121], (ii) interactions between chromatin modifiers and associated TFs [122], and (iii) kinase-TF interactions [123]. We computed the expected number of direct interactors among the genes in the aQTL region for a specific TF based on the total number of interactors of the TF genomewide, the number of genes in the aQTL, and the total number of genes. Statistical significance was computed using Fisher's exact test.

3.5.9 Validation of predicted locus-TF association

We downloaded gene expression profiles obtained by Smith *et al*. [116] for a strain carrying the RM allele of *IRA2* in the BY4742 background (RM@IRA2), a strain carrying the BY allele of *IRA2* in the RM11-1a background (BY@IRA2), and six replicates each of the BY and RM parental strains (GEO accession number GSE9376). We only used the date for cells grown in glucose as the carbon source. The reference sample used in all cases was pooled parental mRNA (see above). Therefore, to obtain an estimate for the differential expression between RM@IRA2 and BY, we subtracted the mean log-ratio of the BY replicates from the RM@IRA2 log-ratios,

$$y_g^{\mathrm{BY}\rightarrow\mathrm{RM@IRA2}} = \left[\log_2\left(\frac{[mRNA_g](RM@IRA2,\mathrm{glucose})}{[mRNA_g](\mathrm{pool})}\right) - \log_2\left(\frac{[mRNA_g](BY,\mathrm{glucose})}{[mRNA_g](\mathrm{pool})}\right)\right]$$

and performed multivariate (ridge) regression of these values on the BY promoter affinities for all TFs. We also performed the equivalent analysis where the roles of RM and BY were reversed. Finally, to average over strain background, we took the difference between the two regression coefficients for each TF to be our statistic for differential activity. To determine statistical significance, we performed 1,000 random permutations of all genes to determine the standard error of an empirical null distribution, and used it to compute a p-value. A FDR of 5% corresponded to a p-value of $10^{-4.20}$.

Chapter 4

Applications of the aQTL method

Our aQTL approach [90] is generally applicable whenever a matrix of connection strength between regulators and targets, independent of the phenotype matrix, is available as prior information. In collaboration with other Bussemaker lab members, I have extended this approach to other biological contexts: (i) the promiscuous binding of TFs to high-occupancy target (HOT) regions and (ii) post-transcriptional regulatory networks.

4.1 Transcriptional factors dynamically congregate at non-coding RNA gene in *Saccharomyces cerevisiae*

This section has been adapted from a manuscript co-authored by Lucas D. Ward, Junbai Wang, Eunjee Lee, and Harmen J. Bussemaker, which is currently in revision. My contribution to this work was to perform QTL mapping of the amplitude of the hotspot phenomenon for Ste12p for yeast segregants data (Figure 4.5).

4.1.1 Introduction

A canonical description of transcriptional regulation involves the binding of transcription factors (TFs) to *cis*-regulatory promoter and enhancer elements, resulting in recruitment of the RNA polymerase complex. However, whole-genome studies of TF binding have revealed that affinity of DNA for a TF is not sufficient for occupancy by TFs [124-126]. A number of mechanisms constrain the binding of TFs to only a subset of the sites that genomic affinity alone would predict, such as chromatin state [127] and variability in local TF concentration [128].

Perhaps more surprising has been the recent discovery of genomic loci at which many TFs congregate despite the underlying DNA sequence having affinity for only a small subset of them. The evidence of such TF colocalization hotspots, or High-Occupancy Target (HOT) regions, was seen in *Drosophila melanogaster* TFs [129]., *C. elegans* genome [130], and mouse embryonic stem cells [131]. A number of previously-proposed mechanisms are consistent with TF colocalization, including multifunctional "transcription factories" or enhanceosomes [132] cross-linked by chromatin loops, a locally permissive chromatin structure [133, 134]. A feature shared by both organisms is that hotspots are associated with increased expression at neighboring genes, but are often located far from traditionally-defined proximal promoters.

The present study was motivated by the fact that, although much more extensive genome-wide protein location data has been collected in yeast than in higher eukaryotes [135-137] no analogous colocalization of sequence-specific regulators has been observed. Significantly, however, in the large-scale compendia by Lee et al. and Harbison et al., the authors normalized across arrays for each probe to account for biases in the immunoprecipitation reaction. Clearly, this normalization would have effectively removed the evidence for any true genomic hotspot pattern shared by many TFs. We therefore performed a detailed re-analysis of the original microarray data in which we have omitted the probe-specific normalization step. A pattern of ubiquitous TF binding at many probed regions was immediately apparent. Remarkably, unlike in fly and mouse, these yeast hotspots are not associated with sequence-predicted affinity for any of the TFs involved. Rather, sequence and functional analysis reveals that the most significant features of co-occupied probed regions are: (i) the extent of nucleosome depletion and (ii) the proximity to noncoding RNA genes, the majority of which encode tRNAs and snoRNAs. Additionally, the TF colocalization hotspots are occupied chiefly in rich-media (YPD) conditions. The phenomenon is abrogated in the majority of environmental perturbation and stress conditions, indicating that the hotspots are a regulated biological entity. Supporting this hypothesis, we show that genetic variation at the locus encoding the regulatory gene RIM15 is associated with the amplitude of the hotspot effect.

4.1.2 Results

4.1.2.1 A majority of TFs preferentially immnuprecipitate with ncRNA genes, not at their target genes

Because we were interested in finding loci that were bound by many TFs, we first considered the ChIP fold-enrichment (FE) of the 195 TFs profiled in rich media (YPD) for each probed region, using the median $\log_2$ fold-enrichment (MFE) as a measure of co-occupancy. Surprisingly, the distribution of co-occupancy across probes was skewed heavily to the right (**Figure 4.1A**), suggesting that a subset of the probed regions was occupied by many factors. To systematically determine whether specific genomic features were associated with co-occupancy, we tested whether the distribution of MFE for probes corresponding to each annotated genomic feature was different from that corresponding to the rest of the genome. The most significantly co-occupied were the 514 probes corresponding to ncRNA genes (difference of means $\Delta\overline{M}$ = 0.27; $p = 6.9 \times 10^{-161}$, Student's t-test; $p < 2.2 \times 10^{-16}$, Wilcoxon-Mann-Whitney test). The more specific ncRNA categories of tRNAs, snoRNAs, and snRNAs were all significantly co-occupied as well.

MacIsaac et al. [95] combined ChIP data with motif analysis and comparative genomics to define which promoter regions correspond to direct target genes for each yeast TF. We were interested in the extent to which these targeted regions were occupied according to our reanalysis of the data, and how this occupancy compared in magnitude to the occupancy we discovered at ncRNA genes (see Methods). We found that occupancy at ncRNA genes was typically on the same order of magnitude as the occupancy at annotated targets (**Figure 4.1B**). Of all TFs, 47 significantly occupied both the ncRNA loci and their annotated targets, while 28 significantly occupied ncRNA genes but not their annotated targets, and only three significantly occupied their annotated targets but not ncRNA genes.

4.1.2.2 Co-occupied loci are nucleosome-depleted

To explore other relationships between genome annotation and TF co-occupancy, we looked for Gene Ontology (GO) enrichment. For every GO category, we compared the distribution of MFE within probes corresponding to promoters of genes in that category with the rest of the probes using both a t-test and a Wilcoxon-Mann-Whitney test. The most enriched GO category is genes encoding components of the cytosolic ribosome (t = 12.3, $p = 5.5 \times 10^{-20}$). Because ribosomal protein (RP) promoters are known to be particularly active [138], we were interested in whether nucleosome depletion at these promoters was making them more susceptible to nonspecific binding by TFs. Indeed, we found that MFE is strongly anticorrelated with nucleosome occupancy (Pearson r = –0.31, $p = 1.3 \times 10^{-128}$; **Figure 4.1C**).

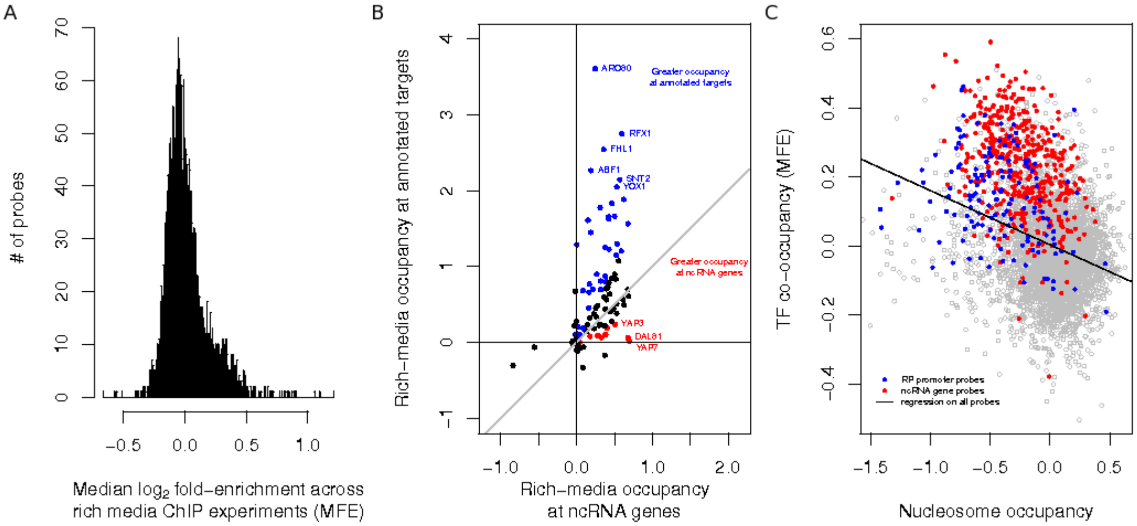

Figure 4.1: Detection of TF co-occupancy. (**A**) Distribution of TF co-occupancy at probes, defined as median $\log_2$ fold enrichment (MFE) across all analyzed rich media experiments from Harbison et al. [135]. (**B**) Occupancy of TFs at their annotated target regions and at ncRNA genes. The occupancy at both is expressed as $\Delta\overline{M}$, the difference between the mean log2 fold enrichment of the probes in question and the mean log2 fold enrichment of all other probes. Significant occupancy is defined as described in Methods. (**C**) TF co-occupancy vs. nucleosome occupancy. Plotted as a black line is a fit of all the data to a linear model y = ax + b, where a = -0.16 and b = 0.003 (r = -0.31). RP promoter probes are colored blue and ncRNA gene probes colored red. Note that both RP promoters and ncRNA genes are significantly co-occupied as well as nucleosome depleted. Residuals corresponding to ncRNA gene probes are higher than residuals corresponding to RP promoter probes (t = 11.8, p = 1.2 x 10-25).

4.1.2.3 Co-occupancy at ncRNA genes is largely eliminated in stress conditions

So far, our analysis has been restricted to rich media (YPD) conditions. It is known that the nuclear localization of many TFs is altered in stress conditions [139]. Examining ncRNA loci in stress conditions reveals dramatically reduced co-occupancy (**Figure 4.2**). Using the median TF occupancy across all non-YPD conditions, the elevation in co-occupancy at ncRNA genes drops from $\Delta\overline{M}$ = 0.25 to $\Delta\overline{M}$ = 0.03. To further investigate this general observation by focusing on occupancy of individual TFs in their rich media and stress conditions. For each particular stress-TF combination, we calculated the occupancy at ncRNA genes relative to all other probes (**Figure 4.3**). As expected from our pooled analysis, in the majority of stress conditions the occupancy at ncRNA genes is greatly reduced. The most notable example of this is Ste12p, which occupies ncRNA genes upon exposure to alpha mating factor, but not in the absence of alpha factor or in the presence of 1-butanol. Dig1p, which is also associated with the mating response, behaves differently: it does not occupy ncRNA genes in rich media, and avoids them in the presence of alpha mating factor and 1-butanol.

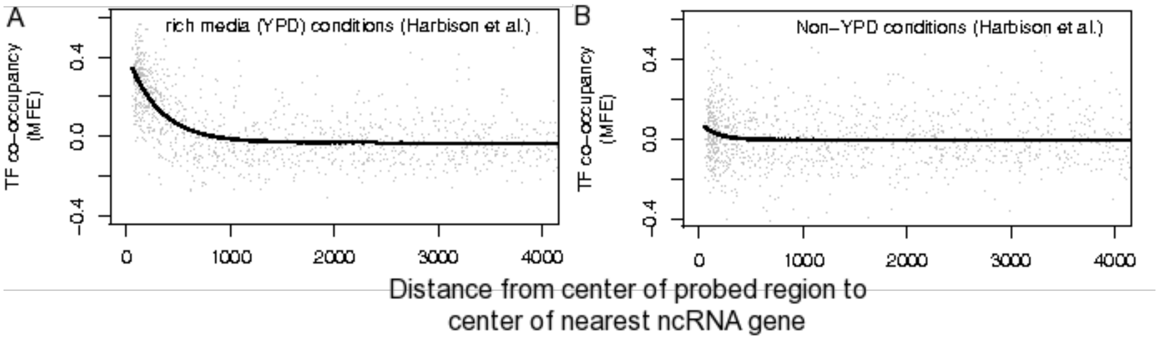

Figure 4.2: condition specific co-occupancy at ncRNA genes. TF co-occupancy, defined as the median $\log_2$ ChIP-chip fold enrichment (MFE), as a function of distance to the nearest ncRNA gene. (**A**) MFE across YPD experiments from Harbison et al. [135] (**B**) MFE across non-YPD experiments from Harbison et al. [135]

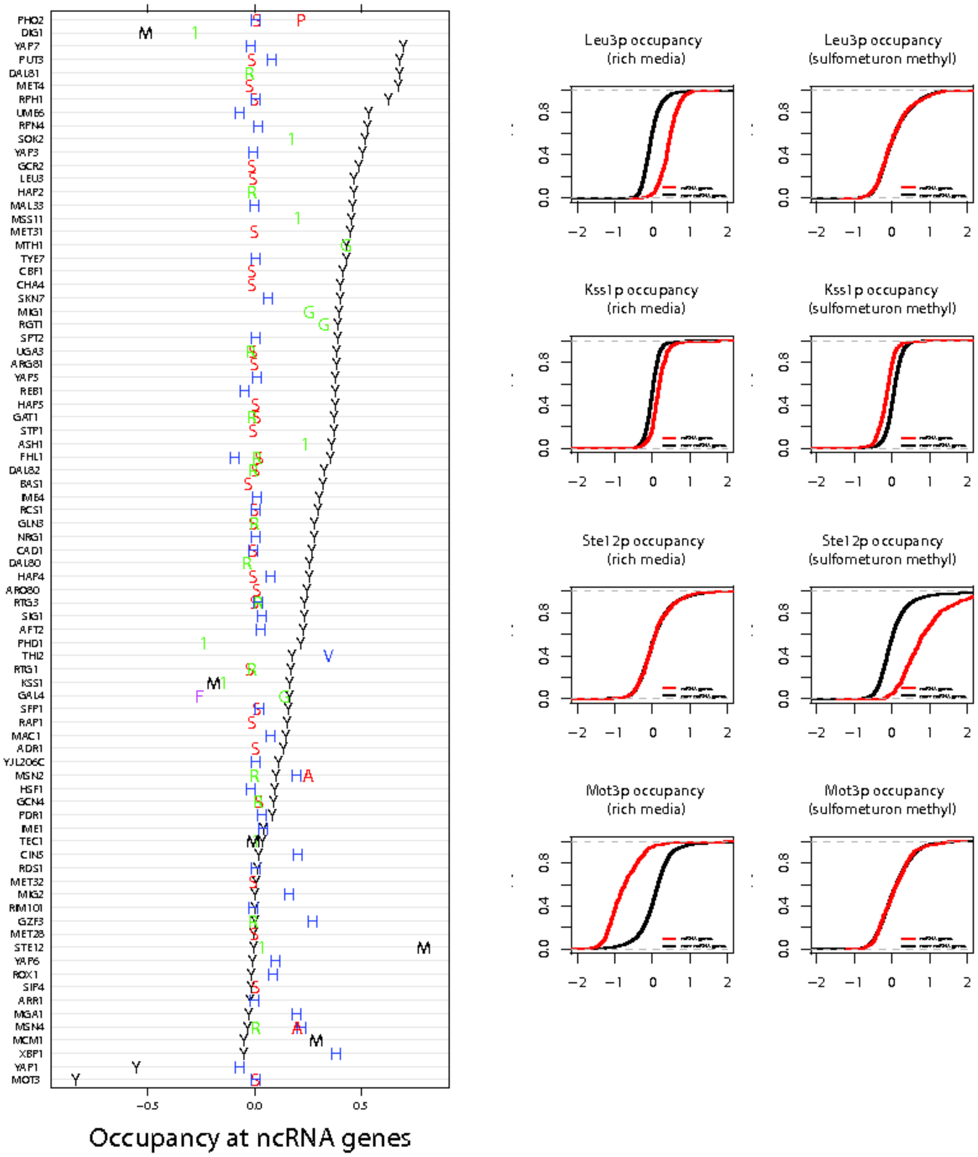

Figure 4.3: Condition specificity of occupancy at ncRNA gene. (**A**) Each row is a TF, and experimental conditions for that TF are plotted on the same row with letters indicating the condition. Conditions are: "Y", rich media; "S", sulfometuron methyl; "R", rapamycin; "H", hydrogen peroxide; "1", 1-butanol; "A", succinic acid; "G", galactose; "V", vitamin deprived medium; "M", alpha mating factor; "F", raffinose; and "P", phosphate deprived medium. Occupancy is expressed as $\Delta \bar{M}$, the difference between the mean log2 fold enrichment of ncRNA gene probes and the mean log2 fold enrichment of all other probes. (**B**) Leu3p, Kss1p, Ste12p, and Mot3p occupancy at ncRNA genes in rich media vs. sulfometuron methyl treatment. For each factor and conditions, an empirical cumulative distribution function is shown contrasting the distribution in log2 fold enrichment (FE) for ncRNA gene probes and all other probes.

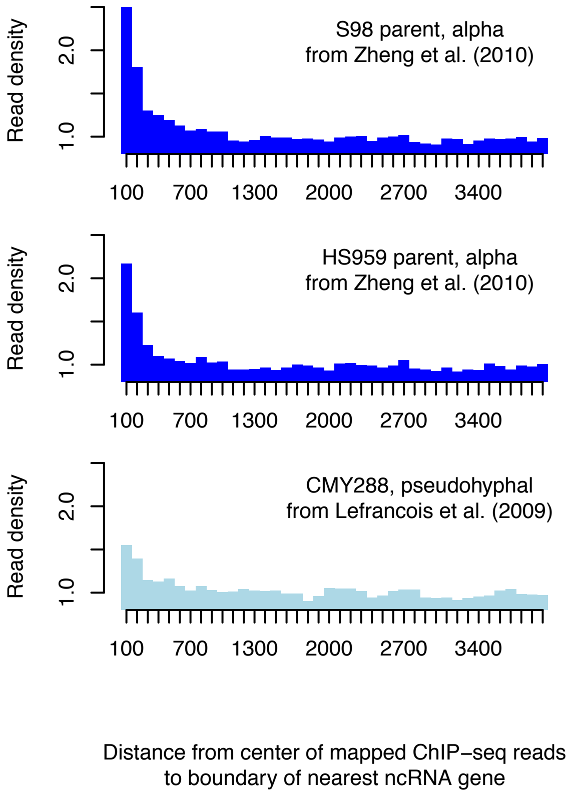

Figure 4.4: ChIP-Seq validation. Density of Ste12p ChIP-seq reads relative to the genome-wide coverage for the two parents tested under exposure to alpha mating factor in Zheng et al [140] and the strain tested under pseudohyphal growth conditions in Lefrancois et al. [141].

4.1.2.4 The RIM15 locus genetically modulates the degree of Ste12p targeting to ncRNA genes

In spite of the fact that ChIP-Seq is currently the state of the art in genomic protein mapping, only a handful of factors have been assayed in yeast. For validation purposes, we compared three Ste12p ChIP-Seq datasets, one of which was performed in pseudohyphal conditions and two in exposure to alpha mating factor [140, 141]. Both showed enrichment near ncRNA genes, although the magnitude was greater during exposure to alpha mating factor, consistent with the experiments of Harbison et al. (**Figure 4.4**). Zheng and colleagues profiled Ste12p occupancy in two diverged yeast strains, S96 and HS959, as well 43 genotyped segregants derived from a cross between these strains, in order to infer *cis*- and *trans*-regulatory polymorphisms affecting binding of the factor. To quantify the amplitude of the hotspot phenomenon for Ste12p in each segregant, we determined the fraction of ChIP-Seq reads mapping to within 100 bp of a ncRNA gene. Mapping this fraction as a quantitative trait, we found that highly significant linkage (LOD score = 6.08) with the hotspot effect for Ste12p occurs on chromosome VI (**Figure 4.5**). This locus contains three genes, one of which is *RIM15*. The regulatory kinase Rim15p is known to integrate signals from several distinct nutrient-sensing pathways (PKA, TORC1, Sch9, and Pho85-Pho80) and be required for entry into the stationary (G0) phase mediated by transcription factors Msn2/4 and Gis1 upon nutrient deprivation [142, 143]. This is consistent with our observation that the hotspot effect for Ste12p is condition-dependent, and suggests that genetic variation in the *RIM15* gene is a plausible modulator of hotspot behavior for Ste12p. Alignment of the S96 and HS959 protein sequences for Rim15p reveals that the *RIM15* allele from strain HS959 encoded five amino-acid changes, four of which are non-conservative when compared to the S96 sequence, as well as one amino-acid deletion.

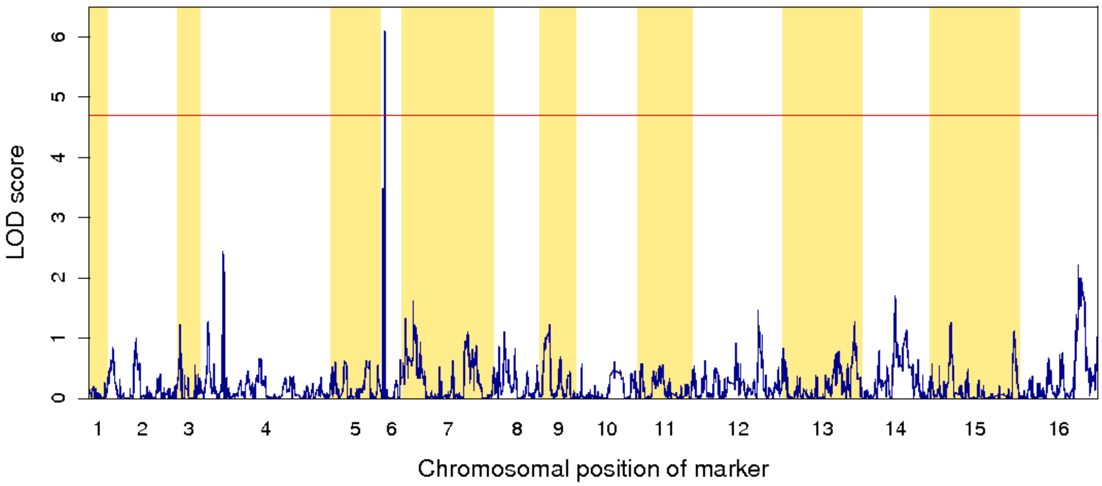

Figure 4.5: QTL profile. Manhattan plot of genome-wide association between genotype and hotspot behavior identifying the RIM15-containing locus.

4.1.3 Discussion

4.1.3.1 Possible mechanisms underlying the dynamic targeting of ncRNA gene by TFs

Several lines of cytological evidence from mammalian cells suggest that transcription by polymerase II occurs at nuclear foci comprising many polymerase molecules and transcription factors, termed "transcription factories" [132]. If such factories exist in yeast, it is conceivable that nucleosome-free regions and ncRNA genes – which are associated with high levels of transcription (by polymerase II and I/III, respectively) – are in close proximity to multiple TFs as a result of transcription factories. Consistently, it was recently discovered that Pol II-associated transcription factors tightly associate with Pol III-transcribed genes in human cells [144].

4.1.3.2 Yeast as a platform to genetically dissect TF colocalization

Although hotspot behavior is familiar to investigators performing ChIP in metazoans, it had not been previously recognized in yeast, which is the most genetically tractable model eukaryote. Our analysis of the Ste12p data by Zheng and colleagues [140] reveals an association between polymorphisms in the RIM15 gene and localization of Ste12p to ncRNA genes during exposure to alpha mating factor. Multiple nutrient-sensing pathways converge on the Rim15p regulatory kinase, which regulates entry into the G0 phase and mediates calorie restriction-dependent life span extension [142, 143, 145]. Based on this previously observed function of Rim15p in combination of our analysis of the Ste12p ChIP-seq data [140], we were able to implicate polymorphisms in the *RIM15* gene as a plausible genetic factor underlying the hotspot behavior for Ste12p. Whether RIM15 is uniquely associated with Ste12p localization or is a general regulator of TF colocalization remains to be determined. The ability of yeast ncRNA genes to serve as a model for the hotspot behavior of TFs opens the door for further genetic screens and biochemical validation.

4.1.4 Methods and Materials

4.1.4.1 Processing of raw ChIP-chip data

The original raw ChIP-chip data [135, 136] were obtained from ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) using accession numbers E-WMIT-1 and E-WMIT-10, respectively. Protocol information for each array (which dye was IP vs. WCE, experimental conditions, etc.) was extracted from the files E-WMIT-1.sdrf.txt and E-

WMIT-10.sdrf.txt, available in the directory http://www.ebi.ac.uk/microarray-as/ae/download/. Raw intensity information was downloaded from the tab-delimited text files in E-WMIT-1.raw.zip and E-WMIT-10.raw.zip available in the FTP directory specified within the aforementioned text files. The column headers in all of these text files were found to be corrupted. Therefore, they were split between nine different formats. Each format was manually curated to locate the correct median foreground and background red and green intensity columns, using the presence of a background-subtracted log ratio column as a validation. Raw intensities were loaded into *R* and Loess normalization was performed on each array (to account for dye-specific response functions) using the normalizeWithinArrays function of the *limma* package [146], resulting in an M (relative intensity) and A (absolute intensity) value for each spot on each array. A number of the arrays were found to have very low variance in their log ratios; arrays with a variance in M after Loess normalization less than 0.05 were discarded. Four summary values were calculated for each probe: a median log ratio (M) and intensity (A) signal across all rich-media (YPD) arrays, and a median log ratio (M) and intensity (A) across all stress arrays. Additionally, for every experimental condition for which multiple replicates were available, a median M and A value across replicates was calculated. The same processing was applied to ArrayExpress data from assaying rabbit IgG control, no-antibody control, and kinase occupancy by tiling array [72, 147], which we used for validation.

4.1.4.2 Comparison of occupancy at annotated targets and at ncRNA genes

Annotated targets for each TF were defined as probes that overlapped or were neighboring (within 100 bp of) regions reported by MacIsaac et al. [95] within their *p*-value threshold of 0.005. After discarding probes that were annotated both as ncRNA probes (according to the criterion described above) and as TF targets, we compared the mean $\log_2$ fold enrichment among ncRNA probes and among annotated targets with that of all other probes. A significant difference in means was defined as a *t*-test passing a *p*-value threshold of 0.05, Bonferroni corrected for the number of tests.

4.1.4.3 Gene Ontology analysis

Functional enrichment of probes by Gene Ontology (GO) categories [148] was determined using the *T-profiler* algorithm [149], implemented using the *YEAST* package from the *BioConductor* platform within the *R* programming environment [150].

4.1.4.4 Regression analysis of gene annotation, nucleosome depletion, and co-occupancy

A linear regression was performed between nucleosome occupancy [151] and MFE across all probes. Residuals corresponding to ncRNA probes and all other probes were then compared using a t-test. Residuals were also analyzed for correlation with predicted affinity for TFs as described above.

4.1.4.5 ChIP-seq analysis

ChIP-seq data from Lefrancois et al. [141] and Zheng et al. [140] were downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). These data include mapped peaks, but not genome-wide mapping of reads; therefore, read alignment results from ELAND were downloaded and processed using MACS [152] as described by the authors in order to obtain a genome-wide landscape of binding, in 10-bp bins. Distances from these bins to ncRNA genes were measured using the SGD genome annotation described above and BEDTools [153]. Bins not overlapping an ncRNA gene with a distance between 1 and 100 bp of an ncRNA gene were considered hotspots, and a quantitative trait for Ste12p localization to hotspots was defined as the ratio of reads mapping to this region to the total number of reads. This quantitative trait was calculated for both parents and each of the segregants for linkage analysis.

4.1.4.6 QTL analysis

To detect genetic contributions to the variation in hotspot behavior for Ste12p by specific loci, we used 55958 markers genotyped for 43 segregants from previous studies [140, 154]. Following these authors, we removed 2498 markers with fewer than five occurrences of either parental genotype and grouped the neighboring markers with the same genotype distribution across segregants, which resulted in 2592 non-redundant markers. We performed Composite Interval Mapping (CIM) of Quantitative Trait Loci (QTLs) as implemented in the R/qtl package [119], which uses multivariate regression on multiple markers for increased precision [97]. Statistical significance was determined by performing 50 independent random permutations for each segregant. The FDR corresponding to a given LOD score threshold was computed as the ratio of the average number of markers above threshold and the number of markers for which linkage was detected. A 5% FDR based on the empirical permutation test corresponded to a LOD score greater than 4.70.

4.2 Harnessing natural sequence variation to dissect post-transcriptional networks in yeast

This section has been adapted from a manuscript co-authored by Mina Fazlollahi, Eunjee Lee, Xiang-Jun Lu, Maria Pilar Gomez-Alcala and Harmen J Bussemaker. It will be submitted for publication in the near future. For this work, Mina Fazlollahi applied the aQTL software that I developed to the regulation of mRNA stability by RNA-binding proteins.

4.2.1 Introduction

Regulation through post-transcriptional activities of RNA binding proteins (RBPs) is critical mechanism for a living cell to control the mRNA levels. It includes assembly, edition, localization and stability of RNAs. However there have been more studies, which focus on DNA biding factors and their interaction network. Detecting binding motifs associated with RBPs are challenging due to more complicated structure of RNAs.

In this study, we searched for potential regulatory elements in mRNA sequences that are recognized by diverse RNA-binding proteins. We used a similar approach by [78] which searches for binding sites in the form of sequence-specific affinity matrices (PSAMs). We use REDUCESuite package (more specifically MatrixREDUCE software) to discover binding motifs. We then used these obtained PSAMs to score mRNA sequences. These score act as prior information for our QTL analysis. Together with the mRNA differential expression levels for the every member of the population, we can infer the activity levels of the RBPs for every individual. We treated these activity levels as quantitative trait to discover loci modulating them as used in [90].

With our PSAM finding analysis we were able to obtain known binding motifs for 15 various RBPs and novel motifs for Scp160p, Sik1p and Tdh3. For the activity QTL (aQTL) analysis we recovered a known locus that contains MKT1 gene, which is shown to modulate Puf3p interaction with its targets [110]. Interestingly, we found that depending on Puf3p interaction with 5′ or 3′ UTRs of its target, there are different loci regulating its activity. For 5′ UTRs, we found a locus on chromosome 2 that contains POP7 gene.

4.2.2 Results

For the PSAM search, we used RNA imunoaffinity purification [155] including total of 132 IP experiments for 45 different RBPs and for the aQTL search we used [91] including mRNA expression levels for 108 segregants obtained by a genetic cross between two haploid Saccharomyces cerevisiae strains: BY (laboratory strain) and RM (a wild isolate form a vineyard in California).

4.2.2.1 PSAM search

We used MatrixREDUCE software form REDUCESuite package that takes as inputs the nucleotide sequences and RBP binding data. To reduce the effect of outliers, we transformed the binding data into the quantiles of the ranks for each IP experiment, thus transforming it into a normally distributed data. As for sequences, we used the annotations from [156] and extracted 5′ and 3′ untranslated regions (UTRs), open reading frames (ORFs) and whole mRNA nucleotide sequences in yeast. For every RBP, we performed our genome-wide motif search on whole mRNAs, ORFs, 5′ and 3′ UTRs separately due to different binding sites for the cases that the protein prefers to bind directly or indirectly.

We optimized these PSAMs by adding flanking sides (maximum of one nucleotide added to either sides at every optimization iteration) to capture any low specificity sites that were not captured during our training step. Out of 45 proteins, As a results, we were able to obtain PSAMs for 20 different RBPs.

We checked the correlation between 132 IP experiment and affinity of each mRNA segment separately and only accepted the RBP- mRNA region combination that is exclusively correlated to its own RBP binding experiments (specificity test). **Figure 4.6A** shows the PSAM logos for 15 out of 20 PSAMs that passed the specificity test. **Figure 4.6B** shows all the 25 RBP-mRNA region combination that passed the specificity test. There was an exception with Scp160p-ORF where the affinity is slightly more correlated to Bfr1p IP experiment (green dots) However Bfr1p is reported to associate with cytoplasmic mRNP complexes containing Scp160p [157]. We observe that there is a large gap between the relevant IP experiments (red dots) and the rest of the IP experiments (blue dots) for the affinity of 3′ UTR of Pub1p, Puf2p, Puf3p, Puf4p and Puf5p. This indicates that these PSAM are highly specific to the binding data for their RBPs.

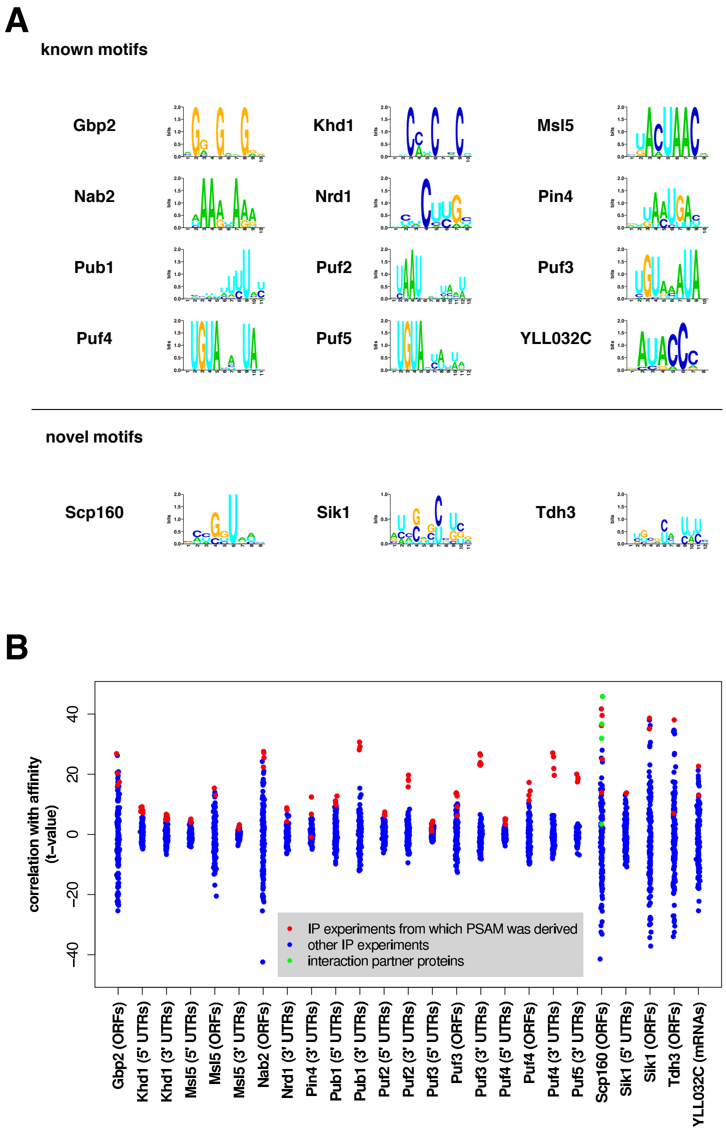

Figure 4.6: RBP PSAMs. (**A**) Recovered and discovered RBP PSAMs logos. These 15 PSAMs passed cross-validation and specificity tests. (**B**) Scatter plot for the factors specificity test where the Pearson t-values of univariate linear fit coefficients between 132 RBP binding experiments and 25 selected PSAM-sequence combinations are presented. Only the factors with at least one self RBP IP experiment t-value (red dots) appearing at the top are shown. The only exception is for Scp160 (ORF) where we have a higher correlation to Bfr1p binding data (green dots). We accepted this PSAM, since Scp160p and Bfr1p are know to interact and are co-imunopercipitated in IP measurements.

4.2.2.2 Inferring RBP activity and linkage analysis

We used segregant-specific genomewide mRNA expression levels to predict RBP activity variation among 108 yeast segregants. For each of 25 RBP-mRNA regions, we first scored all of the genes by calculating the affinity of the PSAM using the chosen sequence (whole mRNA, UTRs or ORFs). Then we performed multivariate regression between all 25 factors affinities and each segregant mRNA expression data. We considered the coefficients of the regression as the quantitative representation of the RBP activities. We used composite interval mapping (CIM) [97] to find aQTL for each factor. To correct for multiple testing, we calculated the lod-odds (LOD) score threshold for 1% false discovery rate (FDR) level by 200 permutations.

4.2.2.3 aQTL discovery

Using mRNA expression levels data to infer the RBPs activity levels, we were able to detect significant aQTLs for 7 different RBPs (9 factors) where 1 of them has been previously reported (**Figure 4.7**) and the rest are new.

Recovered aQTL for Puf3p

Our method was able to recover a locus on chromosome 14 linked to MKT1 for Puf3p when we calculated its activity using 3′ UTRs sequences (**Figure** 4.7C). This locus was previously discovered computationally and experimentally by [110]. They have suggested that MKT1 regulates p-body abundance, which consequently regulates Puf3p target abundance.

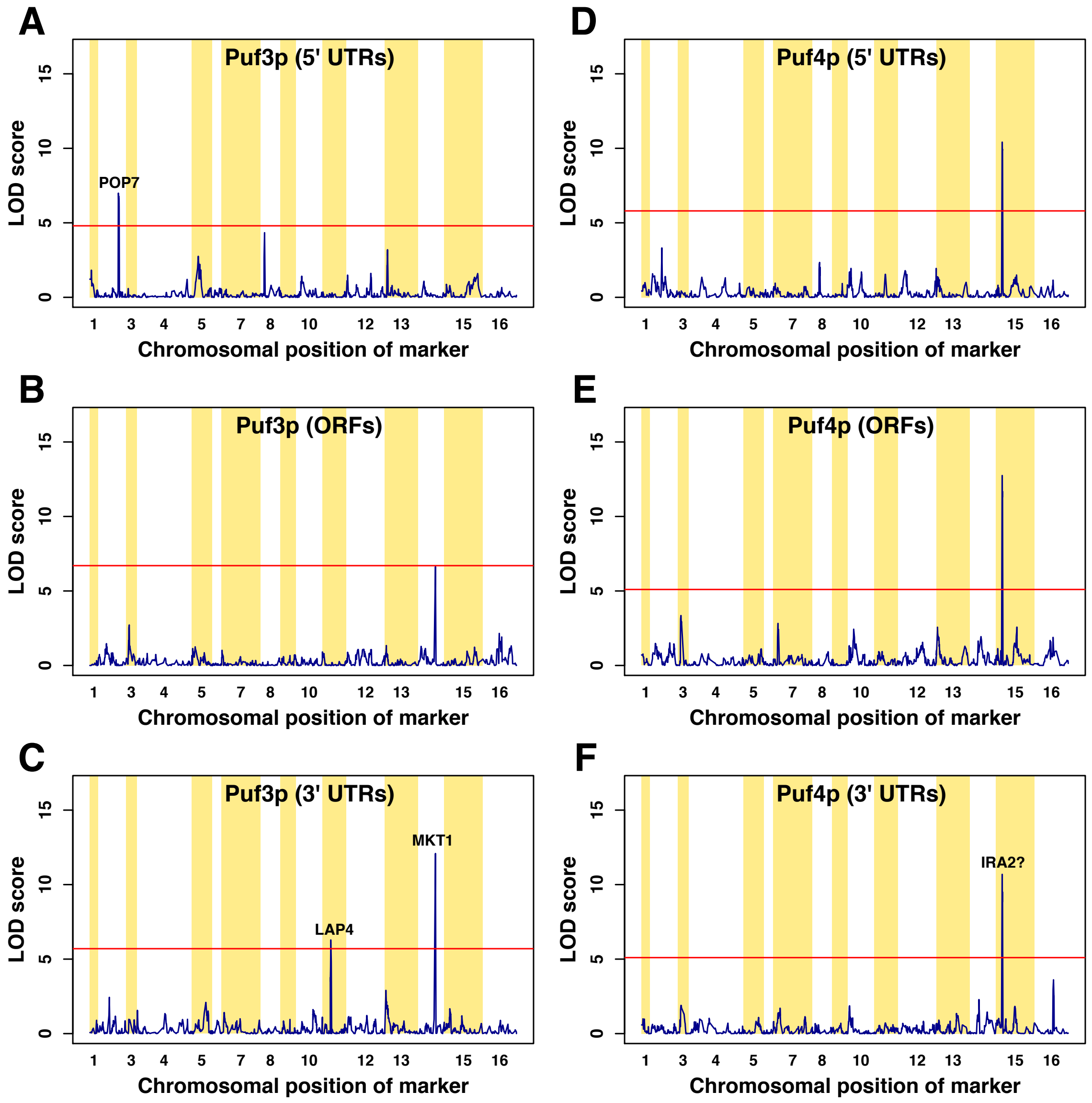

Figure 4.7: aQTL profile of RBPs. **A-F**) Results of the trans-acting genetic modulators of RBP activity of Puf3p and Puf4p mapped using our aQTL method. Significant aQTL regions at a 1% FDR level are marked calculated using 200 independent permutations of expression data. For each factor, these regions survived after filtering out for the 3 groups of genes mentioned earlier. The red line represents the LOD score threshold for the 1% FDR level. Sections A-C show aQTL profile for Puf3p activity and D-F for Puf4p activity on 5′ UTRs, ORF and 3’ UTRs respectively. We were able to detect 2 novel trans-acting loci (chr2 and chr11) and a previously reported one (chr14) For Puf3p. Puf4 activity showed a significant linkage to a locus on chr15 irrespective of which mRNA region we used for activity calculation. This locus includes IRA2.

Puf3p activity modulated differently depending on binding to 5′ UTRs and 3′ UTRs

Previously it was mostly known that Puf3p interact with the 3′ UTRs of its targets and no evidence of functional interaction with the mRNAs 5′ UTRs has been reported. As mention above, Puf3p activity is modulated through a locus on Chr14 when considering its binding to 3′ UTRs of its targets. However considering Puf3p binding to 5′ UTRs, we were able to link its activity level variation to a locus on Chr2 (**Figure 4.7A**). This region contains POP7, which is reported to have positive genetic interaction with Puf3p [158]. These findings indicate that the activity modulation of Puf3p is linked to different genomic locations and networks depending on where it binds to the mRNA.

Puf4p activity modulation is independent of where on target mRNA it binds

Puf4p aQTL profile is shown in **Figure 4.7D-F**. Whether Puf4p binds to 5′ UTRs, ORFs or 3′ UTRs of its targets, its activity regulation is controlled by a locus on chr15. This locus contains REX4 and BRX1. Both of them are involved in pre-rRNA possessing and ribosome assembly. It is known that Puf4 interacts with mRNAs encoding nucleolar rRNA-prossessing factors.

4.2.3 Methods and Materials

4.2.3.1 Experiment data sets

For our RBP PSAM search, we analyzed genome-wide RNA imunoaffinity purification data, which was performed for 45 different RNA binding proteins by [155] (GEO accession number GSE13135). The mRNA levels bound to each RBP were measured using the C-terminal tandem affinity purification (TAP)-tagged proteins and were affinity purified from whole cells grown in YPD. The data includes 2 to 6 replicated for each factor total among cDNA and oligonucleotide microarray thus having total of 132 IP experiments.

For aQTL analysis, we used the genome-wide mRNA expression data (GEO accession GSE9376) using a similar approach done by [90]. 108 haploid segregants mRNA expression levels from a genetic cross between two parental strains BY and RM were measured. The $\log_2$ ratios are between segregants and the BY mRNA differential levels. The pre-processing on the expression data is the same as done by [90]. As for genotype data, we used [19] data for the same segregants for total of 2956 markers.

4.2.3.2 PSAM search

We used the normalized $\log_2$ ratios between the mRNA level bound to a protein and the control from the RBP binding data set. To correct for the effect of outliers, we applied a rank-quantile transformation. For each IP experiment, we first ranked the normalized $\log_2$ values among all genes and then assign the quantile values to the ranked values. This way, we diminish the values points located in the long tails of the distribution.

We used the MatrixREDUCE program from the REDUCE Suite (http://bussemakerlab.org/software/REDUCE) to perform a genomewide fit of a position-specific affinity matrix (PSAM) to the rank-quantiled normalized log2 ratios of IP experiments. For every RBP, we searched for binding motifs on the whole mRNAs, 5′ UTRs, ORFs and 3′ UTRs sequences separately. We obtained the Saccharomyces cerevisiae UTR sequences from a study using RNA-seq method to obtain the transcriptional landscape of the yeast genome by [156]. ORF sequences were downloaded from Saccharomyces Genome Database (SGD; http://www.yeastgenome.org). The final set of PSAMs was obtained from the PSAMs that pass the test for specificity to their own IP experiment among all the experiment and If the affinity of a specific PSAM on UTRs and/or ORFs had the highest correlation to at least one of the relevant IP experiments among the 132 experiments, that RBP-mRNA region combination would pass our test.

4.2.3.3 Inferring segregant-specific RBP activities

From our RBP PSAM search we obtained 25 independent RBP-region factors. As for the previous work by [90], we used the total sequence affinity of the PSAMs as a predictor for mRNA differential expression levels in the low protein concentration region established by [78]. Thus we considered the RBP sequence occupancy to be proportional to the total affinity of a desired PSAM for a sliding window along the whole mRNA, UTRs or ORF sequences. We performed a genome-wide multivariate regression between the 25 factors we had obtained from the RBP PSAM search and specificity test step and every segregant mRNA expression $\log_2$-ratios. For the case of protein level, we performed the regression between the affinities and every segregant protein levels.

4.2.3.4 aQTL mapping

Significant aQTL region were discovered by splitting the multivariate regression coefficient between BY and RM at every marker and testing for the significance of the difference between the distributions of the two groups of coefficients using composite interval mapping (CIM) method for maximum resolution. Thus LOD score was calculated to check for the effect of each locus on the activity of the RBPs. We performed

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

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

Chapter 5

Identifying regulatory mechanisms underlying tumorigenesis

This chapter has been adapted from a manuscript co-authored by Eunjee Lee, Jeroen de Ridder, Lodewyk Wessels and Harmen J. Bussemaker. It is in preparation to be submitted for publication in the near future.

5.1 Abstract

Retroviral insertional mutagenesis (RIM) is a powerful tool for identifying putative cancer genes in mice. To uncover the regulatory mechanisms by which common insertion loci affect downstream processes, we supplemented genotyping data with genomewide mRNA expression profiling data for 97 tumors induced by RIM [68]. We developed LESA: locus expression signature analysis, an algorithm to construct and interpret the differential gene expression signature associated with each common insertion locus. Comparing locus expression signatures to promoter affinity profiles allowed us to build a detailed map of transcription factors whose protein-level regulatory activity is modulated by a particular locus. Surprisingly, we found that the transcriptional response to insertion at the MYC locus is dominated by Myc binding events downstream of transcription start sites. Our analysis also revealed the induction of a large multi-gene chromosomal domain in response to insertions near the PVT1 gene. Taken together, our results demonstrate the

potential of a locus-specific signature approach for identifying mammalian regulatory mechanisms in a cancer context.

5.2 Introduction

Cancer arises as a result of the accumulation of genetic and epigenetic changes, which each deregulate a specific aspect of normal cell function. High-throughput technologies for mapping genetic and chromosomal aberrations have revealed complex changes in genomes of individual tumors [40-43]. Mouse retroviral insertional mutagenesis has been used as an efficient tool for identification of causal mutations in cancer [44, 45]. Such oncogenic mutations may either cause alteration of a gene product or influence the expression levels of one or more genes surrounding the insertion.

A high-throughput screen in mice for mutations collaborating with either p19^{ARF} or p53 deficiency was performed yielding over 10,000 independent insertion sites for more than 500 tumors [61]. To date, more than 10,000 genomic regions where retroviral insertions have been found in close proximity in multiple tumors (referred as Common Insertion Sites or CIS) have been identified in previous studies (Retroviral Tagged Cancer Gene Database (RTCGD), http://variation.osu.edu/rtcgd [62]). However, insertional mutagenesis screens have an important limitation. It is not straightforward to determine which genetic lesion near the CIS is playing a causal role in oncogenesis. Methods to predict target genes affected by insertions based on information about insertions such as orientation have been developed [68]. However, these are based on gene annotation and do not use or provide functional evidence, i.e. whether the mRNA expression level of the predicted target genes is affected by insertions. Analyzing the mRNA expression level of the genes near the insertion site is often not sufficient, as the lesion may exert its effect only at the post-translational level. Furthermore, retroviral insertional mutagenesis screens alone rarely elucidate the regulatory mechanisms that drive tumorigenesis. These limitations highlight the need for new approaches that can integrate the genetic data with functional genomics data and other information in order to identify causal genes and regulatory mechanisms underlying cancer.

For these reasons we carried out genomewide expression profiling in a set of tumors induced by retroviral insertional mutagenesis. We furthermore describe a novel approach that first defines and then analyzes the genomewide mRNA expression response that is induced by the putative causal gene near the insertion locus. We refer to our method as Locus Expression Signature Analysis (LESA). We analyzed genome-wide expression profiles for a panel of splenic tumors induced by retroviral insertional mutagenesis [68]. To identify regulatory mechanisms underlying tumorigenesis, we hypothesized that gene expression is affected by insertional mutations through one of two regulatory mechanisms: (i) regulation by sequence specific transcription factors (TFs) or (ii) changes in chromosomal domain organization leading to changes in gene expression. We construct a map that connects each insertion locus with the TFs whose regulatory activity is affected by it. This map contains both know and new associations. LESA also provides an opportunity to analyze transcriptional regulation mechanisms in great detail. In particular, we show that the MYC expression signature is mainly driven by promoter-proximal downstream binding of the c-Myc transcription factor to the transcription start site (TSS) of its target genes. LESA also identified a large chromosomal domain on chromosome 15 in which most genes are induced by viral insertions at the PVT1 locus. Together, our results demonstrate the ability of LESA to identify candidate regulatory mechanisms in cancer.

5.3 Results

5.3.1 Identifying common insertion sites

Many insertion regions are tagged in multiple independent tumors. To identify common insertion sites (CISs), i.e., regions in the genome that are significantly more frequently mutated by insertions than would be expected by chance, we used a statistical framework based on Gaussian kernel convolution (GKC), which estimates a smoothed density distribution of inserts over the entire genome [67] Next, we applied rule-based mapping (RBM), which maps individual insertions to target genes based on orientation-dependent windows and information regarding repetitive occurrence of insertions at a given locus across tumors [68]. We identified 87 insertion sites, and assigned a potential modulator gene that generates the tumors in collaboration with the loss of p53 and p19^{ARF}. For further analysis, we selected 13 insertion loci that occur in more than 10 tumors. The loci include known oncogenes such as *NOTCH1, c-MYC* and *n-MYC.*

5.3.2 Locus Expression Signature Analysis (LESA)

Our hypothesis is that an insertional mutation perturbs the function or expression level of the proximal causal gene contributing tumorigenesis in collaboration with loss of p53 or p19^{ARF}, and the perturbed causal gene in turn influences the function or expression level of downstream target genes in the signaling pathway (**Figure 5.1A**). Therefore, the contribution to tumorigenesis by a causal mutation might have a characteristic gene expression signature associated with it. Based on this assumption, we inferred locus expression signatures (LES) representing the average difference in mRNA expression between the tumors carrying insertions at a particular locus and all other tumors (**Figure 5.1B**).

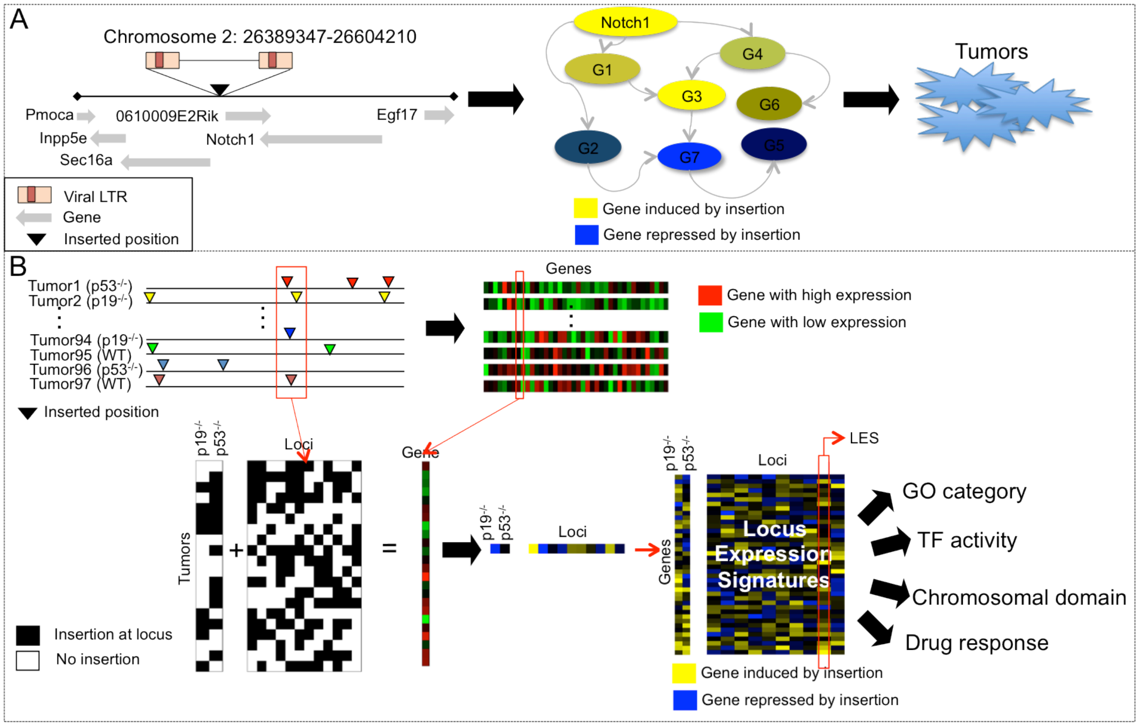

Figure 5.1: The overview of LESA. (**A**) The proviral insertion in to host genome perturbs the function or expression level of the proximal causal gene contributing tumorigenesis in collaboration with loss of p53 or p19^{ARF}, and the perturbed causal gene in turn influences the function or expression level of downstream target genes in the signaling pathway (**B**) The contribution to tumorigenesis by a causal mutation might have a characteristic gene expression signature associated with it. Based on this assumption, we inferred locus expression signatures (LES) representing the average difference in mRNA expression between the tumors carrying insertions at a particular locus and all other tumors using common insertion sites data and gene expression data for each tumor. These LES are used for further analyses including finding significant GO category, differential TF activities, chromosomal domain and drug responses to identify regulatory mechanisms underlying tumors.

In a previous analysis of the insertion data *per se* [61], insertion loci and genetic background (e.g.p19$^{ARF-/-}$, p53$^{-/-}$, and wild-type) were found to co-occur across tumors, which is an indication for collaboration between these lesions in driving tumorigenesis. To deal with this dependency, we used a multivariate linear model to explain the variation in mRNA expression level for each gene in terms of the pattern of insertions in a given tumor (**Figure 5.1B**). The coefficients from this fit, aggregated across all genes, constitute our collection of locus expression signatures. For each insertion locus, we compared the variance of the signature value across all genes with its null distribution (see Methods). We found that for most loci the information contained in their signature is highly statistically significant (**Figure 5.2B**). This encouraged us to dissect the regulatory mechanisms that give rise to these signatures.

5.3.3 Regulation of individual genes as identified by LESA.

The absolute value of the locus expression signature is highest for genes whose expression level is most strongly affected by the associated genetic mutation. As expected for the signatures we inferred for the contribution of the p53$^{-/-}$ and p19$^{ARF-/-}$ deletion backgrounds, respectively, Trp53 had the most negative value in the p53 signature, while the p19^{ARF} isoform Cdkn2a showed one of the most negative p19^{ARF} signature values (**Table 5.1**). We also found an association between the loss of p53 and the increase of Cdkn2a expression, suggesting a feedback mechanism compensating for the loss of p53. Among the remaining signatures, i.e., those associated with common insertion loci, we detected strong cis-association for *NOTCH1*, *RASGRP1*, *GFI1*, and *RRAS2*, with retroviral insertions either increasing or decreasing the mRNA expression level of the gene designated as their primary target (**Table 5.1**). To analyze the extent to which an insertion could have a wider local effect, we inspected the expression signature values near the insertion locus. The expression of multiple genes proximal to *CCND3* and *RUNX1* was highly induced by retroviral insertion (**Figure 5.2A**). In other cases, insertions may affect genes at the protein level rather with no apparent change at the mRNA level (as exemplified by the *MYC* expression signature in **Figure 5.2A**), resulting in the change in the transcription level of functionally related or known direct target genes. Examples in Table 1 include the *CCND3* locus influencing the mRNA of its know target Ccnd2 [159], and the *MYC* locus that of Ccnd2 [160]. We found that the gene Adam19 has a high *NOTCH1* signature value, suggesting a positive feedback loop: Adam19 is a metalloprotease known to activate Notch1 by cleaving it [161].

p19		**p53**		**Notch1**		**Myc**		**Med20/Ccnd3**	
Slpi	0.77	**Cdkn2a**	1.66	Gm12253	2.05	Myl1	0.94	Emb	0.66
RP23-395H4.4	0.62	**Cdkn2a**	1.45	**Notch1**	1.92	Gimap7	0.64	Lysmd2	0.51
NM_027222	0.61	Butr1	1.27	Dtx1	1.86	Grb7	0.57	Ly6k	0.50
Reg1	0.58	Ifi27l2a	1.03	Aldh1b1	1.65	Chst2	0.56	Itgb7	0.49
Try10	0.58	Isg15	1.03	NR_002860	1.45	Tns4	0.53	Gimap7	0.48
Amy2	0.57	Sparc	0.90	**Adam19**	1.42	**Ccnd2**	0.52	Il18r1	0.47
Vpreb3	0.56	Vpreb3	0.82	Spsb4	1.35	Cd160	0.51	Thy1	0.45
Pnlip	0.56	Usp18	0.77	Rag1	1.35	Bmp7	0.47	Kcnf1	0.44
Il4i1	0.55	Retnlg	0.65	Cd163l1	1.28	Rapsn	0.46	Gimap3	0.44
Pnliprp1	0.55	NM_027222	0.64	Ctla4	1.25	Gfi1	0.44	Rapsn	0.43
⋮		⋮		⋮		⋮		⋮	
Treml1	-0.23	NM_175332	-0.55	S100a10	-1.09	Prg2	-0.43	NM_027222	-0.46
Cdr2	-0.23	Xist	-0.55	Try10	-1.11	Rras2	-0.43	H2-T22	-0.47
Cdkn2a	-0.24	Lck	-0.56	Prss2	-1.11	Ddc	-0.46	**Ccnd2**	-0.48
Serpine2	-0.24	Sh2d2a	-0.58	Myl1	-1.17	S100a9	-0.46	Irf4	-0.49
Ednra	-0.25	Hdc	-0.62	Amy2	-1.18	Pp11r	-0.47	Plac8	-0.49
Ddc	-0.25	Thy1	-0.64	Clps	-1.20	Gpc1	-0.50	Adam19	-0.50
Ppbp	-0.28	Cd3d	-0.68	Zg16	-1.22	Cd163l1	-0.62	Notch1	-0.54
Mpp4	-0.31	NM_009831	-0.69	Pnlip	-1.27	Ly6d	-0.64	Gm12253	-0.57
Rsad2	-0.36	Cd3g	-0.71	Pnliprp1	-1.46	Ccnd1	-0.68	Myl4	-0.57
Gfi1	-0.39	**Trp53**	-1.77	RP23-395H4.4	-1.50	Satb1	-0.80	Myl1	-0.61

Table 5.1: The genes with the highest absolute value of the locus expression signature.

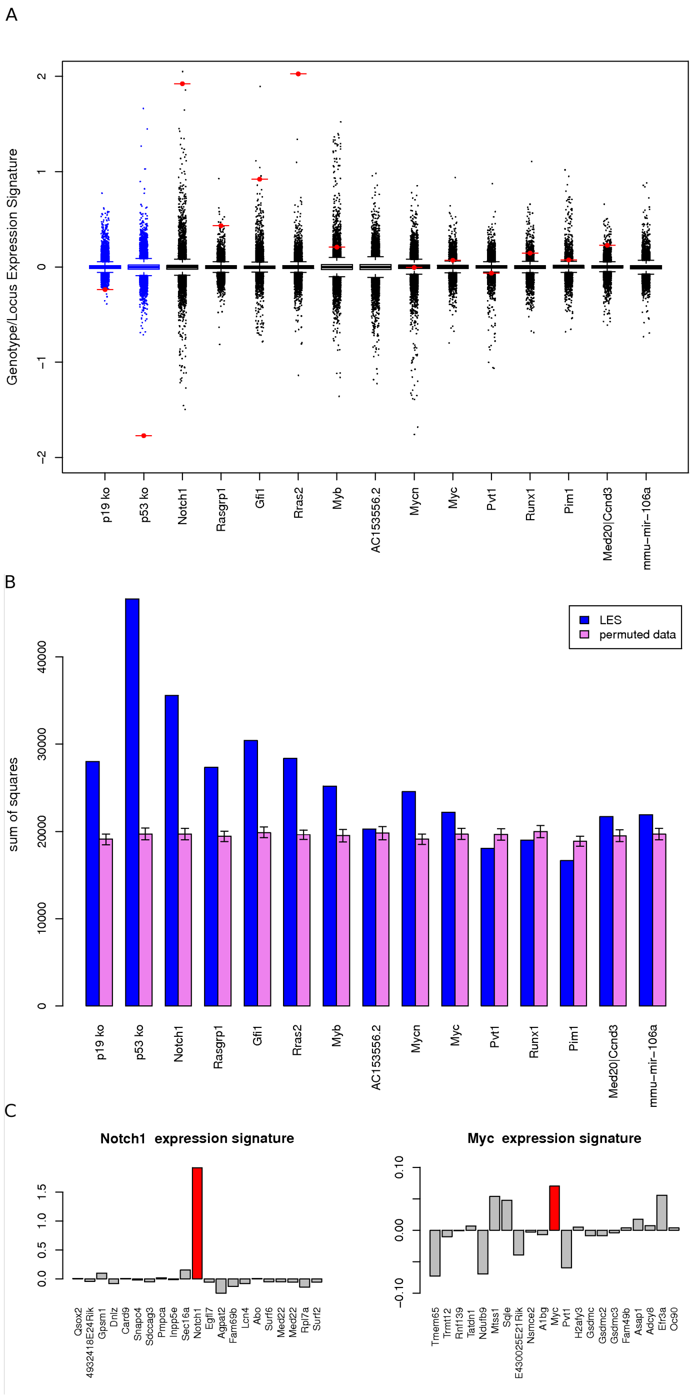

Figure 5.2: Single-gene expression changes identified by LESA. (**A**) Box plot of locus-expression signatures for each genotype (blue) and each insertion locus (black). Red spot represents the LES value of the near target gene. (**B**) Sum of squares of t-value corresponding to each LES (blue) and permuted values (pink) (**C**) LES values for genes in the surrounding regions of insertions for NOTCH1 and MYC insertions.

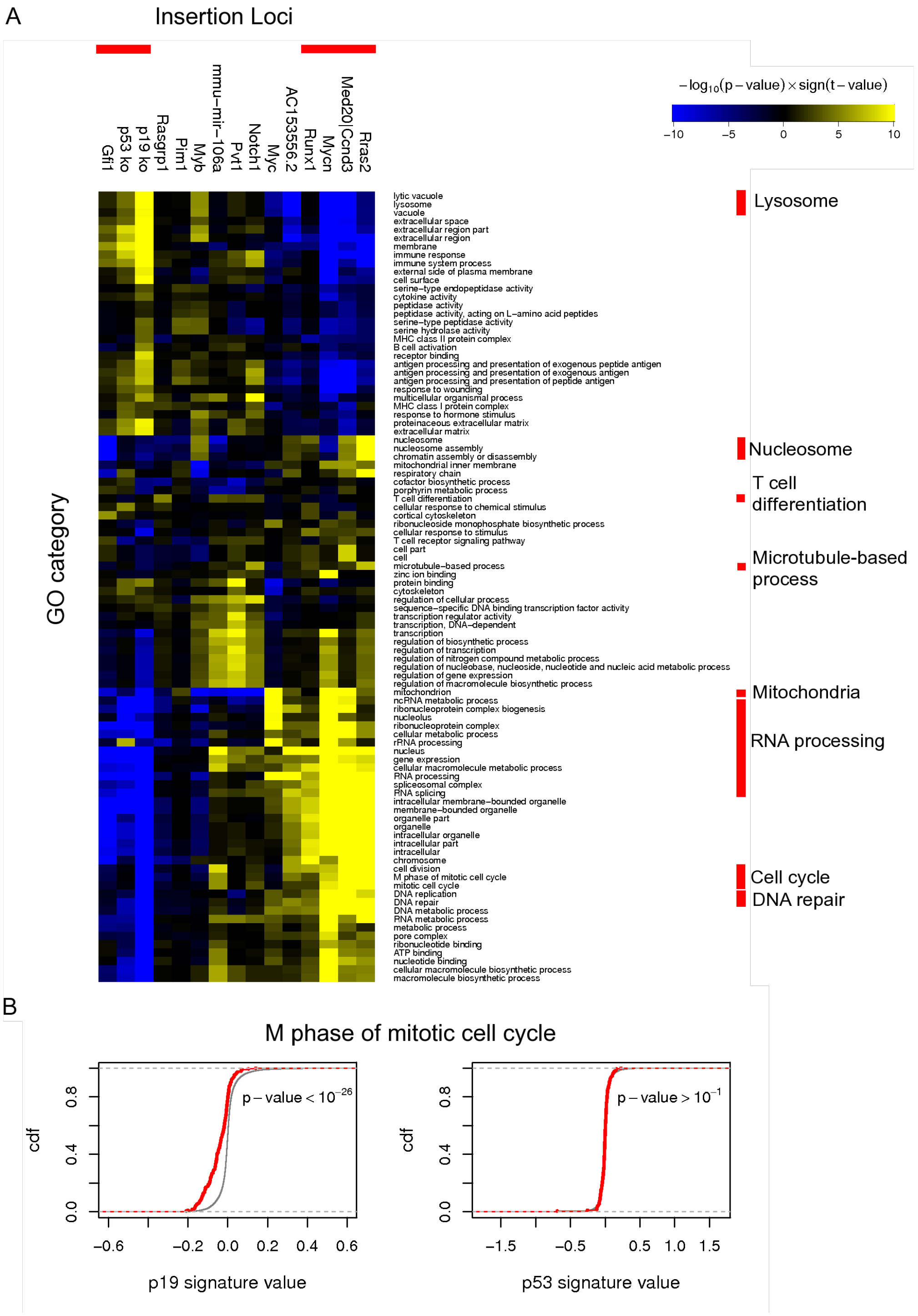

Figure 5.3: Functional annotation using GO category. (**A**) Heatmap of significant GO categories for each LES at FDR 1%. (**B**) CDF plot of p19 and p53 LES for genes in M phase mitotic cell cycle

5.3.4 Expression signatures can elucidate the biological functions affected by insertions at specific loci.

To explore the functional significance of the locus expression signatures, we used gene ontology (GO) terms to identify the biological process, molecular function, and cellular component categories enriched in each locus expression signature. We compared the distribution of the locus expression signature values in each particular GO category with that of the remaining genes using the Wilcoxon-Mann-Whitney (WMW) test. Since the GO categories are hierarchically organized, with overlapping gene sets that are mutually redundant, we used a forward selection scheme [149] to select a non-redundant set of significantly associated GO categories.

The resulting functional map provides several useful insights (Figure 3A). First, the GO categories associated with the p53$^{-/-}$ effect on expression are a subset of those associated with the p19$^{ARF-/-}$ background. For example, the “DNA repair” genes are suppressed in tumors lacking either p19^{ARF} or p53, consistent with the known role of these genes as activators of DNA repair [162]. By contrast, the “M phase of mitotic cell cycle” genes are enriched only in the p19^{ARF} expression signature (**Figure 5.3B**). This agrees with the fact that while p19^{ARF} acts upstream of p53 by enhancing p53 activity [163], it may also suppress tumorigenesis independently of p53 [164, 165].

“Lysosome” genes are significantly associated with the p19^{ARF}, MYCN, AC153556, and MED20/CCND3 signatures. The lysosome controls cell death and lysosomal alterations are common in cancer cells [166]. The mechanism mediating the effect of these insertions on the expression level of lysosomal genes is not clear. In addition, we detected association with "mitochondrion" genes for several loci (*p19*^{ARF}, *NOTCH1*, *MYB*, *MYCN*, and *MYC)*. Interestingly, p19$^{\mathrm{ARF}}$ and Myc are both known to regulate apoptosis via the release of cytochrome c from mitochondria [167, 168]. Several genes related to mitochondria-dependent apoptosis, including Hspd1 [169], Bnip3l[170], and cytochrome c oxidase genes Cox7a2 and Cox6c have among the lowest *MYB* expression signature values among genes within this category, suggesting an unknown role for Myb as a mitochondria-dependent apoptosis regulator. Furthermore, we found locus specific GO categories such as "T cell differentiation" for RASGRP1, consistent with the role of Rasgrp1 in T cell signaling [171], and "microtubule-based process" for RRAS2, supporting its role in the regulation of cytoskeleton and its participation in related function such as adhesion, migration, and invasion [172-174].

Our functional analysis revealed that, by and large, the insertion loci fall into two major groups, with opposite characteristics and opposite effects on expression across multiple pathways: *MYCN*-like (e.g., *MYCN*, *RRAS2* and *MED20*/*CCND3*) and *p19*^{ARF}-like (e.g., *p19*^{ARF}, *p53* and *GFI1*). This observation implies that insertions either help or oppose each other in enhancing or abrogating tumorigenesis. The downstream consequences of loss of the tumor suppressor gene *p53* or *p19*^{ARF} are worsened by insertions near loci such as *MYCN*, *RRAS2* and *MED20*/*CCND3*, but mitigated by insertion near *GFI1*. Indeed, we found a strong negative correlation between insertion near *MYCN* and mRNA expression of the *MYC* gene (t-value=19.94, p-value=8.9×10-24), whose enhanced protein expression contributes to almost every aspect of tumor cell biology [175].

Using the functional map in **Figure 5.3**, we can predict whether the effect of a particular insertion effect is tumor-suppressive or oncogenic. This is especially useful for loci without any apparent *cis*-association at the mRNA level. For example, it is unclear whether insertion near *MYB*, *RUNX1*, and *MED20*/*CCND3* up-regulates or down-regulates the associated proximal gene. We found that the RUNX1 and *MED20*/*CCND3* signatures were similar to that of *MYCN*. This confirms that insertion near *RUNX1* up-regulates Runx1 mRNA expression (**Figure 5.2A**), which may act to abrogate the tumorigenesis; the reduced expression of Ccnd2 (**Table 5.1**) and induced expression of Ccnd3 (**Figure 5.2A**) predicted by LESA is consistent with a previous observation that over-expression of Ccnd3 results in reduced tumor development and strong reduction in Ccnd2 [159]. Conversely, the *MYB* signature is similar to that of p19^{ARF}, suggesting a role in tumorigenesis. Myb protein activity, while subtle at the mRNA level (**Figure 5.2A**), may be enhanced by the retroviral insertion. Indeed, *MYB* has been found to be oncogenically altered in human leukemia [176].

5.3.5 Detecting changes in protein-level TF activity associated with mutations.

Having surveyed the gene function landscape associated with each insertion locus based on its genomewide expression signature, we next wanted to identify the specific trans-acting regulatory mechanisms underlying their influence on the formation of tumors. We previously demonstrated that when the binding specificity of a TF is known (in the form of a weight matrix), quantitative changes in its regulatory activity can be inferred by performing genome-wide linear regression of a single genomewide differential mRNA expression profile on the predicted total in vitro binding affinity of cis-regulatory regions

[115], the regression coefficients corresponding to changes in TF activity. Here, we used a compendium of position weight matrices (PWMs) for 130 vertebrate TFs from the JASPAR database [177], and 11 family-level PWMs for well-characterized structural TF families [178]. We converted each PWM to an approximate position specific affinity matrix (PSAM) by assuming base frequencies to be proportional to relative binding affinities at each position within the binding site [114].

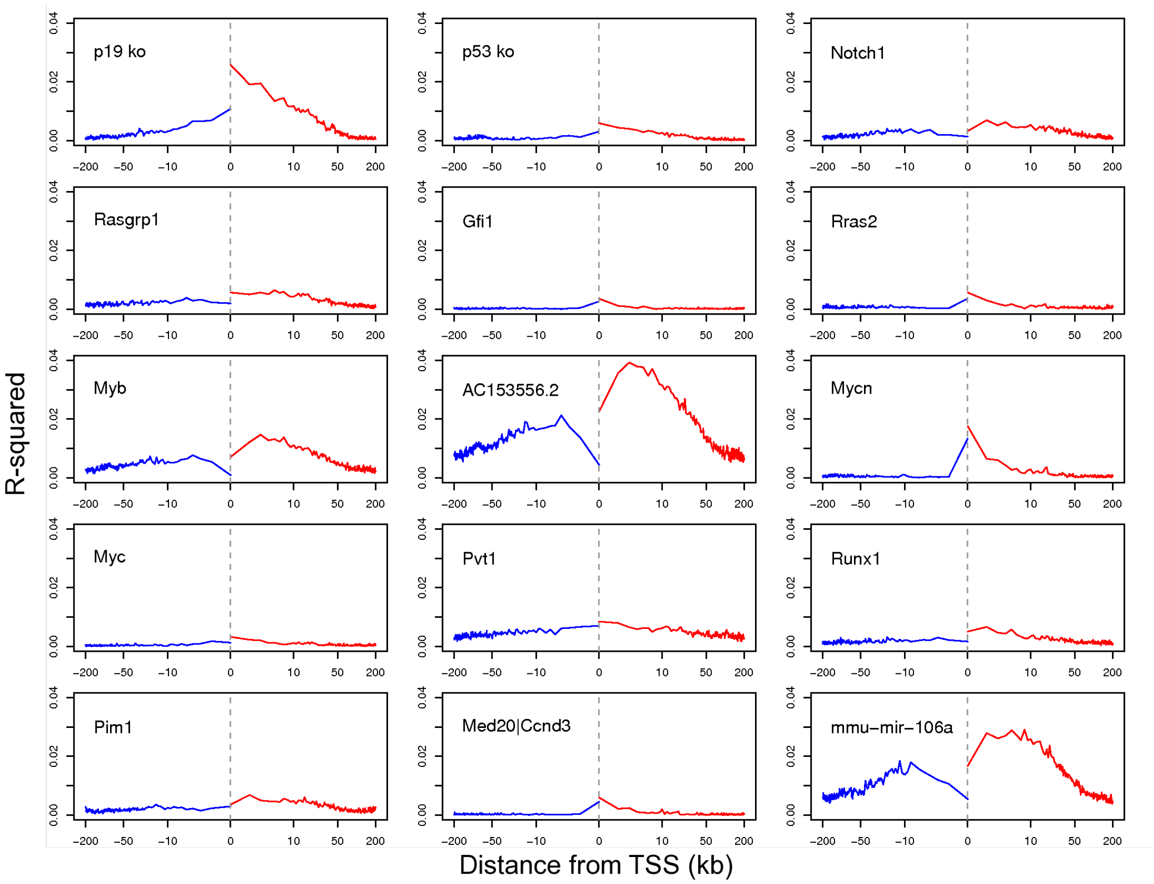

Figure 5.4: Effect of DNA base composition on LES. R^2 from a linear regression model between each LES and the frequency of DNA base composition including each DNA composition and CpG composition. Red represents downstream and blue represents upstream from transcription start sites (TSS). The DNA base composition for each gene and each window is calculated based on 1kb window sequence from TSS, then used as independent variables in linear model with LES to calculated R^2.

Using a sequence window from 200kb upstream to 200kb downstream of transcription start site (TSS) as the potential cis-regulatory region, we assumed the contribution to the transcription rate by any particular binding site to be proportional both to its binding affinity and to a positional weight decaying exponentially with the distance from the transcriptional start site (TSS); the corresponding length scale parameter was inferred from the data (see **Materials and Methods** for details). We found a relatively strong correlation between the locus expression signature value for each gene and the percentage of A, C, G, and T in its regulatory region (**Figure 5.4**). To avoid confounding due to these low-complexity signals, we inferred TF activities from the residuals of a linear regression of the signature on base composition (see **Materials and Methods**).

5.3.6 A map of cancer-related TFs specifically regulated by viral insertions.

Despite the fact that using a weight matrix to represent the target specificity of a transcription factor has some limitations, especially when it comes to distinguishing between TFs with closely similar sequence specificity, our method allowed us to make a number of interesting observations. Controlling for multiple testing – using stringent p-value thresholds of 1.0×10^{-6} and 7.9×10^{-10} for familial and individual TF-locus associations, respectively, corresponding to a false discovery rate of 0.1% – we identified a total of 22 TF-locus associations (**Figure 5.5B)**. These include known relationships, such as that of insertions at the p19$^{\text{ARF}}$ and *MYCN* loci activating REL family members NFKB1, NF-κB, RELA, and REL (p-value=9.2×10^{-13} and 2.5×10^{-11}, respectively), which are known to promote the oncogenic phenotype such as angiogenesis, proliferation and invasion/metagenesis [179]; Mycn is also known to suppress the mRNA expression level of p50 subunit of NF-κB [180]. Our algorithm also detected that one or more members of the Trp family of transcription factors (which includes Myb) are responsible for the transcriptional response to insertion at the MYB locus. Furthermore, the activity of the basic helix-loop-helix zipper (bHLH-zip) family is significantly affected by the loss of p53 as well as viral insertion near the *MYC*, *MYCN*, or *PVT1* locus (**Figure 5.5B**). Some TFs are modulated by multiple mutations, such as E2F1, which mediates the transcriptional response to both the $p19^{ARF}$ background and to insertion at the *CCND3* locus (p-value = 1.8×10^{-18} and 1.2×10^{-10}, respectively), consistent with the previous observation that human ARF binds to E2F1 to inhibit its transcriptional activity [181]. Other TFs only respond to insertion at one of the loci in our panel, such as the factor Pou5f1 to the *NOTCH1* locus. Furthermore, an association found between the Myc locus and the factor SP1 supports their known cooperative interaction to activate transcriptions [182, 183].

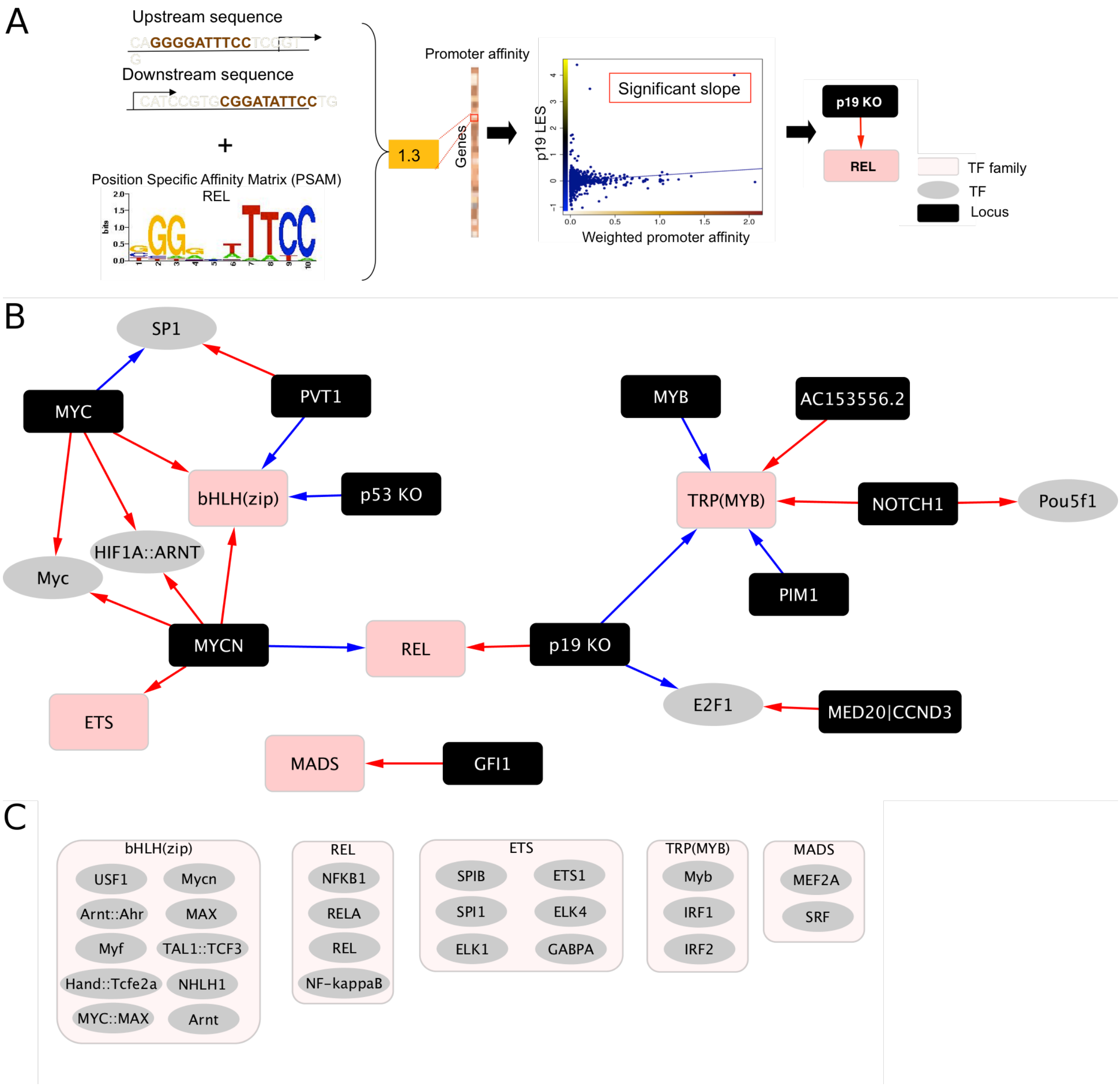

Figure 5.5: TF-locus associations. (**A**) Detecting TF-locus associations (**B**) Overview of significant familial and individual TF-locus associations at FDR 0.1%. (**C**) TFs that is included in each TF family.

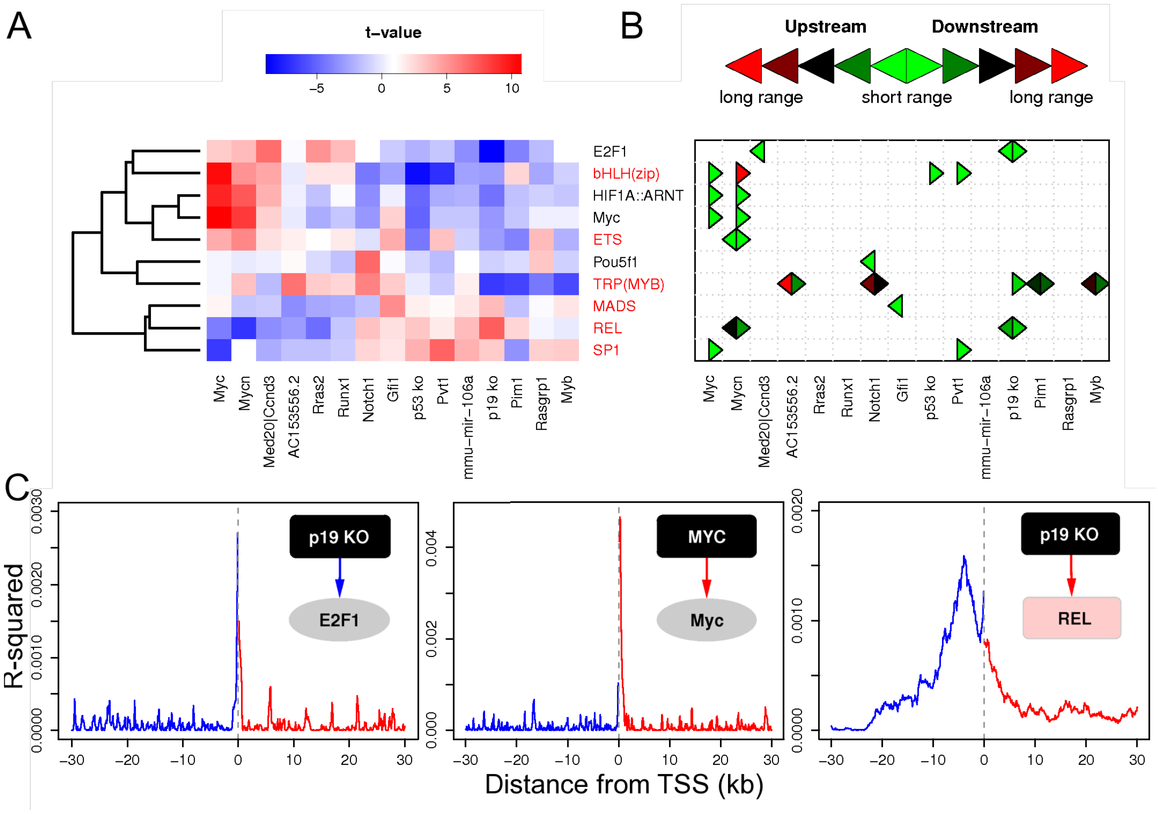

Figure 5.6: Spatial style of regulation (**A**) Heatmap of t-value from a linear model between TF affinity and LES (**B**) Optimal parameter selected for each upstream and downstream sequence, showing that both the distance (proximal vs. distal) and direction (upstream vs. downstream) of regulation differs greatly among between TFs (**C**) Examples of regulation model: proximal upstream as well as downstream regulation, downstream dependent regulation, and distal regulation.

5.3.7 TFs differ dramatically in their spatial style of regulation.

We have shown above that using prior information about the network of interactions between TFs and their target genes (in the form of cis-regulatory sequence and weight matrices) can reveal the contributions of each common insertion locus to the differential activity of specific TFs or TF families (**Figure 5.6A**). As part of this procedure we (exponentially) weighted the strength of individual binding sites surrounding the transcription start site (TSS) in such a way that those closer to the TSS contributed more heavily. Two parameters corresponding to the distance over which the contribution decreased to half the maximum value, one for sites upstream of the TSS and one for downstream sites, were fit to the data (see details in Materials and Methods). We were interested in exploring the detailed information provided by these parameters about the range over which cis-regulatory logic acts, as there is ongoing controversy about the size and orientation of the window around the TSS that is optimal to use when looking for TF binding sites.

Surprisingly, both the distance (proximal vs. distal) and direction (upstream vs. downstream) of regulation differs greatly among between TFs, in a way that is largely, but not entirely, dependent on the identity of the locus driving the TF activity change (**Figure 5.6B**). For example, the *p19^{ARF}* expression signature is best explained in terms of E2F1 binding sites when proximal occurrences upstream as well as downstream of the TSS are counted as functional (**Figure 5.6C**). By contrast, we uncover a striking asymmetry in the contribution from REL family binding sites, with only predicted binding sites upstream of the TSS (both proximal and distal) being useful for predicting the genomewide differences in expression due to the *p19^{ARF}* background (**Figure 5.6C**).

5.3.8 Myc as a regulator of transcriptional elongation.

Of particular interest among our findings was that the association between predicted binding affinity for Myc and the expression response to insertions at the *MYC* locus is limited to the proximal, downstream cis-regulatory region, with only binding sites within ~300 bp downstream of the TSS contributing, and not binding sites upstream, not even close to the TSS (**Figure 5.6C**). A plausible mechanistic explanation for this observation is that Myc acts as a transcriptional elongation control factor. Indeed, it was recently found that Myc can bind to positive transcription elongation factor b (P-TEFb) and thereby stimulate elongation at specific genes in tumor cells [184]. Since there is no significant correlation between Myc mRNA expression level and the presence of an insertion at the *MYC* locus, insertions near *MYC* are likely to act post-transcriptionally, possibly through disruption of a protein domain mediating the interaction with P-TEFb.

According to our analysis, viral insertions near *MYC* change the protein-level of the Myc transcription factor, which in turns modulates genomewide mRNA expression level, but only via binding downstream of, and proximal to, the TSS. To validate our findings, we analyzed gene expression profiles obtained after Myc inactivation using doxycycline treatment [185]. Analyzing the genomewide profile of differential mRNA expression between Myc-inactivated and mock-treated cells in exactly the same way as the *MYC* locus expression signature (**Figure 5.7A**), we found significant differential activity both for bHLH factors at the family level and for the bHLH factor Myc at the single-TF level. Since again our analysis indicated that the response to Myc inactivation is caused by Myc binding sites downstream of the TSS (**Figure 5.7B**), we conclude that it points to an intrinsic aspect of how Myc interacts with the transcriptional machinery in the cell.

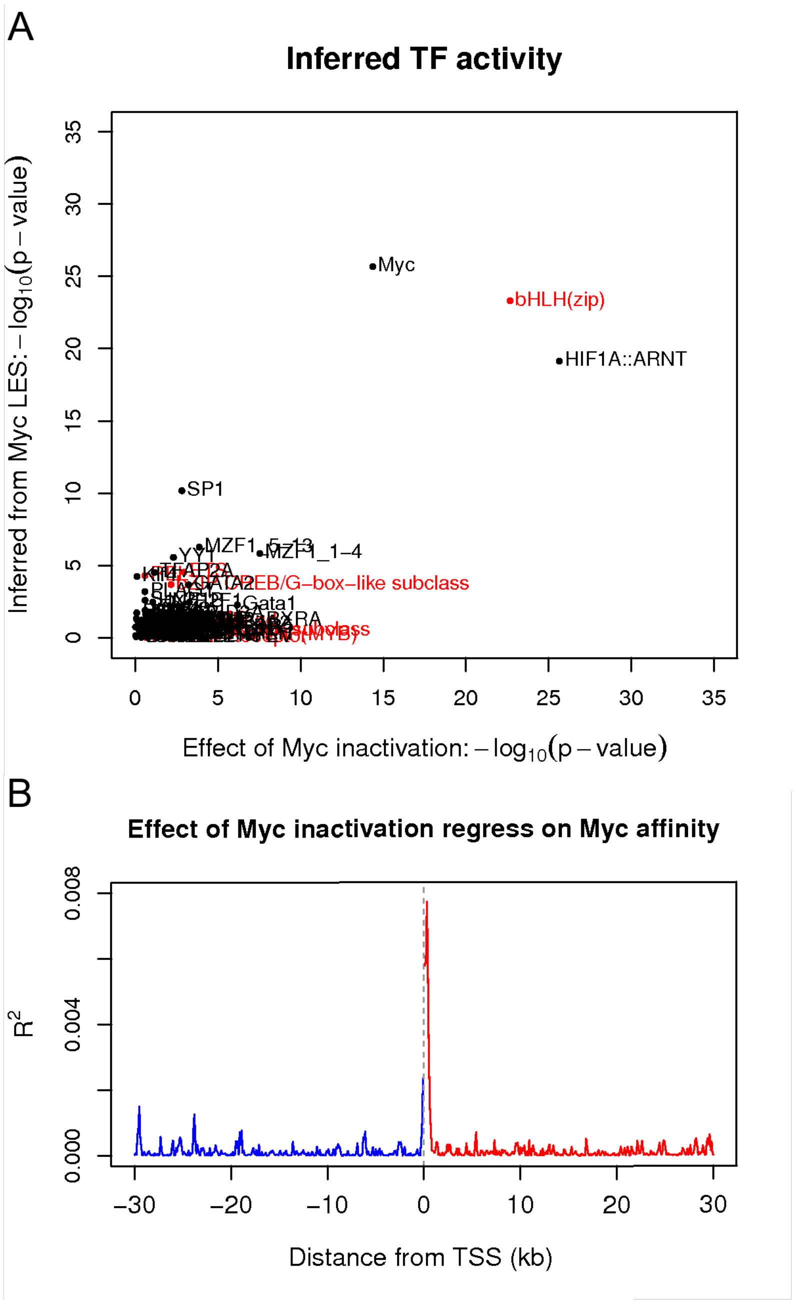

Figure 5.7: Validation of regulation under MYC expression signature. (**A**) Scatter plot of –log(p-value) for each TF. The significance levels of differential TF activities in response to MYC inactivation (x-axis) and in response to MYC insertion (y-axis) are shown. (**B**) R^2 plot from a linear model between each window affinity of Myc and the differential gene expressions in response to Myc inactivation, showing detected differential gene expressions in response to Myc inactivation are also mainly determined through binding of the Myc transcription factor on promoter-proximal downstream of its target genes.

5.3.9 Pvt1 regulates a multi-gene chromosomal domain on chromosome 15.

So far, we have analyzed TF-mediated changes in mRNA expression in a manner that is oblivious to how genes are distributed over the chromosomes. However, regional regulation of multiple neighboring genes through locus control regions has been documented [186, 187], and the rich array of locus-locus interactions observed in a typical mammalian cell [188] suggests that local modulation subnuclear organization might provide a mechanisms for coordinately controlling adjacent genes. We reasoned that are our locus expression signatures provide a natural opportunity to investigate this. Therefore, we adopted a "domainogram" approach from a previous study [189] to detect chromosomal domains within which the distribution of a given locus expression signature was non-random (**Figure 5.8**). For each window of *n* consecutive genes, we tested whether the distribution of signature values differed from the distribution of genes in the rest of the whole genome using a Wilcoxon-Mann-Whitney test. This allowed us to visualize the statistical significance of the local enrichment of signature values for all possible combinations of window position and window size (see **Materials and Methods** for details).

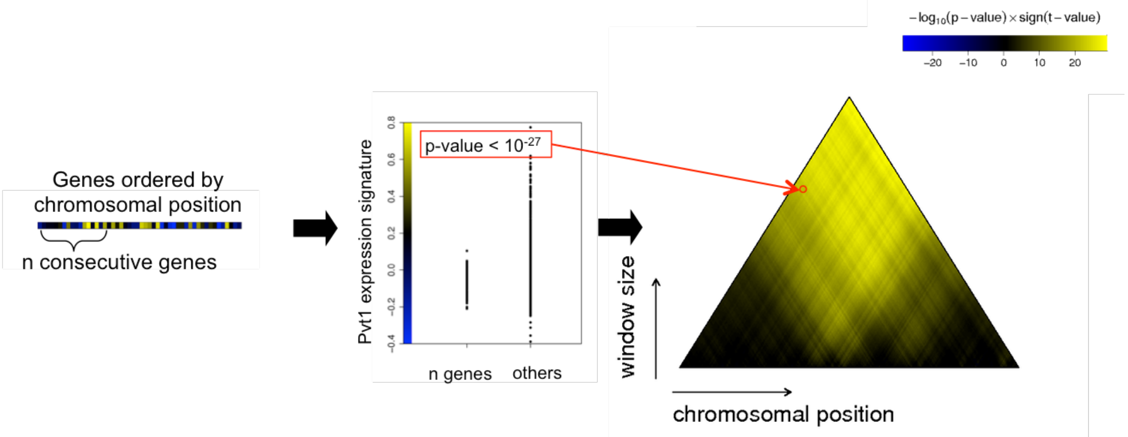

Figure 5.8: Detection of chromosomal domain organization. For each LES and each window of *n* consecutive genes, we tested whether the distribution of LES values differs from the distribution of genes in the whole genome using Wilcoxon-Mann-Whitney test, resulting in domainogram. It shows domainogram of *PVT1* expression signature on chromosome 15.

For each insertion locus, we generated domainograms across all chromosomes. Several insertion loci show domain behavior. For the PVT1 locus, this identified a region on chromosome 15 within which genes are collectively induced (**Figure 5.8)**. Significantly, the *PVT1* locus is also located on the chromosome 15, suggesting a locus-control-region type of regulatory mechanism. We observed a significant bias in DNA base composition and CpG content on chromosome 15 (results are not shown). However, these low-complexity features were not found to be predictive of the PVT1 signature (p-value > 0.1). Moreover, the *PVT1* domain structure remained unchanged after we explicitly removed low-complexity signals (results are not shown). We conclude that the domain structure of the transcriptional response to viral insertion at PVT1 is not driven by low-complexity signals per se, and therefore may result from a local reorganization of the nucleus.

5.3.10 Relating drug response to locus expression signatures.

There is a close analogy between how a viral insertion and a drug affect the cell state at various levels, including that of the transcriptome. In fact, our study could be viewed as an attempt to uncover the “mode-of-action" of viral insertion at a specific locus. To systematically search for drugs that might move the cell state in a direction opposite to the change in cell state caused by to a specific mutation, we analyzed data from the Connectivity Map, encompassing 7,000 genomewide expression responses to 1309 different compounds [190]. We performed genome-wide linear regression between locus expression signatures and drug response profiles to identify drug-locus associations and predict the drug response of tumors with specific viral insertions (**Figure 5.9**).

We found a total of 281 significant drug-locus relationships at a false discovery rate of 1% (**Figure 5.10A**). A negative correlation between locus signature and drug response indicates that the drug might mitigate the effect of the insertion on tumorigenesis; a positive correlation indicates that the drug might exacerbate it. We found that 10 out of 15 loci were significantly associated with at least one significant candidate therapeutic drug. Interestingly, even though the overall patterns of drug association are similar for *MYC* and *MYCN*, these loci each had specific drug associations. The *MYC* signature has significant negative correlation with the response to PI3K inhibitors (i.e. wortmannin and LY-294002), while the response to histone deacetylases (i.e. valproic acid, trichostatin A, vorinostat, MS-275 and HC toxin) significantly antagonizes the MYCN signature (**Figure 5.9A**). Furthermore, our analysis predicts tumors with insertions at mmu-mir-106a to be responsive to topoisomerase inhibitors (i.e. camptothecin, mitoxantrone, doxorubicin, daunorubicin and irinotecan), a well-known class of anti-cancer drugs. Taken together, our results suggest a strategy for predicting drug responsiveness from knowledge of the genotype of an individual tumor. We found most drug-locus associations to be independent of the cell-type in which the drug response was measured. Some notable exceptions to this trend are shown in **Figure 5.10B**. In HL60 cells, the response to genistein, which acts as a tyrosine kinase inhibitor and was identified as an angiogenesis inhibitor [191], shows a significant negative correlation with both the *MYCN* and *MYC* expression signatures. By contrast, the response to genistein in MCF7 cells correlates with the p19$^{ARF-/-}$ signature.

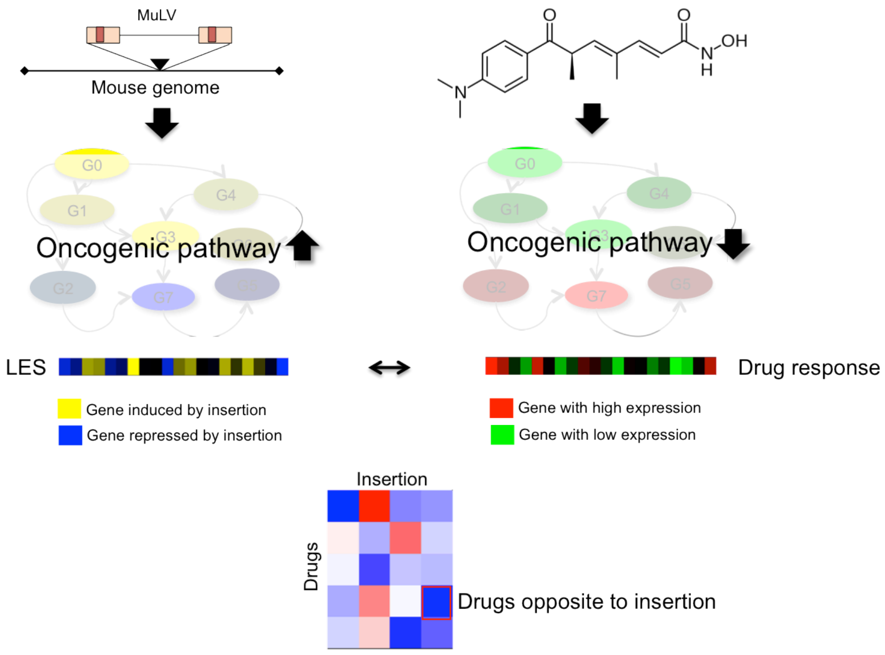

Figure 5.9: Overview of our strategy for relating drug response to locus expression signatures. There is a close analogy between how a viral insertion and a drug affect the cell state at the level of transcriptome. We perform genome-wide linear regression between locus expression signatures and drug response profiles to identify drug-locus associations. The negative association represents cases where the effect of drugs is opposite to that of insertion, and therefore may have a therapeutic effect.

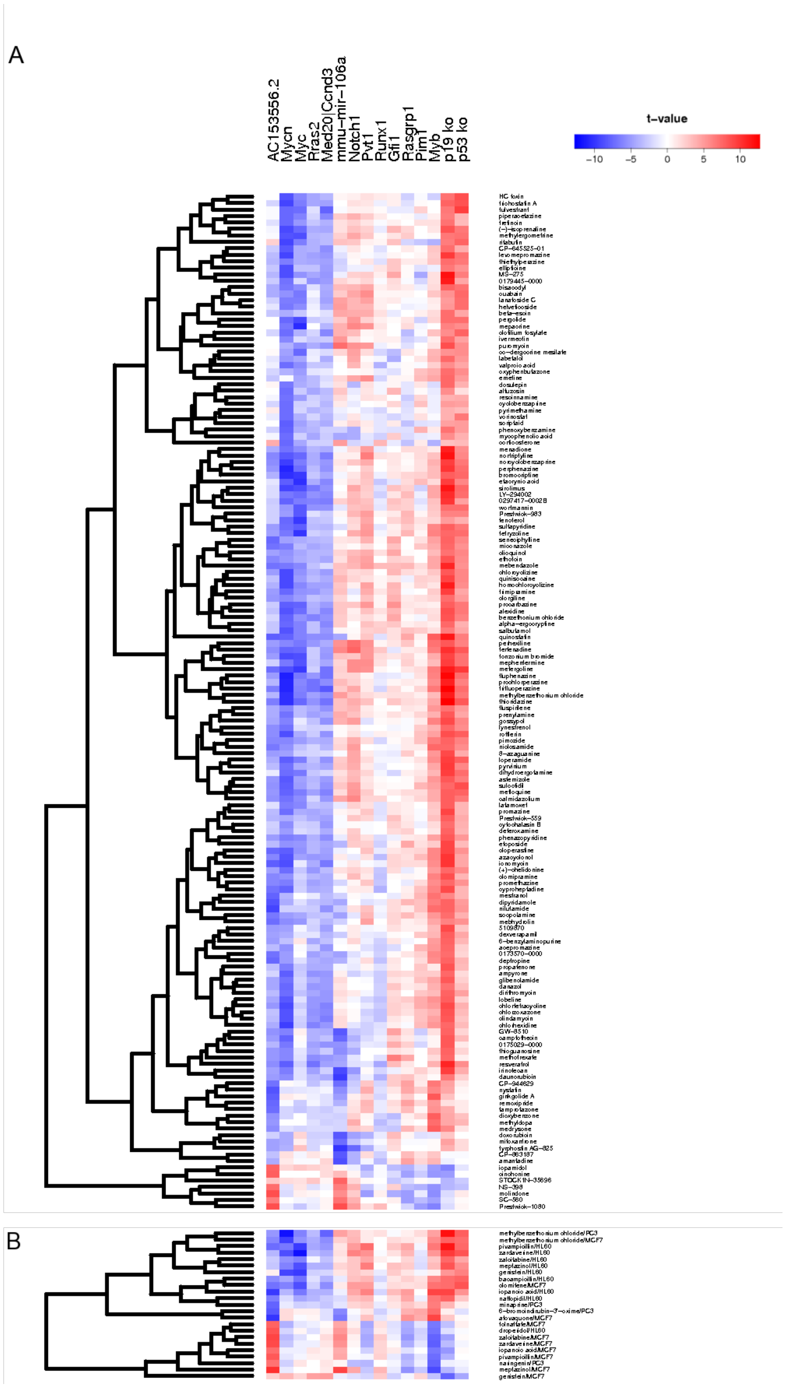

Figure 5.10: Heatmap of t-value showing drug-locus associations. (**A**) Significant Drug-locus associations at FDR 1%. It is based on expression profiles averaged over the cell types in response to the same drug (**B**) Cell type specific drug-locus associations

5.4 Discussion

We have demonstrated here that using parallel genotyping data and genomewide mRNA expression data from mouse tumors generated by retroviral insertional mutagenesis greatly increases our ability to identify the regulatory mechanisms underlying tumorigenesis and the biological pathways affected by them. We use a mechanistic framework that allows us to attribute function to causal mechanisms in a way that would not be possible using retroviral insertional mutagenesis data alone. The information provided by the insertion loci and the gene expression data is highly complementary and synergistic. On the one hand, the insertional mutagenesis screen benefits from the selective pressure that is present during tumor growth. On the other hand, the genome-wide expression profiles provide a high-dimensional downstream readout of the cell state.

The locus expression signature analysis (LESA) approach presented in this paper allows us to exploit the genomewide transcriptional response to a change in the structure and/or expression level of the causal gene near the insertion locus, even if the latter is initially unknown. First, by integrating the locus expression signature with gene function annotation, we could map the influence of insertion at each common locus on various biological processes relevant to tumorigenesis. Second, systematic application of LESA uncovered a large number of connections between common insertion loci and their downstream gene regulatory mechanisms – some of which were known, and many of which were novel.

In the first place, our results include many locus-TF associations, in which the regulatory activity of a particular TFs or TF family is modulated by viral insertion at a particular locus. By construction, these associations are causal as well as functional. It should be noted that our method fully accounts for post-translational regulation of TF activity at the protein level. For example, we found Myc activity at the protein level to change in response insertions near the *MYC* locus, even though the mRNA expression level of the *MYC* gene does not change. Perhaps more surprisingly, in a different type of analysis, application of LESA helped reveal changes in chromosomal domain organization affected by viral insertions. For instance, we found that viral insertion near the *PVT1* gene affects a large number of neighboring genes also located on chromosome 15. This suggests viral insertion near *PVT1* could locally change nuclear organization. This observation could be the consequence of genomic instability, which is a characteristic of almost all human cancers [192]. However, it is unlikely because our expression signature represents a consensus across many tumors. Furthermore, by relating our locus signatures directly to drug response profiles from the Connectivity Map, we were able to identify drugs that might specifically counteract the deregulatory effect of mutations at particular loci. A previous study [193] performed a large-scale integration of expression signatures of human disease from public data with drug response profiles from Connectivity map, and confirmed the usefulness of drug responses profiles to predict novel therapeutic indication.

The mammalian genome is a complex landscape: some regions are very gene-rich, whereas others are devoid of genes [194-196]. This variation in gene density is linked to large-scale variation in DNA base composition [195, 196]. GC content in turn is positively correlated with expression level [197]. To capture any such effects of low sequence complexity, we considered DNA base composition as well as CpG content. We detected a high correlation between sequence composition and signature value for several loci including *MYB*, *AC153556.2*, and mmu-mir-106a~363. For most loci, the region downstream of the transcription start site (TSS) showed higher correlations than that upstream of the TSS. At the same time, loci such as *MYCN*, *MYC* and *MED20*/*CCND3* do not show any correlation with low-complexity sequence signals (**Figure 5.4**). The mechanism underlying these observations is not clear, and it would be interesting to further investigate them. To avoid confounding in our TF-centric analysis, we removed the low-complexity sequence effect in our analysis.

Many of functional regulatory sites occur outside the proximal promoter for mammalian [198], and distal binding events up to 1Mb away from the TSS have been shown to contribute to p300 occupancy [199]. How the contribution of binding sites to transcriptional control depends on its proximity to the TSS has been previously quantified [200]. We considered 200kb upstream and downstream sequence from transcription start site (TSS) as the cis-regulatory region. However, to allow for variable relationships between binding positions and transcriptional control, we employed TF-specific length scale parameters and fit these by maximizing the correlation between the weighted TF affinity profiling and the locus expression signature. Surprisingly, we found that for various TFs including Myc, the transcriptional response to insertion at the locus is almost exclusively mediated by binding events downstream of the TSS, implying a role in transcriptional elongation or post-transcriptional regulation. Taken together, our analyses have yielded a more comprehensive and detailed view of the oncogenic signaling networks and cis-regulatory mechanisms underlying tumorigenesis.

5.5 Materials and Methods

5.5.1 Retroviral Insertional mutation data

Retroviral insertional mutagenesis screens have been performed previously [61]. Mice were infected with Murine leukemia virus (MuLV) at postnatal day 1 and monitored for tumor growth. MuLV infection accelerated lymphomagenesis in these mice. Mice developed tumors almost exclusively in spleen, thymus and lymph nodes. The gene expression levels of a subset of 97 retrovirally induced splenic lymphomas in p19$^{ARF-/-}$ (n=31), p53$^{-/-}$ (n=19) and wild-type (n=53) mice were measured and analyzed.

5.5.2 Gene expression data

Gene expression data were collected using Illumina MouseWG6-V2 beadchips, and normalized using variance-stabilizing transformation (VST) and robust spline normalization (RSN) [201]. Illumina probes with no corresponding RefSeq ID were discarded, leaving 45,281 measurements. Averaging over probes mapping to the same RefSeq ID resulted in expression values for 19,010 genes.

5.5.3 Common insertion loci

The effect of viral insertions on their nearby targets is dependent on the relative position and orientation of the target transcript as well as the orientation of the viral integration. To exploit this context information, we employed a rule-based mapping (RBM) procedure [68]. RBM assigns each insertion to one or more putative target transcripts based on a set of rules that were distilled from literature. The unique list of transcripts that results from this procedure is used to generate a binary profile that, for each tumor, indicates if a transcript is a putative target or not. We observed that for proximal transcripts the same binary profile frequently results. These were therefore treated as a single profile. Only those transcript-insertion associations found in at least three tumors were considered in our analysis.

5.5.4 Genome sequence and gene annotation

We obtained mouse genome sequence from UCSC via the *BSgenome.Mmusculus.UCSC.mm9* package in *BioConductor* [150]. We downloaded the corresponding genome annotation coordinates directly from *genome.ucsc.edu* (version mm9).

5.5.5 Locus expression signature analysis (LESA)

The variation in relative mRNA expression level across all tumors (represented as the log2-ratio relative to the mean across all tumors) was analyzed independently for each gene. To obtain locus expression signature values across all loci for a given gene, we performed analysis of variance (ANOVA) for the mRNA expression in terms of background genotype and insertion status at all loci considered, by fitting the following multivariate linear model:

$$A_{gt} = \beta_{g,p19^{ARF}} B_{t,p19^{ARF}} + \beta_{g,p53} B_{t,p53} + \sum_{m \in M} \beta_{gm} I_{mt}$$

Here, $A_{g,\mathrm{t}}$ represents the relative mRNA expression level for tumor t and gene g. $B_{t,p19^{ARF}}$ and $B_{t,p53}$ (taking values 0 or 1) indicate whether the genetic background of tumor t is $\text{p19}^{\text{ARF-/-}}$ and $\text{p53}^{\text{-/-}}$, respectively, while I_{mt} indicates whether an insertion was present at locus m in tumor t. After this analysis had been completed for all genes, the expression signature for each background and for each viral insertion locus was constructed by combining the values of the pertinent regression coefficient β as a vector across all genes.

5.5.6 Information content of locus expression signatures

To assess how much information about downstream transcriptional regulation was contained in a given signature, without the need to specify a particular regulatory mechanism, we summed the squares of the t-values t_{gm} corresponding to the regression coefficients β_{gm}:

$$\chi_m^2 = \sum_g t_{gm}^2$$

To determine the statistical significance of the χ^2-statistic, we constructed a null distribution by performing 100 independent random permutations of all tumors, in a way that preserved the correlation structure between insertion loci.

5.5.7 Forward selection of GO categories

Functional annotation of genes in terms of Gene Ontology (GO) categories was performed using a variation of the T-profiler algorithm [149], including an iterative procedure to select a non-redundant set of gene sets from those that have a significantly different expression distribution from the other genes. For each GO category, we applied the Wilcoxon-Mann-Whitney (WMW) test to detect differences in distribution between the locus expression signature value of genes within the GO category and that of the other genes. At each step, we subtracted the mean signature value of the genes in the gene set with the lowest p-value from all genes in that gene set. The p-values were then recalculated and the procedure repeated until even the most significantly regulated gene group had a p-value > 10^{-5}, which corresponds to a false discovery rate (FDR) <0.1%. Statistical significance was determined by performing independent random permutation of the signature values for each gene. The FDR corresponding to a given p-value threshold was computed as the ratio of the number of GO categories with a p-value below threshold, averaged over 50 randomized data sets, and the number of GO categories with p-value below threshold. A 1% FDR based on the empirical permutation test corresponds to a WMW test p-value <10^{-4}.

5.5.8 Low-complexity sequence features

To eliminate the potential confounding contribution from low-complexity sequence features to locus expression signatures (LES), we calculated the frequency of each base and the CpG dinucleotide across the transcribed region for each gene. Next, we computed the residuals from a multiple linear regression of each LES on these five frequencies (without an intercept). We used these in further TF-locus association analyses.

5.5.9 Weight matrices

TFs from the same structural family tend to bind to similar DNA target sequences [202]. Our analysis explicitly recognizes that it is often not possible to implicate a specific TF based on its binding specificity alone. The JASPAR database [203] contains models at the level of structural families with the shared DNA binding structures[178]. We downloaded 11 PWMs representing a distinct TF family from the JASPAR FAM database as well as 130 PWMs that are supposed to represent individual TFs from JASPAR CORE database.

5.5.10 TF binding affinity

We used the convert2psam utility from REDUCE Suite version 2.0 software package (bussemakerlab.org) to convert each PWM from JASPAR to a position-specific affinity matrix or PSAM [114]; pseudo-counts equal to one were added to the PWM at each position, and the resulting base counts were divided by that of the most frequent base at each position to get an estimate for the relative affinity associated with each point mutation away from the optimal binding sequence. The resulting PSAM collection was used to compute a weighted promoter affinity for each gene. All putative individual binding sites in the genomic region from 200kb upstream to 200kb downstream of the TSS of each gene with a predicted relative affinity of at least 0.1 were identified and scored using the *AffinityProfile* utility in the REDUCE Suite. To obtain a total weighted upstream affinity for a given value of the regulatory scale parameter λ, we summed the affinity of all upstream binding sites using a weight exp(-d/λ), where d is the (absolute) distance of a given binding site from the TSS. A total downstream affinity was computed in an analogous manner.

5.5.11 Mapping locus-to-TF network connectivity

To treat the attribution of a regulatory role to TFs with similar sequence specificity in a conservative manner, we applied a two-step procedure. First, we performed multivariate linear regression of low-complexity-signal-corrected LES values (see above) on the weighted binding affinities for each TF family. Next, the residuals from this fit were regression on affinity profiles for individual TF. We computed p-values using ordinary linear regression. Statistical significance was determined by performing 1000 independent random permutation of for each gene. A 0.1% false discovery rate (FDR) based on the empirical permutation test for family-level and individual PSAMs corresponded to a p-value $<1.0\times10^{-6}$ and $<7.9\times10^{-10}$, respectively.

5.5.12 Validation of *MYC* result

We downloaded gene expression profiles obtained by [185] for transgenic mice that conditionally express the human MYC cDNA in T-cell lymphocytes (GEO accession number GSE10200). In this transgenic mouse, doxycycline treatment suppresses MYC expression. The authors measured gene expression at different doxycycline concentrations, and found that a doxycycline threshold level of 0.05ng/ml was required to maintain the tumor phenotype. We only used the two most extreme doxycycline concentrations of 0ng/ml and 20ng/ml. To obtain an estimate for the differential expression level in response to inactivation of Myc, we subtracted the treatment/reference log2-ratio at 0ng/ml from that at 20ng/ml. These values served as the dependent variable in the regression on TF affinity profiles.

5.5.13 Identifying regulation at the level of multi-gene chromosomal domains

To test whether the distribution of locus expression signature values across any given number of adjacent genes differed from that of genes in the whole genome, we adapted a procedure used by [189]. Within each chromosome, we first sorted genes by their TSS. For every possible set of n adjacent genes, we performed a WMW test. The resulting p-values were visualized as triangular graph, with window position indicated on the horizontal and window size on the vertical axis.

5.5.14 Mapping drug-locus associations

Genomewide mRNA expression data for cultured human cells treated with bioactive small molecules were downloaded from the Connectivity Map website (www.broadinstitute.org/cmap/). This collection contains 7056 expression profiles for 1309 distinct compounds. The experiments were carried out on two different Affymetrix GeneChip designs (HG-U133A and HT-HG_U133A), and in four different cell lines (the breast cancer epithelial cell line MCF7, the prostate cancer epithelial cell line PC3, nonepithelial leukemica cell line HL60, and nonepithelial melanoma cell line SKMEL5). We followed the preprocessing and normalization steps described in [190] to obtain the expression log2-ratio between drug treatment and control. To combine the human drug response expression data with our mouse-based locus expression signatures, we needed to map human Affymetrix probe IDs to mouse RefSeq IDs. For this, we used human-mouse orthology tables downloaded from UCSC (genome.ucsc.edu/). We averaged over the Affymetrix probes mapping to the same mouse RefSeq ID, resulting in 9757 genes shared between both data sets.

To obtain robust results, we filtered out non-informative genes using two criteria. First, only mouse genes showing a high variance across tumors (upper 50 percentile) were retained. Second, we deleted human genes whose expression was detected in neither treatment nor control. Next, we calculated averages of gene expression levels across profiles for the same drug in different cell types, resulting in 1309 drug signatures. Multivariate linear regression of each of these on the locus expression signatures was performed. To determine the statistical significance of each putative drug-locus association, we performed 100 random permutations of drug signatures and repeated the analysis. A 1% false discovery rate (FDR) corresponded to a regression coefficients whose t-value has an absolute value >7. To explore the impact of physiological context on drug-locus associations, we calculated cell-type-specific drug responses by averaging only over replicate profiles for the same cell type, resulting 3587 cell-type-specific drug response signatures.

Chapter 6

Concluding Remarks

In this dissertation, I have presented two approaches to dissect the mechanisms underlying genetic variants in gene expression through a TF-centric point of view. The following concluding remarks address future directions for this research and potential applications of our approaches.

6.1 Hypotheses to be experimentally validated

Both studies in this dissertation led to hypotheses that may be tested experimentally, and for which some supporting evidence has come to light.

6.1.1 Genetic modulators of transcription factor activity

In Section 3.3.6, we predicted that the region on chromosome 2 contains a genetic modulator of several transcription factors including Fkh1, Fkh2, Swi5, Ace2 and Stb1 based on protein-protein interaction. The locus contains the *CDC28* gene, which encodes a cyclin-dependent kinase. In particular, we predicted that the sign of the aQTL linkage to the *CDC28* locus for Fhk2p is the opposite of that for Fkh1p (Figure 3.8A): whereas the transcriptional targets of Fkp1p are more highly expressed in segregants carrying the BY allele at the *CDC28* locus, the opposite is true for the targets of Fkh2p (Figure 3.8B). Even though no amino-acid change is detected by alignment of the BY and RM protein sequences for Cdc28p, alignment of BY and RM coding regions for *CDC28* gene reveals each point mutation in 3'UTR and 5'UTR. While the locus coincides with a previously identified eQTL hotspot and *AMN1* in this region was experimentally validated to control transcriptional rates [32], *CDC28* has not previously been identified as a *trans*-acting genetic modulator, and none of the TFs we identified except Ace2 has been previously implicated with the hotspot. Therefore, the effect of *CDC28* polymorphisms on TF activities could be tested by measurement of expression changes of allele replacement strains at *CDC28*, and one would expect the target genes of these TFs are differentially expressed in response to allele swap of *CDC28*. Additionally, we detected novel *trans*-acting polymorphisms in the TF-encoding gene *STB5*, *RFX1*, and *HAP4* in Section 3.3.5. To test this observation, one could measure the expression levels of genes in allele replacement strains at these TFs genes, and expect their target genes are affected by it.

6.1.2 Regulatory mechanisms underlying tumorigenesis

In Section 5.3, we predicted several regulatory mechanisms underlying tumorigenesis. One of our findings is the existence of a large region on chromosome 15 in which most genes are induced by insertions near *PVT1* (see Section 5.3.9). Indeed, *PVT1* is located on chromosome 15, suggesting a regional effect of viral insertion on the expression of multiple genes. Interestingly, *PVT1* whose full name is plasmacytoma variant translocation 1 is the site of reciprocal translocations to immunoglobulin loci, resulting in 'variant' translocations, T(2:8) or T(8:22) in Burkitt's lymphoma [204]. In most murine plasmacytomas, t(15:12) translocations, analogous to the T(8:14) translocations in Burkitt's lymphoma, t(6:15) translocations are observed [205, 206]. These previous studies suggest one possible mechanism by which a retroviral insertion near *PVT1* locus can induce translocation and changes chromosomal territories, resulting in induced expression levels for the large number of genes on chromosome 15. The experiment visualizing locations around *PVT1* locus by fluorescence in situ hybridization (FISH) for the tumors with *PVT1* insertion could be performed, and compared to other tumors to see whether the chromosomal domains are changed for these tumors. Furthermore, we detected the chromosomal domain on chromosome 15 in which many genes are repressed in response to *PIM1* insertion (**Figure 6.1**). Interestingly, genome-wide translocation sequencing reveals that double-strand breaks (DSBs) at c-myc oncogene on chromosome 15 were preferentially targeted some chromosomal regions, including *PIM1* on chromosome 17 [207]. Chromatin interaction profiling using Hi-C experiment [188] or FISH experiment could be performed to validate this observation.

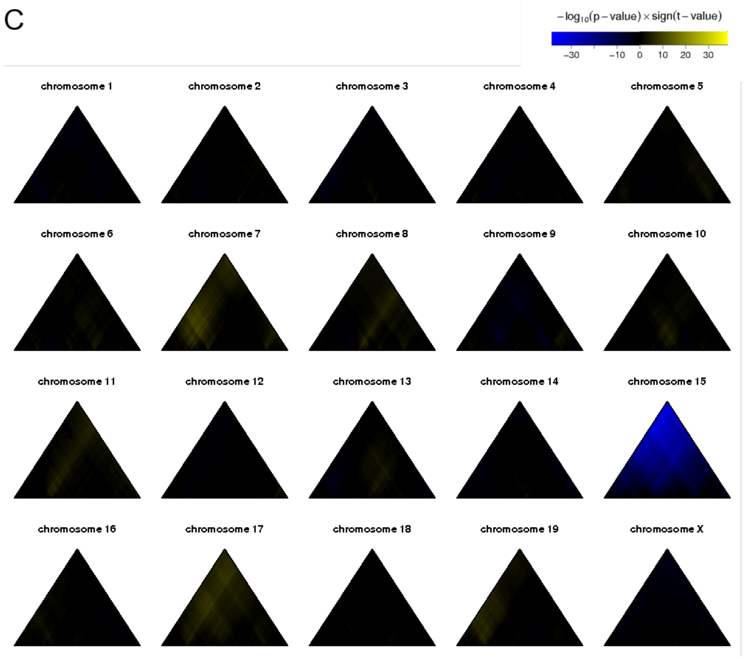

Figure 6.1: Domainogram of each chromosome for PIM1 expression signature.

In Section 5.3.11, we investigated the relationship between drug responses and locus-specific expression signatures, and detected several insertional mutation-specific drug responses. Interestingly, even though the overall patterns of drug association are similar for *MYC* and *MYCN*, these loci each had specific drug associations. The *MYC* signature has significant negative correlation with the response to PI3K inhibitors (i.e. wortmannin and LY-294002), while the response to histone deacetylases (i.e. valproic acid, trichostatin A and HC toxin) significantly antagonizes the MYCN signature. Furthermore, our analysis predicts tumors with insertions at mmu-mir-106a to be responsive to topoisomerase inhibitors (i.e. camptothecin, mitoxantrone, doxorubicin, daunorubicin and irinotecan), a well-known class of anti-cancer drugs. The response of tumor carrying these insertions to each drug could be observed to validate our observation.

6.2 Applying to other data sets

6.2.1 Application of aQTL approach

Our aQTL approach in Chapter 3 is generally applicable whenever a matrix of connection strength between regulators and targets, independent of the phenotype matrix, is available as prior information. In collaboration with other Bussemaker lab members, I have extended this approach to other biological contexts, including the promiscuous binding of TFs to high-occupancy target (HOT) regions and post-transcriptional regulatory networks (see Chapter 4). Furthermore, an obvious extension of the analysis in Chapter 3 would be to apply the approach to other organisms' eQTL datasets, for example, human. The 270 individuals in the four HapMap populations were previously genotyped [208] from International HapMap project. The gene expression measurements from genes spanning the genome of EBV-transformed lymphoblastoid cell lines were obtained from the same individuals used in the phase 1 and phase II of the project [31]. Additionally, RNA-seq data that enable the analysis of transcript variation at unprecedented resolution from unrelated Nigerian individuals are available [209]. Furthermore, recent high-resolution, and high-throughput protein binding microarray (PBM) [210, 211] array data for defining the *in vitro* sequence specificity of a substantial subset of mammalian transcription factors could be used. Application of our aQTL approach would provide insights about transcriptional regulation of human.

6.2.2 Application of LESA

Recently, large-scale projects to map chromosomal aberrations, mutations and gene expression of cancer were launched by several groups. Genomic data has been collected for thousands of tumors at high-resolution using array comparative genomic hybridization (aCGH) [40], high density single nucleotide polymorphism (SNP) microarrays [41], and massively parallel sequencing [43]. Measurements of gene expression levels in parallel with these genomic data could be used to identify drivers of tumorigenesis and connected them to many of their targets and biological functions by applying our analysis. The genomic data could be treated in a way analogous to what we did for common insertion sites data in our analysis. Then, we would calculate expression signatures modulated by each genomic mutation in cancer. The resulting expression signatures could be used for further analyses such as GO analysis, identifying differential

TF activities and chromosomal domains. This application would facilitate identification of regulatory mechanisms underlying tumorigenesis.

6.3 Integrating other layers of regulation

The human genome is a store of information. The three billion bases encode, either directly or indirectly, the instructions for synthesizing all the molecules that form each human cell, tissue and organ. The genomic elements within the DNA sequence orchestrate the development and function of a human through multiple layers of regulations, including chromatin state [1], transcriptional rate [2], splicing [3], mRNA localization [4], mRNA stability [5], translational rate [6], and protein stability [7]. The Encyclopedia of DNA Elements (ENCODE) Project [212, 213] aims to provide a more biologically informative representation of the human genome by using high-throughput methods, including data about the degree of DNA methylation and chemical modifications to histones that can influence the rate of transcription. ENCODE also examines long-range chromatin interactions, such as looping, that alter the relative proximities of different chromosomal regions in three dimensions and also affect transcription. An obvious extension of the analysis described in Chapter 3 and Chapter 5 would be the modification of algorithms to consider the additional layer of regulation using data now available from ENCODE project.

REFERENCES

1. Cairns, B.R., *The logic of chromatin architecture and remodelling at promoters.* Nature, 2009. **461**(7261): p. 193-8.

2. Lee, T.I. and R.A. Young, *Transcription of eukaryotic protein-coding genes.* Annual review of genetics, 2000. **34**: p. 77-137.

3. Licatalosi, D.D. and R.B. Darnell, *RNA processing and its regulation: global insights into biological networks.* Nature reviews. Genetics, 2010. **11**(1): p. 75-87.

4. Martin, K.C. and A. Ephrussi, *mRNA localization: gene expression in the spatial dimension.* Cell, 2009. **136**(4): p. 719-30.

5. Garneau, N.L., J. Wilusz, and C.J. Wilusz, *The highways and byways of mRNA decay.* Nature reviews. Molecular cell biology, 2007. **8**(2): p. 113-26.

6. Gebauer, F. and M.W. Hentze, *Molecular mechanisms of translational control.* Nature reviews. Molecular cell biology, 2004. **5**(10): p. 827-35.

7. Ciechanover, A. and A.L. Schwartz, *The ubiquitin-proteasome pathway: the complexity and myriad functions of proteins death.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(6): p. 2727-30.

8. Maskos, U. and E.M. Southern, *Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ.* Nucleic acids research, 1992. **20**(7): p. 1679-84.

9. DeRisi, J.L., V.R. Iyer, and P.O. Brown, *Exploring the metabolic and genetic control of gene expression on a genomic scale.* Science, 1997. **278**(5338): p. 680-6.

10. Hughes, T.R., et al., *Functional discovery via a compendium of expression profiles.* Cell, 2000. **102**(1): p. 109-26.

11. Chee, M., et al., *Accessing genetic information with high-density DNA arrays.* Science, 1996. **274**(5287): p. 610-4.

12. Wang, D.G., et al., *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.* Science, 1998. **280**(5366): p. 1077-82.

13. Ren, B., et al., *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.

14. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews. Genetics, 2009. **10**(1): p. 57-63.

15. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nature methods, 2007. **4**(8): p. 651-7.

16. Jansen, R.C. and J.P. Nap, *Genetical genomics: the added value from segregation.* Trends Genet, 2001. **17**(7): p. 388-91.

17. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* Journal of molecular biology, 1961. **3**: p. 318-56.

18. Schwartz, D., *Genetic Studies on Mutant Enzymes in Maize. III. Control of Gene Action in the Synthesis of Ph 7.5 Esterase.* Genetics, 1962. **47**(11): p. 1609-15.

19. Brem, R.B., et al., *Genetic dissection of transcriptional regulation in budding yeast.* Science, 2002. **296**(5568): p. 752-5.

20. Cheung, V.G., et al., *Natural variation in human gene expression assessed in lymphoblastoid cells.* Nat Genet, 2003. **33**(3): p. 422-5.

21. Schadt, E.E., et al., *Genetics of gene expression surveyed in maize, mouse and man.* Nature, 2003. **422**(6929): p. 297-302.

22. DeCook, R., et al., *Genetic regulation of gene expression during shoot development in Arabidopsis.* Genetics, 2006. **172**(2): p. 1155-64.

23. Li, J. and M. Burmeister, *Genetical genomics: combining genetics with gene expression analysis.* Human molecular genetics, 2005. **14 Spec No. 2**: p. R163-9.

24. Jansen, R.C., *Studying complex biological systems using multifactorial perturbation.* Nature reviews. Genetics, 2003. **4**(2): p. 145-51.

25. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(16): p. 9440-5.

26. Brem, R.B. and L. Kruglyak, *The landscape of genetic complexity across 5,700 gene expression traits in yeast.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(5): p. 1572-7.

27. Rockman, M.V. and L. Kruglyak, *Genetics of global gene expression.* Nat Rev Genet, 2006. **7**(11): p. 862-72.

28. Ronald, J., et al., *Local regulatory variation in Saccharomyces cerevisiae.* PLoS Genet, 2005. **1**(2): p. e25.

29. Myers, A.J., et al., *A survey of genetic human cortical gene expression.* Nature genetics, 2007. **39**(12): p. 1494-9.

30. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-8.

31. Stranger, B.E., et al., *Population genomics of human gene expression.* Nature genetics, 2007. **39**(10): p. 1217-24.

32. Yvert, G., et al., *Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors.* Nat Genet, 2003. **35**(1): p. 57-64.

33. Morley, M., et al., *Genetic analysis of genome-wide variation in human gene expression.* Nature, 2004. **430**(7001): p. 743-7.

34. Zhu, J., et al., *Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.* Nat Genet, 2008. **40**(7): p. 854-61.

35. Lee, S.I., et al., *Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification.* Proc Natl Acad Sci U S A, 2006. **103**(38): p. 14062-7.

36. Biswas, S., J.D. Storey, and J.M. Akey, *Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis.* BMC Bioinformatics, 2008. **9**: p. 244.

37. Sun, W., T. Yu, and K.C. Li, *Detection of eQTL modules mediated by activity levels of transcription factors.* Bioinformatics, 2007. **23**(17): p. 2290-7.

38. Ye, C., et al., *Using network component analysis to dissect regulatory networks mediated by transcription factors in yeast.* PLoS Comput Biol, 2009. **5**(3): p. e1000311.

39. Kliebenstein, D.J., et al., *Identification of QTLs controlling gene expression networks defined a priori.* BMC Bioinformatics, 2006. **7**: p. 308.

40. Pinkel, D., et al., *High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.* Nature genetics, 1998. **20**(2): p. 207-11.

41. Garraway, L.A., et al., *Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma.* Nature, 2005. **436**(7047): p. 117-22.

42. Lin, W.M., et al., *Modeling genomic diversity and tumor dependency in malignant melanoma.* Cancer research, 2008. **68**(3): p. 664-73.

43. Pleasance, E.D., et al., *A comprehensive catalogue of somatic mutations from a human cancer genome.* Nature, 2010. **463**(7278): p. 191-6.

44. Uren, A.G., et al., *Retroviral insertional mutagenesis: past, present and future.* Oncogene, 2005. **24**(52): p. 7656-72.

45. Kool, J. and A. Berns, *High-throughput insertional mutagenesis screens in mice to identify oncogenic networks.* Nature reviews. Cancer, 2009. **9**(6): p. 389-99.

46. Shore, S.K., R.V. Tantravahi, and E.P. Reddy, *Transforming pathways activated by the v-Abl tyrosine kinase.* Oncogene, 2002. **21**(56): p. 8568-76.

47. Lipsick, J.S. and D.M. Wang, *Transformation by v-Myb.* Oncogene, 1999. **18**(19): p. 3047-55.

48. Graves, B.J., R.N. Eisenman, and S.L. McKnight, *Delineation of transcriptional control signals within the Moloney murine sarcoma virus long terminal repeat.* Molecular and cellular biology, 1985. **5**(8): p. 1948-58.

49. Gunther, C.V. and B.J. Graves, *Identification of ETS domain proteins in murine T lymphocytes that interact with the Moloney murine leukemia virus enhancer.* Molecular and cellular biology, 1994. **14**(11): p. 7569-80.

50. Reisman, D., *Nuclear factor-1 (NF-1) binds to multiple sites within the transcriptional enhancer of Moloney murine leukemia virus.* FEBS letters, 1990. **277**(1-2): p. 209-11.

51. Sun, W., B.J. Graves, and N.A. Speck, *Transactivation of the Moloney murine leukemia virus and T-cell receptor beta-chain enhancers by cbf and ets requires intact binding sites for both proteins.* Journal of virology, 1995. **69**(8): p. 4941-9.

52. Hwang, H.C., et al., *Identification of oncogenes collaborating with p27Kip1 loss by insertional mutagenesis and high-throughput insertion site analysis.*

Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(17): p. 11293-8.

53. Johansson, F.K., et al., *Identification of candidate cancer-causing genes in mouse brain tumors by retroviral tagging.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(31): p. 11334-7.

54. Li, J., et al., *Leukaemia disease genes: large-scale cloning and pathway predictions.* Nature genetics, 1999. **23**(3): p. 348-53.

55. Lund, A.H., et al., *Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice.* Nature genetics, 2002. **32**(1): p. 160-5.

56. Mikkers, H., et al., *High-throughput retroviral tagging to identify components of specific signaling pathways in cancer.* Nature genetics, 2002. **32**(1): p. 153-9.

57. Slape, C., et al., *Retroviral insertional mutagenesis identifies genes that collaborate with NUP98-HOXD13 during leukemic transformation.* Cancer research, 2007. **67**(11): p. 5148-55.

58. Stewart, M., et al., *Insertional mutagenesis reveals progression genes and checkpoints in MYC/Runx2 lymphomas.* Cancer research, 2007. **67**(11): p. 5126-33.

59. Suzuki, T., et al., *New genes involved in cancer identified by retroviral tagging.* Nature genetics, 2002. **32**(1): p. 166-74.

60. Theodorou, V., et al., *MMTV insertional mutagenesis identifies genes, gene families and pathways involved in mammary cancer.* Nature genetics, 2007. **39**(6): p. 759-69.

61. Uren, A.G., et al., *Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks.* Cell, 2008. **133**(4): p. 727-41.

62. Akagi, K., et al., *RTCGD: retroviral tagged cancer gene database.* Nucleic acids research, 2004. **32**(Database issue): p. D523-7.

63. Dudley, J.P., et al., *What retroviruses teach us about the involvement of c-Myc in leukemias and lymphomas.* Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K, 2002. **16**(6): p. 1086-98.

64. Cuypers, H.T., et al., *Murine leukemia virus-induced T-cell lymphomagenesis: integration of proviruses in a distinct chromosomal region.* Cell, 1984. **37**(1): p. 141-50.

65. van Lohuizen, M., et al., *Predisposition to lymphomagenesis in pim-1 transgenic mice: cooperation with c-myc and N-myc in murine leukemia virus-induced tumors.* Cell, 1989. **56**(4): p. 673-82.

66. Lazo, P.A., J.S. Lee, and P.N. Tsichlis, *Long-distance activation of the Myc protooncogene by provirus insertion in Mlvi-1 or Mlvi-4 in rat T-cell lymphomas.* Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(1): p. 170-3.

67. de Ridder, J., et al., *Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens.* PLoS computational biology, 2006. **2**(12): p. e166.

68. de Jong, J., et al., *Computational identification of insertional mutagenesis targets for cancer gene discovery.* Nucleic acids research, 2011. **39**(15): p. e105.

69. Iyer, V.R., et al., *Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.* Nature, 2001. **409**(6819): p. 533-8.

70. van Steensel, B., J. Delrow, and S. Henikoff, *Chromatin profiling using targeted DNA adenine methyltransferase.* Nature genetics, 2001. **27**(3): p. 304-8.

71. Yuan, G.C., et al., *Genome-scale identification of nucleosome positions in S. cerevisiae.* Science, 2005. **309**(5734): p. 626-30.

72. Pokholok, D.K., et al., *Genome-wide map of nucleosome acetylation and methylation in yeast.* Cell, 2005. **122**(4): p. 517-27.

73. Liu, J. and G.D. Stormo, *Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions.* Nucleic acids research, 2005. **33**(17): p. e141.

74. Mukherjee, S., et al., *Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays.* Nature genetics, 2004. **36**(12): p. 1331-9.

75. Berg, O.G. and P.H. von Hippel, *Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters.* Journal of molecular biology, 1987. **193**(4): p. 723-50.

76. Granek, J.A. and N.D. Clarke, *Explicit equilibrium modeling of transcription-factor binding and gene regulation.* Genome biology, 2005. **6**(10): p. R87.

77. Djordjevic, M., A.M. Sengupta, and B.I. Shraiman, *A biophysical approach to transcription factor binding site discovery.* Genome research, 2003. **13**(11): p. 2381-90.

78. Foat, B.C., A.V. Morozov, and H.J. Bussemaker, *Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.* Bioinformatics, 2006. **22**(14): p. e141-9.

79. Tanay, A., *Extensive low-affinity transcriptional interactions in the yeast genome.* Genome research, 2006. **16**(8): p. 962-72.

80. Bussemaker, H.J., B.C. Foat, and L.D. Ward, *Predictive modeling of genome-wide mRNA expression: from modules to molecules.* Annual review of biophysics and biomolecular structure, 2007. **36**: p. 329-47.

81. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(25): p. 14863-8.

82. Tavazoie, S., et al., *Systematic determination of genetic network architecture.* Nature genetics, 1999. **22**(3): p. 281-5.

83. Pe'er, D., et al., *Inferring subnetworks from perturbed expression profiles.* Bioinformatics, 2001. **17 Suppl 1**: p. S215-24.

84. Pe'er, D., A. Regev, and A. Tanay, *Minreg: inferring an active regulator set.* Bioinformatics, 2002. **18 Suppl 1**: p. S258-67.

85. Segal, E., et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.

86. Boorsma, A., et al., *Inferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression.* PloS one, 2008. **3**(9): p. e3112.

87. Bussemaker, H.J., H. Li, and E.D. Siggia, *Regulatory element detection using correlation with expression.* Nature genetics, 2001. **27**(2): p. 167-71.

88. Keles, S., M. van der Laan, and M.B. Eisen, *Identification of regulatory elements using a feature selection method.* Bioinformatics, 2002. **18**(9): p. 1167-75.

89. Gao, F., B.C. Foat, and H.J. Bussemaker, *Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data.* BMC Bioinformatics, 2004. **5**: p. 31.

90. Lee, E. and H.J. Bussemaker, *Identifying the genetic determinants of transcription factor activity.* Molecular systems biology, 2010. **6**: p. 412.

91. Smith, E.N. and L. Kruglyak, *Gene-environment interaction in yeast gene expression.* PLoS biology, 2008. **6**(4): p. e83.

92. Drake, T.A., et al., *Genetic loci determining bone density in mice with diet-induced atherosclerosis.* Physiological genomics, 2001. **5**(4): p. 205-15.

93. Foat, B.C., R.G. Tepper, and H.J. Bussemaker, *TransfactomeDB: a resource for exploring the nucleotide sequence specificity and condition-specific regulatory activity of trans-acting factors.* Nucleic acids research, 2008. **36**(Database issue): p. D125-31.

94. Foat, B.C., et al., *Profiling condition-specific, genome-wide regulation of mRNA stability in yeast.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(49): p. 17675-80.

95. MacIsaac, K.D., et al., *An improved map of conserved regulatory sites for Saccharomyces cerevisiae.* BMC Bioinformatics, 2006. **7**: p. 113.

96. Gaisne, M., et al., *A 'natural' mutation in Saccharomyces cerevisiae strains derived from S288c affects the complex regulatory gene HAP1 (CYP1).* Current genetics, 1999. **36**(4): p. 195-200.

97. Zeng, Z.B., *Precision mapping of quantitative trait loci.* Genetics, 1994. **136**(4): p. 1457-68.

98. Tanaka, K., et al., *IRA2, a second gene of Saccharomyces cerevisiae that encodes a protein with a domain homologous to mammalian ras GTPase-activating protein.* Molecular and cellular biology, 1990. **10**(8): p. 4303-13.

99. Cherry, J.R., et al., *Cyclic AMP-dependent protein kinase phosphorylates and inactivates the yeast transcriptional activator ADR1.* Cell, 1989. **56**(3): p. 409-19.

100. Gorner, W., et al., *Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity.* Genes & development, 1998. **12**(4): p. 586-97.

101. Kasten, M.M. and D.J. Stillman, *Identification of the Saccharomyces cerevisiae genes STB1-STB5 encoding Sin3p binding proteins.* Molecular & general genetics : MGG, 1997. **256**(4): p. 376-86.

102. Pic-Taylor, A., et al., *Regulation of cell cycle-specific gene expression through cyclin-dependent kinase-mediated phosphorylation of the forkhead transcription factor Fkh2p.* Molecular and cellular biology, 2004. **24**(22): p. 10036-46.

103. Ho, Y., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.* Nature, 2002. **415**(6868): p. 180-3.

104. Hollenhorst, P.C., G. Pietz, and C.A. Fox, *Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: implications for regulating the cell cycle and differentiation.* Genes & development, 2001. **15**(18): p. 2445-56.

105. Morillon, A., et al., *Regulation of elongating RNA polymerase II by forkhead transcription factors in yeast.* Science, 2003. **300**(5618): p. 492-5.

106. Badis, G., et al., *Diversity and complexity in DNA recognition by transcription factors.* Science, 2009. **324**(5935): p. 1720-3.

107. Williams, A.J., et al., *Isolation and characterization of a novel zinc-finger protein with transcription repressor activity.* The Journal of biological chemistry, 1995. **270**(38): p. 22143-52.

108. Das, D., Z. Nahle, and M.Q. Zhang, *Adaptively inferring human transcriptional subnetworks.* Molecular systems biology, 2006. **2**: p. 2006 0029.

109. Suthram, S., et al., *eQED: an efficient method for interpreting eQTL associations using protein networks.* Molecular systems biology, 2008. **4**: p. 162.

110. Lee, S.I., et al., *Learning a prior on regulatory potential from eQTL data.* PLoS genetics, 2009. **5**(1): p. e1000358.

111. Kellis, M., et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature, 2003. **423**(6937): p. 241-54.

112. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic acids research, 1997. **25**(17): p. 3389-402.

113. Foat, B.C., et al., *Profiling condition-specific, genome-wide regulation of mRNA stability in yeast.* Proc Natl Acad Sci U S A, 2005. **102**(49): p. 17675-80.

114. Bussemaker, H.J., B.C. Foat, and L.D. Ward, *Predictive modeling of genome-wide mRNA expression: from modules to molecules.* Annu Rev Biophys Biomol Struct, 2007. **36**: p. 329-47.

115. Foat, B.C., R.G. Tepper, and H.J. Bussemaker, *TransfactomeDB: a resource for exploring the nucleotide sequence specificity and condition-specific regulatory activity of trans-acting factors.* Nucleic Acids Res, 2008. **36**(Database issue): p. D125-31.

116. Smith, E.N. and L. Kruglyak, *Gene-environment interaction in yeast gene expression.* PLoS Biol, 2008. **6**(4): p. e83.

117. Brem, R.B. and L. Kruglyak, *The landscape of genetic complexity across 5,700 gene expression traits in yeast.* Proc Natl Acad Sci U S A, 2005. **102**(5): p. 1572-7.

118. Hoerl, A.E. and R.W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Technometrics, 1970. **12**(1): p. 55-67.

119. Broman, K.W., et al., *R/qtl: QTL mapping in experimental crosses.* Bioinformatics, 2003. **19**(7): p. 889-90.

120. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.

121. Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.

122. Steinfeld, I., R. Shamir, and M. Kupiec, *A genome-wide analysis in Saccharomyces cerevisiae demonstrates the influence of chromatin modifiers on transcription.* Nat Genet, 2007. **39**(3): p. 303-9.

123. Ptacek, J., et al., *Global analysis of protein phosphorylation in yeast.* Nature, 2005. **438**(7068): p. 679-84.

124. Farnham, P.J., *Insights from genomic profiling of transcription factors.* Nat Rev Genet, 2009. **10**(9): p. 605-16.

125. Liu, X., et al., *Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection.* Genome Res, 2006. **16**(12): p. 1517-28.

126. Wade, J.T., et al., *Genomic analysis of protein-DNA interactions in bacteria: insights into transcription and chromosome organization.* Mol Microbiol, 2007. **65**(1): p. 21-6.

127. Morse, R.H., *Getting into chromatin: how do transcription factors get past the histones?* Biochem Cell Biol, 2003. **81**(3): p. 101-12.

128. de Wit, E. and B. van Steensel, *Chromatin domains in higher eukaryotes: insights from genome-wide mapping studies.* Chromosoma, 2008.

129. Moorman, C., et al., *Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster.* Proc Natl Acad Sci U S A, 2006. **103**(32): p. 12027-12032.

130. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science, 2010. **330**(6012): p. 1775-87.

131. Chen, X., et al., *Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.* Cell, 2008. **133**(6): p. 1106-17.

132. Sutherland, H. and W.A. Bickmore, *Transcription factories: gene expression in unions?* Nat Rev Genet, 2009. **10**(7): p. 457-66.

133. MacArthur, S., et al., *Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions.* Genome Biol, 2009. **10**(7): p. R80.

134. Li, X.Y., et al., *The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding.* Genome Biol, 2011. **12**(4): p. R34.

135. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.

136. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.

137. Venters, B.J., et al., *A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces.* Mol Cell, 2011. **41**(4): p. 480-92.

138. Warner, J.R., *The economics of ribosome biosynthesis in yeast.* Trends Biochem Sci, 1999. **24**(11): p. 437-40.

139. Ferrigno, P. and P.A. Silver, *Regulated nuclear localization of stress-responsive factors: how the nuclear trafficking of protein kinases and transcription factors contributes to cell survival.* Oncogene, 1999. **18**(45): p. 6129-34.

140. Zheng, W., et al., *Genetic analysis of variation in transcription factor binding in yeast.* Nature, 2010. **464**(7292): p. 1187-91.

141. Lefrancois, P., et al., *Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing.* BMC Genomics, 2009. **10**: p. 37.

142. Pedruzzi, I., et al., *TOR and PKA signaling pathways converge on the protein kinase Rim15 to control entry into G0.* Mol Cell, 2003. **12**(6): p. 1607-13.

143. Swinnen, E., et al., *Rim15 and the crossroads of nutrient signalling pathways in Saccharomyces cerevisiae.* Cell Div, 2006. **1**: p. 3.

144. Raha, D., et al., *Close association of RNA polymerase II and many transcription factors with Pol III genes.* Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3639-44.

145. Wei, M., et al., *Life span extension by calorie restriction depends on Rim15 and transcription factors downstream of Ras/PKA, Tor, and Sch9.* PLoS Genet, 2008. **4**(1): p. e13.

146. Smyth, G.K. and T. Speed, *Normalization of cDNA microarray data.* Methods, 2003. **31**(4): p. 265-273.

147. Pokholok, D.K., et al., *Activated signal transduction kinases frequently occupy target genes.* Science, 2006. **313**(5786): p. 533-536.

148. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

149. Boorsma, A., et al., *T-profiler: scoring the activity of predefined groups of genes using gene expression data.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W592-5.

150. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol, 2004. **5**(10): p. R80.

151. Bernstein, B.E., et al., *Global nucleosome occupancy in yeast.* Genome Biol, 2004. **5**(9): p. R62.

152. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.

153. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

154. Mancera, E., et al., *High-resolution mapping of meiotic crossovers and non-crossovers in yeast.* Nature, 2008. **454**(7203): p. 479-85.

155. Hogan, D.J., et al., *Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system.* PLoS biology, 2008. **6**(10): p. e255.

156. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing.* Science, 2008. **320**(5881): p. 1344-9.

157. Lang, B.D., et al., *The brefeldin A resistance protein Bfr1p is a component of polyribosome-associated mRNP complexes in yeast.* Nucleic acids research, 2001. **29**(12): p. 2567-74.

158. Wilmes, G.M., et al., *A genetic interaction map of RNA-processing factors reveals links between Sem1/Dss1-containing complexes and mRNA export and splicing.* Molecular cell, 2008. **32**(5): p. 735-46.

159. Rojas, P., et al., *Cyclin D2 and cyclin D3 play opposite roles in mouse skin carcinogenesis.* Oncogene, 2007. **26**(12): p. 1723-30.

160. Bouchard, C., et al., *Direct induction of cyclin D2 by Myc contributes to cell cycle progression and sequestration of p27.* EMBO J, 1999. **18**(19): p. 5321-33.

161. Brou, C., et al., *A novel proteolytic cleavage involved in Notch signaling: the role of the disintegrin-metalloprotease TACE.* Molecular cell, 2000. **5**(2): p. 207-16.

162. Harris, S.L. and A.J. Levine, *The p53 pathway: positive and negative feedback loops.* Oncogene, 2005. **24**(17): p. 2899-908.

163. Sherr, C.J., *Divorcing ARF and p53: an unsettled case.* Nature reviews. Cancer, 2006. **6**(9): p. 663-73.

164. Weber, J.D., et al., *p53-independent functions of the p19(ARF) tumor suppressor.* Genes & development, 2000. **14**(18): p. 2358-65.

165. Kelly-Spratt, K.S., et al., *p19Arf suppresses growth, progression, and metastasis of Hras-driven carcinomas through p53-dependent and -independent pathways.* PLoS biology, 2004. **2**(8): p. E242.

166. Kroemer, G. and M. Jaattela, *Lysosomes and autophagy in cell death control.* Nat Rev Cancer, 2005. **5**(11): p. 886-97.

167. Zindy, F., et al., *Myc signaling via the ARF tumor suppressor regulates p53-dependent apoptosis and immortalization.* Genes Dev, 1998. **12**(15): p. 2424-33.

168. Juin, P., et al., *c-Myc-induced sensitization to apoptosis is mediated through cytochrome c release.* Genes Dev, 1999. **13**(11): p. 1367-81.

169. Cohen-Sfady, M., et al., *Heat shock protein 60, via MyD88 innate signaling, protects B cells from apoptosis, spontaneous and induced.* Journal of immunology, 2009. **183**(2): p. 890-6.

170. Chen, Y., et al., *Dual autonomous mitochondrial cell death pathways are activated by Nix/BNip3L and induce cardiomyopathy.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(20): p. 9035-42.

171. Priatel, J.J., et al., *RasGRP1 regulates antigen-induced developmental programming by naive CD8 T cells.* J Immunol. **184**(2): p. 666-76.

172. Keely, P.J., et al., *R-Ras signals through specific integrin alpha cytoplasmic domains to promote migration and invasion of breast epithelial cells.* The Journal of cell biology, 1999. **145**(5): p. 1077-88.

173. Huang, Y., et al., *Role of TC21/R-Ras2 in enhanced migration of neurofibromin-deficient Schwann cells.* Oncogene, 2004. **23**(2): p. 368-78.

174. Jeong, H.W., J.O. Nam, and I.S. Kim, *The COOH-terminal end of R-Ras alters the motility and morphology of breast epithelial cells through Rho/Rho-kinase.* Cancer research, 2005. **65**(2): p. 507-15.

175. Adhikary, S. and M. Eilers, *Transcriptional regulation and transformation by Myc proteins.* Nat Rev Mol Cell Biol, 2005. **6**(8): p. 635-45.

176. Ramsay, R.G. and T.J. Gonda, *MYB function in normal and cancer cells.* Nat Rev Cancer, 2008. **8**(7): p. 523-34.

177. Bryne, J.C., et al., *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D102-6.

178. Sandelin, A. and W.W. Wasserman, *Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics.* Journal of molecular biology, 2004. **338**(2): p. 207-15.

179. Basseres, D.S. and A.S. Baldwin, *Nuclear factor-kappaB and inhibitor of kappaB kinase pathways in oncogenic initiation and progression.* Oncogene, 2006. **25**(51): p. 6817-30.

180. van 't Veer, L.J., R.L. Beijersbergen, and R. Bernards, *N-myc suppresses major histocompatibility complex class I gene expression through down-regulation of the p50 subunit of NF-kappa B.* EMBO J, 1993. **12**(1): p. 195-200.

181. Eymin, B., et al., *Human ARF binds E2F1 and inhibits its transcriptional activity.* Oncogene, 2001. **20**(9): p. 1033-41.

182. Kyo, S., et al., *Sp1 cooperates with c-Myc to activate transcription of the human telomerase reverse transcriptase gene (hTERT).* Nucleic acids research, 2000. **28**(3): p. 669-77.

183. Gartel, A.L., et al., *Myc represses the p21(WAF1/CIP1) promoter and interacts with Sp1/Sp3.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(8): p. 4510-5.

184. Rahl, P.B., et al., *c-Myc regulates transcriptional pause release.* Cell, 2010. **141**(3): p. 432-45.

185. Shachaf, C.M., et al., *Genomic and proteomic analysis reveals a threshold level of MYC required for tumor maintenance.* Cancer research, 2008. **68**(13): p. 5132-42.

186. Caron, H., et al., *The human transcriptome map: clustering of highly expressed genes in chromosomal domains.* Science, 2001. **291**(5507): p. 1289-92.

187. Versteeg, R., et al., *The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes.* Genome research, 2003. **13**(9): p. 1998-2004.

188. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.

189. de Wit, E., et al., *Global chromatin domain organization of the Drosophila genome.* PLoS Genet, 2008. **4**(3): p. e1000045.

190. Lamb, J., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 2006. **313**(5795): p. 1929-35.

191. Markovits, J., et al., *Inhibitory effects of the tyrosine kinase inhibitor genistein on mammalian DNA topoisomerase II.* Cancer research, 1989. **49**(18): p. 5111-7.

192. Negrini, S., V.G. Gorgoulis, and T.D. Halazonetis, *Genomic instability--an evolving hallmark of cancer.* Nature reviews. Molecular cell biology, 2010. **11**(3): p. 220-8.

193. Sirota, M., et al., *Discovery and preclinical validation of drug indications using compendia of public gene expression data.* Science translational medicine, 2011. **3**(96): p. 96ra77.

194. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

195. Mouchiroud, D., et al., *The distribution of genes in the human genome.* Gene, 1991. **100**: p. 181-7.

196. Duret, L., D. Mouchiroud, and C. Gautier, *Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores.* Journal of molecular evolution, 1995. **40**(3): p. 308-17.

197. Lercher, M.J., et al., *A unification of mosaic structures in the human genome.* Human molecular genetics, 2003. **12**(19): p. 2411-5.

198. Nerenz, R.D., M.L. Martowicz, and J.W. Pike, *An enhancer 20 kilobases upstream of the human receptor activator of nuclear factor-kappaB ligand gene mediates dominant activation by 1,25-dihydroxyvitamin D3.* Mol Endocrinol, 2008. **22**(5): p. 1044-56.

199. McLean, C.Y., et al., *GREAT improves functional interpretation of cis-regulatory regions.* Nat Biotechnol. **28**(5): p. 495-501.

200. MacIsaac, K.D., et al., *A quantitative model of transcriptional regulation reveals the influence of binding location on expression.* PLoS computational biology, 2010. **6**(4): p. e1000773.

201. Lin, S.M., et al., *Model-based variance-stabilizing transformation for Illumina microarray data.* Nucleic acids research, 2008. **36**(2): p. e11.

202. Luscombe, N.M., et al., *An overview of the structures of protein-DNA complexes.* Genome biology, 2000. **1**(1): p. REVIEWS001.

203. Bryne, J.C., et al., *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.* Nucleic acids research, 2008. **36**(Database issue): p. D102-6.

204. Shtivelman, E. and J.M. Bishop, *Effects of translocations on transcription from PVT.* Molecular and cellular biology, 1990. **10**(4): p. 1835-9.

205. Huppi, K. and D. Siwarski, *Chimeric transcripts with an open reading frame are generated as a result of translocation to the Pvt-1 region in mouse B-cell tumors.* International journal of cancer. Journal international du cancer, 1994. **59**(6): p. 848-51.

206. Cory, S., et al., *Variant (6;15) translocations in murine plasmacytomas involve a chromosome 15 locus at least 72 kb from the c-myc oncogene.* The EMBO journal, 1985. **4**(3): p. 675-81.

207. Chiarle, R., et al., *Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells.* Cell, 2011. **147**(1): p. 107-19.

208. Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.

209. Pickrell, J.K., et al., *Understanding mechanisms underlying human gene expression variation with RNA sequencing.* Nature, 2010. **464**(7289): p. 768-72.

210. Badis, G., et al., *Diversity and Complexity in DNA Recognition by Transcription Factors.* Science, 2009.

211. Berger, M.F., et al., *Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences.* Cell, 2008. **133**(7): p. 1266-76.

212. Bernstein, B.E., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

213. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.