Alma Mater Studiorum · Università di Bologna

DOTTORATO DI RICERCA
BIOTECNOLOGIE CELLULARI E MOLECOLARI
Ciclo XXII
Settore scientifico disciplinare di afferenza: Vet/02

EDGE PERTURBATION IN THE TRANSCRIPTION NETWORK OF YEAST

Presentata da: Dott.Mirko Francesconi

Coordinatore Dottorato:
Chiar.mo Prof.
Lanfranco Masotti

Relatore :
Chiar.mo Prof.
Monica Forni

Esame finale
2010

Abstract

Biological systems are defined by their components, but also by the interactions between these components. One type of interaction is a regulatory interaction, whereby one component (a gene or its encoded protein) affects the production or activity of a second component. For example, sequence-specific DNA-binding proteins (transcription factors) physically associate with the genome to regulate the expression of genes. Recently, the physical associations of transcription factors with the genome have been mapped on a global scale (Boyer et al., 2005; Carroll et al., 2006; Harbison et al., 2004; MacArthur et al., 2009; Ouyang et al., 2009; yong Li et al., 2008). One way to conceptualize these physical associations is as a graph of genes (nodes) connected by potential regulatory interactions (edges) in a transcription regulatory network (Barabási and Oltvai, 2004; Thieffry et al., 1998).

High-coverage collections of gene-deletion strains or RNA interference reagents have allowed the phenotypic consequences of gene perturbations to be systematically studied in several organisms (Boutros et al., 2004; Giaever et al., 2002; Kamath et al., 2003), also in combination with environmental perturbations (Hillenmeyer et al., 2008). The determinants of gene importance have been widely investigated (Bloom et al., 2006; Pál et al., 2006; Wall et al., 2005), and approaches have been developed that globally predict which genes, when mutated, give rise to which phenotypic changes (Lee et al., 2004, 2008; Peña-Castillo et al., 2008).

In contrast, the importance of interactions between genes (Gao et al., 2004) and the effects of perturbing regulatory interactions (Isalan et al., 2008) are much less well understood. Mutations in regulatory regions are, however, not of low importance. Rather, they are probably the main source of phenotypic variation within and between species (Carroll, 2008; Prud'homme et al., 2007; Wray, 2007) For example, in humans it is likely that most disease-associated polymorphisms alter gene regulation rather than protein coding sequences (Hindorff et al., 2009).

Transcription factors (TFs) normally have short and degenerate sequence-binding preferences (Stormo, 2000). Across a typical eukaryotic genome, therefore these sequences will be found in very large numbers, and indeed genome-wide mapping studies confirm that most TFs are found physically associated at very many locations in a genome (Harbison et al., 2004; yong Li et al., 2008; Zhang et al., 2005). Many of these sites are likely to be of little functional importance (MacArthur et al., 2009; yong Li et al., 2008), but what distinguishes the functional importance of a site?

Features that could influence the importance of a binding site (TFBS) include the proximity of the site to a target gene, and the affinity of the

site for a transcription factor. The biased distribution of TFBSs towards transcription initiation sites in both yeast and mammals (Harbison et al., 2004; Johnson et al., 2009; Tabach et al., 2007; Xie et al., 2005) supports a role for position affecting importance. However the belief that only high scoring, high affinity binding sites are functionally important for gene regulation has been recently challenged (Segal et al., 2008; Tanay, 2006).

One approach to identify functionally important sites is to use evolutionary conservation. Although non-coding regions show generally higher rates of evolution than coding sequences , comparisons between closely related species can be used to discover and analyze sites that have been maintained by purifying selection (Kellis et al., 2003; Odom et al., 2007). To date, although many studies have demonstrated a high rate of TFBS turnover between species (Dermitzakis and Clark, 2002; Doniger and Fay, 2007; Moses et al., 2006), few features are known that influence the conservation of binding sites. Some studies report stronger negative selection against mutations in overlapping (Kim et al., 2009; Mustonen et al., 2008). Different promoter nucleosome positioning has also been found to be associated with different TFBS location and turnover rate (Tirosh and Barkai, 2008). However, despite these observations, a comprehensive study on which TFBSs and regulatory interactions are more important and why is still lacking.

The aim of this work is to partially address this shortcoming using the transcription network of yeast as a model system. Using natural variation both within (Liti et al., 2009) and between (MacIsaac et al., 2006) species I identify features that predict the importance of individual binding sites and regulatory edges on a global scale.

I find that the conservation of a binding site is more strongly influenced by the regulator that recognizes it than by the importance of the target gene. Conservation is also influenced by multiple contextual features of a promoter, including distance from the start site, and the potential for compensatory regulation among sites. Indeed redundancy reduces the importance of binding sites, just as it does for genes. I find that binding site mutations that are likely to influence the expression of multiple genes either by being located in a divergent promoter or by influencing the expression of a regulatory gene are more likely to be detrimental. Moreover that sites bound by TFs higher in the regulatory hierarchy are of greater overall importance. Further, I show that less important sites are enriched in sub-telomeric regions with high mutation rates.

By integrating these features together I construct a model of binding site and network edge importance that shows good predictive performance across an entire genome. These results show that a few simple properties can be used to understand the deleterious effect of mutations that perturb edges

rather than genes in a network. I anticipate that a similar approach may be useful for understanding binding site importance in other species, including in humans.

Contents

Chapter 1

Introduction

1.1 Biological systems: a network perspective

In recent years, thanks to a considerable advance in high- throughput technologies (the so called "omics" technologies) it has become possible to collect large amount of data in a systematic, parallel, and unbiased manner at every biological level: genomic, transcriptomic, proteomic. This new capability in data collection allows now to analyze biological systems as whole. This global approach represents an improvement over the reductionist approach in which each component of the system is analyzed separately. A biological system in fact is defined by its components but also by the connection between these components.

In an useful conceptualization, a biological system can be modeled as a network (or graph) in which components are represented by nodes and interactions are represented by edges between nodes. Networks can be undirected when the interactions between nodes are symmetrical, for example in protein-protein interaction networks the relationship "protein B interacts with protein A" is equivalent to "protein A interacts with protein B". Networks can also be directed when the relationship between nodes is not symmetrical: for example in gene regulatory networks the relationship "gene A regulates gene B" is not equivalent to "gene B regulates gene A". In this case the link is usually represented by an arrow that starts from gene A and points to gene B.

Network models have been applied at many levels of biolgical sytems: protein-protein interactions, metabolism, gene regulation (see Barabási and Oltvai, 2004 for a review). Despite their simplicity, network models can be extremely useful in capture global emergent properties of the system that

cannot be inferred from the properties and functions of its single elements.

1.2 Network structure

First studies mainly focused on topological properties of biological networks that are the properties related to network structure. Several measures have been introduced to quantify topological properties (see box 1.2). Algorithms for the analysis of recurrent network motifs have been developed (Alon, 2007; Milo et al., 2002) providing a description of network organization at local level. For example recurrent network motifs that provide the basic "building blocks" of yeast transcriptional regulatory network have been identified (Lee et al., 2002) as shown in figure 1.1.

Global organization of biological networks has also been described (Babu et al., 2004; Barabási and Oltvai, 2004). From these studies emerge that some features appear to be shared by the majority of biological systems across all organisms. In particular biological network are characterized by as a hierarchical and modular structure in which motifs combine in higher level modules and network of modules. Another feature is the presence of a few highly connected nodes (that are called hubs) that are reponsible of connecting the whole structure, beside many node with low number of connection (Barabási and Oltvai, 2004; Ravasz et al., 2002; Ravasz and Barabási, 2003).

Network models are an extremely simplified representation of the underlying biology but nevertheless proved to be able to capture and explain important functional properties of biological systems and of its components. It has been shown for example that gene lethality is largely associated with the topological position of its protein product in the protein protein interaction network of yeast (Jeong et al., 2001). Recently network hierarchical structure has also been found to relate to dynamical properties of its nodes such as the rate of mRNA and protein production and degradation, or gene expression noise (Jothi et al., 2009). Examples of functional properties related to structure of the network have been found also in metabolism (Stelling et al., 2002; Wunderlich and Mirny, 2006)

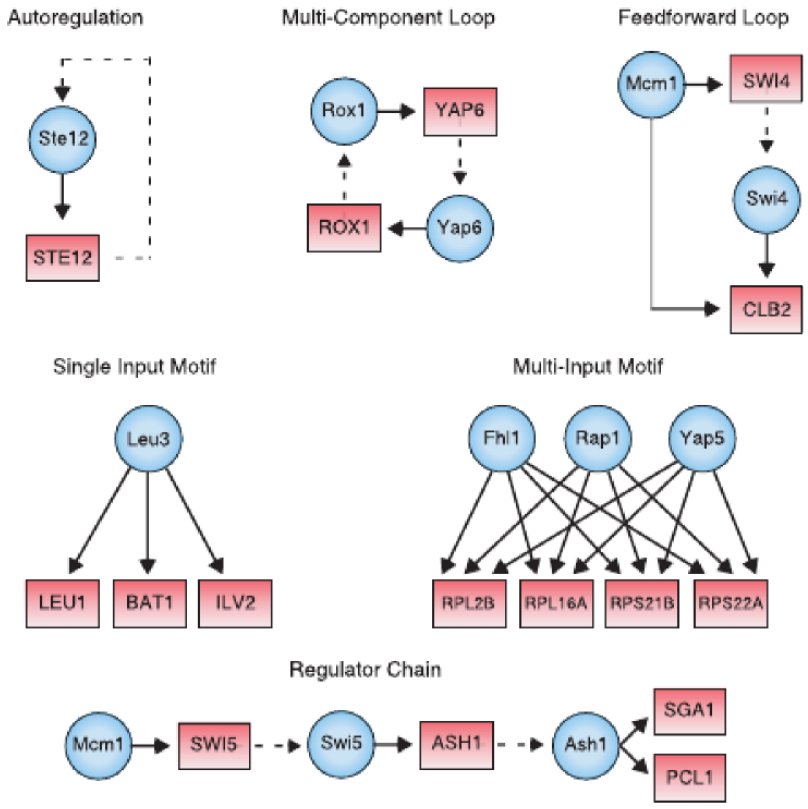

Figure 1.1: Network motifs in transcriptional regulatory network of *Saccharomyces cerevisiae* (taken from Lee et al., 2002)

Box 1.2. Network topological measures

Node degree or connectivity. The number of edges that connect a node with the other nodes in the network. In directed networks the number of edges that point to a node is defined as in degree while the number of edges start from a node is defined out degree.

Node degree distribution. Degree distributio $P(k)$ Represents the probability that a given node in a network have k edges. Usually biological networks and other real networks do not show normal distribution of node degree. They usually show a heavy tailed distribution that highlights the presence of a non negligible number of nodes that are extremely connected that are defined as network hubs.

Shortest path and mean path length. Distance between two nodes in a network is measured using shortest path length that is the minimal number of edges that separate the two nodes. Mean path length is the average over shortes path length between all pairs of nodes in a network.

Clustering coefficient. Many networks show groups of nodes that are very well connected to each other this can be quantified by the clustering coefficient: for each node i the number of existing link n among its k_i first neighbors over the total maximum possible number of links is defined as:$C_i = \frac{2n_i}{k_i(k_i-1)}$
so the clustering coefficient of the network G with N nodes can be defined as: $C(G) = \frac{1}{N}\sum_{i=1}^{N} C_i$

1.3 Network dynamics

Biological systems are constantly changing in time to develop, divide and in response to environmental stimuli. This important aspect is not captured by a topological analysis alone that rather represents a unique picture of all possible interactions superimposed. Several studies have investigated how biological networks change over time. For example combining gene expression data with genome wide location data of transcription factors allowed to describe the dynamics of transcriptional regulatory network of yeast during cell cycle progression (Lee et al., 2002) or under different environmental conditions (Luscombe et al., 2004). In higher multicellular eukaryotes network

changes occur during development of organism in differentiated tissues. A recent study examined how the network of protein protein interaction in human changes with various tissue and give insights on how the tissue specific module are connected to the conserved core of house keeping proteins (Bossi and Lehner, 2009)

1.4 Network perturbation

One of most useful approaches to understand a biological system is by systematically study how it reacts to perturbations. Different types of perturbations can affect biological systems: genetic perturbations, such mutations or polymorphisms - or environmental perturbation such as physical (heat) or chemical (drugs) stimulus but also stochastic perturbation i.e. noise. With perturbation strategy it is possible to identify which parts of the systems are important for which responses (Giaever et al., 2002; Hillenmeyer et al., 2008; Ideker et al., 2001; Kamath et al., 2003). By using mathematical models it is also possible to infer how system elements are connected and reconstruct a global network structure (Bansal et al., 2007; Bonneau et al., 2006; Gardner et al., 2003). Once the network is reconstructed it possible to predict system behavior and phenotypic output under different perturbations (Bonneau et al., 2007).

It is also possible to address questions on how network constrains the behaviour of single elements in the system and how network properties can explain systems behavior and phenotypic output under perturbations. For example it has been recently shown that selective pressure acts in minimizing the noise level of components that are essential or important for normal growth because they are part of large interaction complexes and it is important to preserve stoichiometric proportions of the interacting elements (Fraser et al., 2004; Papp et al., 2003). Selection also acts in minimizing genes that are harmful when overexpressed (Lehner, 2008) and recently it has been speculated that this depends on the property of these genes to make many low affinity out of target interactions when overexpressed (Vavouri et al., 2009).

Two main strategies are used to analyze perturbations in biological systems: first, it is possible to analyze the effect of systematic experimentally induced perturbations on a system (Boutros et al., 2004; Deutschbauer et al., 2005; Giaever et al., 2002; Kamath et al., 2003). Second it is possible to study natural perturbations, for example natural genetic variation that arise within populations or between species during evolution using comparative genomics (Boffelli et al., 2003; Kellis et al., 2003; Pál et al., 2006)

1.4.1 Systematic node perturbation

Many studies have systematically investigated phenotypic consequences of node perturbation (i.e. perturbations of genes or proteins). High-coverage collections of gene-deleted strains have been generated in yeast (Deutschbauer et al., 2005; Giaever et al., 2002) or systematic analysis of iRNA based inhibition of gene function have been performed in C elegans (Kamath et al., 2003) and Drosophila melanogaster (Boutros et al., 2004). The phenotypic consequence of perturbation due to overexpression of a gene rather than inhibition has also been analyzed in yeast (Gelperin et al., 2005; Sopko et al., 2006).

From these large scale studies on gene perturbation it has been possible to compile catalogs of genes that are essential or important for grow or genes that are responsible for determined phenotypes. Strikingly it has been found that essential genes are a minor fraction of the total genes in yeast (17%) (Winzeler et al., 1999) and that only 10% of iRNA inhibited genes in C. elegans show a phenotype (Kamath et al., 2003). Perturbing combinations of genes in yeast (Tong et al., 2004) and C. elegans (Lehner et al., 2006) allowed the identification and the study of genetic interactions.

Genetic perturbations in combination with environmental perturbations have also been studied in yeast (Hillenmeyer et al., 2008). From this study emerges that more than 70% of genes show a phenotype in combination with chemical or physical perturbations. Probably the most important finding of these system levels studies is that gene importance cannot be completely understood only studying gene specific properties or function because the interaction with both the genetic and the environmental context in which they operate play a fundamental role.

1.4.2 Phenotype prediction of node perturbation with network models

The availability of large scale functional dataset also allowed the development of methods that globally predict which genes when mutated induce which phenotypic changes. The most used methods exploit the large collection of datasets from different experimental techniques such as gene expression, protein-protein interactions and phenotype annotations. Each dataset provides a piece of evidence that two genes are functionally interacting. Using bayesian statistical methods it is possible to combine all these evidences into a unifying probabilistic functional network (Lee et al., 2004, 2008). In the integerated network a link between two genes represents the likelihood that these two genes are functional interacting. With this network it is pos-

sible to predict gene function and to identify new genes that were previously unknown to be involved in biological pathways or processes with a superior accuracy and more predictive power than any of the individual datasets from which it derives (Lehner, 2007). With these functional newtorks it is also possible to predict loss of function phenotypes of genes using a guilty by association approach (McGary et al., 2007). This approach have been successfully applied in unicellular microorganisms such as *S. cerevisiae* (Lee et al., 2004) in the more complex nematode *C. elegans* (Lee et al., 2008) and has now been tested in mouse (Peña-Castillo et al., 2008).

1.4.3 Studying natural perturbations using comparative genomics

Natural genetic variation occurring within populations or between species represent a very powerful tool to investigate biological systems. Comparative genomics can be used to individuate genomic regions conserved during evolution that correspond to functional elements (Kellis et al., 2003). Using genomic data it is possible to estimate the selective force acting on functional elements analyzing their the evolutionary rate. A large number of studies focused on coding sequences: it has been found that evolutionary rates of proteins vary over several orders of magnitude. The factors that cause such large differences of evolutionary rate have been widely investigated (Pál et al., 2006). The effect of structural properties have been determined (Bloom et al., 2006) as well as the effect of expression levels of the protein and its dispensability(Wall et al., 2005). Also genomic location and position in biological networks have been found to be important determinants (see (Pál et al., 2006) for a review). Comparative genomic studies have also important consequences on human health, in fact comparing observed rate of protein evolution can also be useful to estimate the contribution of mutations to diseases (Pál et al., 2006). Detrimental mutation that are more likely to cause a disease in fact are subjecte to higher puryfying selection during evolution. This approach has proven to be useful and several tools are now available to prioritize single nucleotide polymophisms in coding regions in the study of genetic componet of human diseases (Ng and Henikoff, 2003; Ramensky et al., 2002).

1.4.4 Edge perturbations

All the studies cited above are focused on systematic node perturbation in biological systems or on natural genetic perturbations on genes. They thus give

insights on node importance in biological networks. In contrast, the importance of interaction between genes and the effect of perturbing edges rather than nodes in biological systems is much less understood. An illuminating example is a recent pioneering study where gene regulatory network in E. coli has been rewired (Isalan et al., 2008). In this work 598 recombinations of promoters with genes coding for transcription factors and sigma factors of E. coli have been added to wild type genetic background. Thus new links where added to existing regulatory network. A totally unexpected result was that the great majority (95%) of rewired networks were tolerated by the bacteria and even more surprisingly some of the rewired bacteria had a fitness advantage over the wild type under several environments. This study highlights our little understanding on the evolutionary constraints on gene regulatory network edges and thus also on the possible phenotypic consequences on edge perturbation.

1.5 Transcriptional regulatory networks

Living beings are the result of the coordinated expression of thousands of genes and this process is fundamental for all the organism to ensure cellular homeostasis, replication, development and proper response to external stimuli. Such a complex process involves several levels of regulation, from mRNA production to protein post-translational modifications. The most important regulatory step usually occurs at the level of mRNA transcription (Lu et al., 2007). The regulation of transcription depends on regulatory proteins - the transcription factors (TFs) - that recognise short DNA sequences across the genome. These sequences can be found in the vicinity of the gene trascriptional start site (TSS) and they are called promoter, or, especially in higher eukaryotes, even far away from the TSS and they are called enhancers. The TFs bound to regulatory regions recruit chromatin modifier complexes and transcriptional machinery to start the transcriptional process of the target gene. How a collection of TFs associates with genes along a genome can be described as a transcriptional regulatory network (TRN), where the nodes represent the TFs and the target genes and the edges represent the regulatory interactions.

1.5.1 Experimental determination of a transcriptional regulatory network

Network level studies of transcriptional regulation have become feasible thanks to an experimental technique that combines chromatin immunoprecipitation

and microarray technology called ChIP-chip (see (Aparicio et al., 2005) for a review). This technique allows to map the binding of individual proteins across the genome in vivo. The procedure consists on formaldehyde based crosslinking of proteins to DNA in vivo and subsequent sonication and immunoprecipitation of the protein of interest. The immunoprecipitate is subsequently reverse-crosslinked and the DNA fraction is then analyzed with microarrays to determine bound sequences. In 2002 a systematic genome-wide location analysis of 106 yeast transcription factors using a c-Myc epitope tagging system allowed to build a map of transcription factors and associated genes that represents all the regulatory potential af a transcription network(Lee et al., 2002). Integrating binding location data with gene expression data allowed the authors to reconstruct dynamics of activation of different network modules at different phases of cell cycle or under different environmental conditions. This work initiated the study of eukaryote trascriptional regulation at a system level. Recently a new experimental technique have been introduced to analize DNA binding in vivo that is called Chip-seq It is again based on chromatin immuno precipitation but make use of new ultra high put sequencing tecnologies to analyzed the precipitated DNA (Johnson et al., 2007). With both Chip-Chip and Chip-seq genomic scale maps of regulatory protein - DNA interactions and histone DNA interaction are being produced at high rate and for increasingly complex organisms including humans.

1.5.2 Bioinformatic discovery of transcription factor binding sites

An important step in the study of the transcriptional networks is the deciphering the regulatory code that accounts for the binding of the TFs to the regulatory regions. From experimental studies on TF binding emerged that different sequences can be bound by a TF with a wide range of affinity. This means that motifs bound by TFs are degenerate and tolerate variability of nucleotide at some positions. In this way it is possible to modulate the strength of the binding without abolishing it. The most simple model that has been used to represent a TF motif is a consensus sequence. This model indicates the most conserved positions with A,T,C,G corresponding to adenine, thymine, cytosine and guanine, and it uses other symbols to identify less specific positions that can vary among two three or four nucleotides across the different binding sites according to IUPAC nomenclature (Cornish-Bowden, 1985). A more sensible model has been introduced that represents TFs motifs by means of a position specific score matrix (PSSM)

(table 1.1).

Table 1.1: Log-likelihood position specific scoring matrix (PSSM) for yeast transcription factor PHO2

	0	1	2	3	4	5
A	1.255	-4.501	-5.552	1.495	1.620	0.730
C	-8.907	0.285	-4.227	-8.993	-9.193	-0.720
T	-0.539	1.249	1.598	-7.097	-9.254	-2.523
G	-8.562	-8.681	-9.436	-1.105	-8.303	0.767

This TF model can be constructed using alternatively the frequency of each base at each position that is found in a collection of experimental BSs, or the log likelihood computed from the base frequencies (see (Stormo, 2000; Stormo and Fields, 1998) for reviews). With a log likelihood matrix model a binding site can be given a score as follow:

$$s = \sum_{i=1}^{N} \log_2(p_i) \tag{1.1}$$

where:

N = length of the sequence

p_i = likelihood of the i symbol in the PSSM.

If the backgroung frequency of the bases in the genome is not random it has to be taken into account in the calculation of score that becomes:

$$s = \sum_{i=1}^{N} \log_2(p_i/b_i) \tag{1.2}$$

where:

b_i = likelihood of the i symbol in the background model.

The information content of a PSSM can be calculated. This represent how different is a PSSM from the background distribution.

The information content I_i of a particular i position in the model is:

$$I_i = \sum_{b=A}^{T} p_{bi} \log_2(p_{bi}/p_b) \tag{1.3}$$

where i refers to the position b refers to each base and p_{bi} is the likelihood of each base at that position and p_b is the background frquency of b.

The total information content (IC) of a motif it is defined as:

$$IC = \sum_{i=1}^{L} I_i \tag{1.4}$$

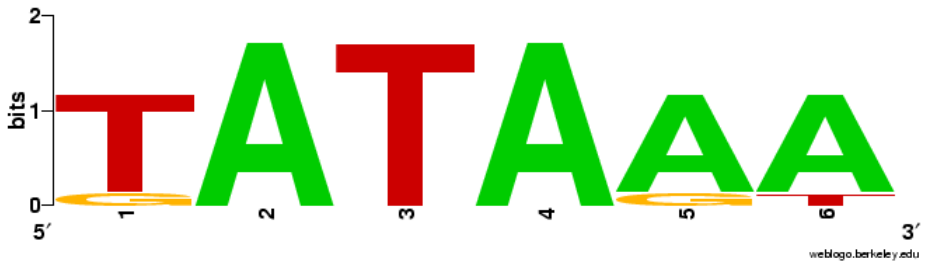

Figure 1.2: Motif logo: a graphical representation of motif model. The height of each letter is proportional to its information content

where L is the length of the motif.

A graphical representation of the motif called motif logo has been developed (Schneider and Stephens, 1990) in which each base contribution to the information content at each position is represented by letters in which the height is proportional to the information content (Figure 1.2).

The representation of the binding motif is only one aspect of the problem, while the other important aspect is the discovery of binding sites. The PSSM model needs a collection of sequences to be constructed. The input sequences can be derived directly from precise binding experiments with the transcription factor of interest, in which the exact position of binding site is know. Once the PSSM has been determined it can then be used to scan sequences to search for binding sites instances.

An alternative approach is to search for binding sites and to construct the PSSM model at the same time. The strategy in this case is to have a collection of related sequences, for instance upstream regulatory region of genes that are co-expressed in a particular condition. Than an *ab initio* motif discovery algorithm has to be applied to search for significantly overrepresented motifs in the input sequence . Considerable efforts have been posed in the last years to develop such algorithms. They can be divided into two main groups: enumerative methods that explore all the possible motifs up to a certain length (Corà et al., 2004; Sinha and Tompa, 2002, 2003) and local search algorithms that include expectation maximization and gibbs sampling (Lawrence and Reilly, 1990; Lawrence et al., 1993; Thijs et al., 2002) (see (Tompa et al., 2005) for a comparison of the methods). The outputs of these algorithm are the putative regulatory motifs and their location in the input sequences.

Computational analyses at single genome level have been succesfully applied to identify regulatory elements associated with sets of related genes, but this approach does not have sufficient power to allow a comprehensive

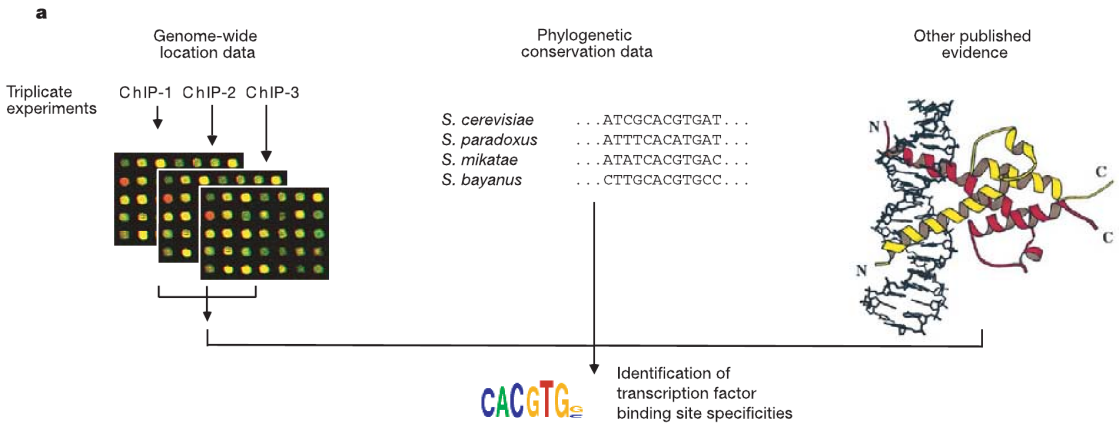

Figure 1.3: Experimental procedure used to identify yeast regulatory code (taken from Harbison et al., 2004)

identification of regulatory elements. Comparative genomics offers a powerful tool to improve the discovery of important functional elements using conservation among related species and it has been succesfully applied from yeast (Kellis et al., 2003) to mammals (Boffelli et al., 2003) as well as for identifying ultra-conserved elements among vertebrates (Boffelli et al., 2004). Many motif discovery algorithms that exlpoit comparative genomics to find conserved cis-regulatory elements have been developed (Cartharius et al., 2005; Corà et al., 2005; Loots et al., 2002; Newberg et al., 2007; Ovcharenko et al., 2004; Siddharthan et al., 2005; Wang and Stormo, 2003).

1.5.3 Yeast transcriptional regulatory map

The first version of S. cerevisiae genomewide regulatory map has been built in 2004 (Harbison et al., 2004) using both an experimental and bionformatic approach (Figure 1.3).

The authors determined genome wide location of the majority of yeast regulators in rich media and other environmental conditions using ChIP-chip analysis of yeast intergenic regions. A combination of six motif discovery algorithms has been applied to the significanly bound intergenic regions to uncover motif specificity for 106 regulators and to draw a map of their binding sites location across the whole genome (Figure 1.4). A refined version of this map has been obtained using two more efficient motif discovery algorithms (MacIsaac et al., 2006). This map represents the transcriptional regulatory code of the budding yeast that is currently used and that has been used also

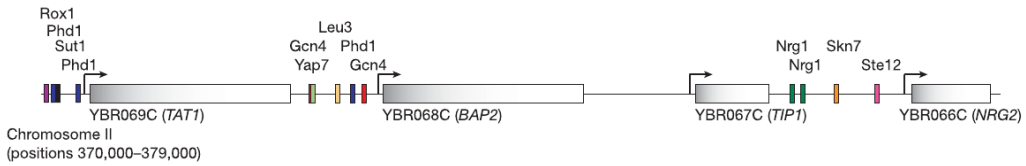

Figure 1.4: Example of genomic map of transcription factor binding sites (taken from Harbison et al., 2004)

in this work (see chapter 5)

1.6 Importance of cis regulatory mutations

The evolutionary significance of cis regulatory variation has been discussed soon after the discovery of the the lac operon and of the gene regulation (JACOB and MONOD, 1961) when the researches speculate on the possible effect of a change in the sequence of the operator for the proper conditions under which an enzyme is produced (MONOD and JACOB, 1961). In 1975 in one of the first examples of comparative genomics analysis (King and Wilson, 1975), it has been realized that human proteins and the homologous proteins of his closely related species chimpanzee share impressive similarity. Thus in this influential paper, the authors speculated that the sparse differences in protein sequences are very unlikely to explain the profound phenotypic differences between human and chimpanzee. They than speculated that evolution must act at a second level that is in cis-regulatory sequences and gene expression that could be in turn responsible for the phenotypic differences between the species. This intuition proved to be substantially correct and today there are many examples of cis-regulatory mutations that have functional consequences for morphology physiology and even behavior in several organisms (Table 1.2). The relative contribution of coding sequence and cis-regulatory sequence to phenotypic evolution is still under debate (Hoekstra and Coyne, 2007) but it is now clear that cis-regulatory sequences play an important role (Carroll, 2000, 2005).

Recently scientists started to debate if cis regulatory mutations confer qualitative different phenotypes from coding sequence mutations (Carroll, 2005, 2008; Hoekstra and Coyne, 2007; Prud'homme et al., 2007; Wray, 2007). In developmental biology there are many example of evolution of morphological traits during development that can be traced back to cis regulatory mutations(Gompel et al., 2005; Prud'homme et al., 2006). This observation support that cis regulatory mutation can give rise to new forms more easily

Table 1.2: Cis-regulatory mutations with interesting phenotypic consequences (taken from Wray, 2007)

Gene	Function of product	Phenotype	Taxon
AVPR1A	Vasopressin receptor	Creative dance performance	Humans
Avpr1a	Vasopressin receptor	Paternal care	Rodents
Cyp6G1	P450 enzyme	Pesticide resistance	Fruitflies
DARC	Chemokine receptor	Resistance to infection with malaria	Humans
e	Pigment synthesis	Colour pattern of abdomen	Fruitflies
hsp70	Heat shock protein	Thermal tolerance	Fruitflies
HTR2A	Serotonin receptor	Obsessive-compulsive behaviour	Humans
IL10	Interleukin	Outcome of infection with HIV and infection with leprosy	Humans
IL10	Interleukin	Susceptibility to schizophrenia	Humans
LCT	Digestive enzyme	Lactose persistence	Humans
LDH	Metabolic enzyme	Cardiac physiology	Killifish
ovo/svb	Transcription factor	Bristle pattern on larvae	Fruitflies
MAOA	Neurotransmitter turnover	Aggressive behaviour	Humans
MMP3	Matrix metalloprotease	Risk of heart disease	Humans
PDYN	Neuropeptide	Memory, emotional status	Humans
pitx1	Transcription factor	Skeletal patterning	Stickleback fish
sc	Transcription factor	Bristle pattern on adult notum	Fruitflies
SLC6A4	Serotonin transporter	Depression, creativity, anxiety	Humans
SLC6A4	Serotonin transporter	Dispersal behaviour	Macaques
tb	Transcription factor	Branching structure	Maize
Ubx	Transcription factor	Bristle pattern on adult legs	Fruitflies
y	Pigment synthesis	Colour pattern of cuticle Mating behaviour	Fruitflies

than coding sequence mutation (Carroll, 2005; Prud'homme et al., 2007) but the idea that morphology evolution is more likely to occur through cis regulatory mutations is still under debate (Hoekstra and Coyne, 2007) as well as the relative contribution of cis regulatory mutation and coding mutations on general phenotypic variation (Wray, 2007).

1.6.1 Impact of cis regulatory variation in human diseases

With modern sequencing and genotyping technology a great amount of data on natural sequence variation within a population can be collected. This data hold promise to provide invaluable unbiased insights into the genetic component of common phenotypes and complex diseases in humans. An early conclusion drawn from the increasing number of genome wide association studies (GWAS) is that a large fraction of estimated heritability of common traits and diseases is still missing, probably due to lack of power for estimating the collective effect of low-size effect genetic variants. Moreover the effect of interactions between genetic variants and between environment and genetics is also difficult to capture (Maher, 2008). Apart from this disappointing observation, what emerges from these studies is that many of the trait associated SNPs (TAS) are common but they typically have a small effect on the phenotype (Hindorff et al., 2009). A functional analysis of TASs show that SNPs in promoter regions are highly significantly enriched in TASs (Hindorff et al., 2009). This finding highlights that the study of cis regulatory variation has important consequence also for understanding the genetic bases of human common disease.

1.6.2 Contribution of cis- and trans- variation to phenotype evolution

In the last years many studies studied how gene expression programs have changed during evolution comparing related species ((Khaitovich et al., 2006; Whitehead and Crawford, 2006). Changes in gene expression programs imply changes in gene regulation. These changes can occur both at cis level in regulatory regions of the affected genes or at trans level that include signaling, chromatin modifiers and transcriptional regulators (Figure 1.5 see Thompson and Regev, 2009, for a review).

Several studies especially in yeast have been done to dissect variability in gene expression due to trans or cis effects both within and between species. Two strategies have been applied to this purpose: one is the cross between

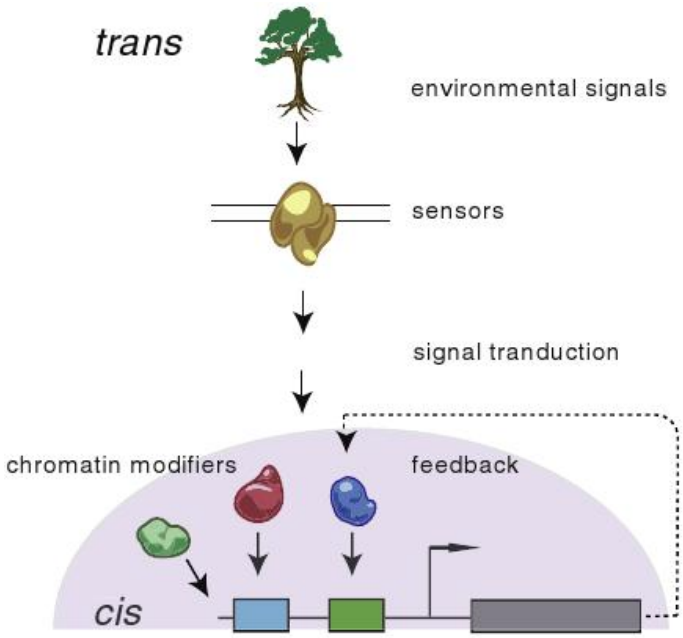

Figure 1.5: Cis- and trans- contribution to gene expression changes

distinct strains and the analysis of the segregants (Brem and Kruglyak, 2005; Brem et al., 2002; Yvert et al., 2003);the other is the analysis of allele-specific expression using intra-specific (Ronald et al., 2005; Sung et al., 2009; Wang et al., 2007) or inter-specific hybrids (Tirosh et al., 2009). From these studies emerge that trans variation is probably more important than cis variation in affecting gene expression within species while, in contrast, cis variation seems more important between species. This apparent paradox can be explained by the difference in dominance effect between cis and trans variation: trans variation is rapidly accumulated it is usually dominant and highly pleiotropic so its effect frequently deviates from additivity (Lemos et al., 2008). so trans effect account for phenotypic variation over short time scales while over longer time scale deleterious trans mutations are more effectively purged by purifying selection. In contrast cis variation is accumulated more slowly but its effect is additive and weakly pleiotropic it can be more easily subjected to positive selection (Thompson and Regev, 2009). An example of trans effect predominance comes from a recent study on the causes of the sporulation efficiency difference between a wild S cerevisiae strain isolated from a oak tree and a vineyard strain (Gerke et al., 2009). By crossing the strains and analysing the segregants, the authors found that the difference in sporulation efficiency is fully explained by allelic variation in three transcription factors IME1, RME1 and RSF1. One of these transcription factors, IME1, carries a

mutation in a cis-regulatory region conserved between several yeast species, while the other mutations cause non-synonymous substitutions in transcription factors coding sequences. This study also shows that the combined effect of the alleles is higher than the sum of the single effects, thus underscoring also the important role of genetic interactions.

1.6.3 Large scale cis-regulatory changes

In some cases large scale changes in cis regulatory regions associated with phenotypic changes have also been found. For example comparing S. cerevisiae and C. albicans, two distantly related yeast species, it has been found that the genes coding for mitochondrial and cytoplasmic ribosomal proteins are coordinately expressed in C. albicans while in S cerevisiae this correlation is not present (Ihmels et al., 2005) this difference in gene expression programs accompanies the change between a preferred aerobic metabolism, present in common ancestor and mantained in C.albicans, to a preferred anaerobic metabolism that instead represent an evolutionary change occured in S. cerevisiae. The authors found that this change is associated to the loss of a transcriptional regulatory motif - the Rapid Growth Element (RGE)- upstream all of genes that encodes mitochondrial ribosomal proteins in S. cerevisiae.

Another study on yeast showed that changes in transcription factor binding sites of the activator of mating response STE12 among different species could explain half of the changes in the genes up-regulated during mating response (Tirosh et al., 2008).

1.7 Challenges in studying cis-regulatory variation

In protein coding regions is easy to identify which mutations can be deleterious only looking at the sequence: we can immediately tell if the mutation cause amino-acid substitution, frame-shifts, or introduce premature stop codons. This is due to fact that the genetic code is fully understood. On the contrary the rules that underlie transcription regulation are still largely unclear (see (Weirauch and Hughes, 2010; Wray, 2007) for reviews). It is quite common for instance that cis-regulatory sequences that show negligible level of conservation determine equivalent output in gene expression (Chan et al., 2009). This phenomenon well represents the challenges for the study of regulatory mechanisms and deeply affect our ability of predicting the effect of

regulatory perturbations. What determines functionality of a regulatory region and allows its conservation in face of poor sequence conservation during evolution has been just started to be delineated.

1.7.1 Transcription factor binding is often non functionally important

Eukaryotic transcription factors (TFs) normally have short and degenerate sequence-binding preferences. Therefore, in a typical eukaryotic genome, these sequences will be found in very large numbers by chance. Genome-wide mapping studies confirm that most TFs are found physically associated at very many locations in a genome (Harbison et al., 2004; yong Li et al., 2008; Zhang et al., 2005). Many of these binding sites are found to be of little or no functional importance. It has been reported for example that CREB (cAMP response element binding protein) binds approximately 4000 human promoters but only a minor fraction of the associated gene are actually induced by cAMP in any cell type (Zhang et al., 2005). Similar conclusions are drawn in Drosophila (yong Li et al., 2008) and even in simpler S. cerevisiae where no significant correlation has been found between transcription factor promoter occupancy and gene expression for the majority (67%) of yeast transcription factors (Gao et al., 2004). A recent theoretical study demonstrates that, in contrast to prokaryotes, single binding sites in eukaryotes do not have sufficient information content to ensure proper gene regulation. The lacking information is achieved by the combinatorial association of several binding sites (Wunderlich and Mirny, 2009). As a consequence of these observations it is now clear that the binding of a transcription factor to a promoter is not sufficient *per se* to determine a regulatory interaction on the gene.

1.7.2 Extensive turnover affect binding site

Turnover or shuffling of binding sites - that is binding site loss in a regulatory region accompanied by a co-occurring gain of binding site in the same region but in a different position - is frequent in yeast, (Doniger and Fay, 2007; Raijman et al., 2008), fly (Moses et al., 2006) and mammals (Dermitzakis and Clark, 2002). A recent study analyzed tissue specif transcriptional regulation between two closely related species such as human and mouse. Despite the conserved function of the tissue specific factors and the conservation of the regulatory program, this study has found that from 41% to 89% of binding events of transcription factors to promoters sites of orthologous genes are species specific. In addition in promoters bound by the same transcription

factors approximately two third of binding sites do not align. A conclusion that can be drawn by these results is that TFBS are indeed important for regulating gene expression but their importance greatly vary and is not clear what distinguishes a functional important binding site that are is likely to affect organism phenotype when perturbed, from a less important one. Which are the determinants of binding site importance is still largely unclear.

1.8 Features affecting binding site importance

Some features that can influence binding site importance have been suggested or identified. Several studies for example found a biased distribution of TFBSs towards transcription initiation sites in both yeast and mammals (Harbison et al., 2004; Tabach et al., 2007; Xie et al., 2005) supporting a role for position in determining importance of binding sites. Moreover in a comparative study of the vertebrate transcriptional repressor of neural specific genes REST across multiple species emerge that more ancient binding sites are closers to the transcription start site and show increased affinity for the transcription factor than the species specific ones.

Higher binding sites strength has been associated to higher conservation also in Drosophila (Kim et al., 2009). However the belief that only high scoring, high affinity binding sites are functionally important for gene regulation has been challenged in a study which found that although stronger binding sites affect gene expression than more low affinity binding sites, the latter can induce significant gene expression and therefore can be important for fine tuning of gene expression (Tanay, 2006). Indeed a recent work that predict gene expression output of the gene network that regulate patterning formation in Drosophila from cis regulatory sequences, found that weaker binding sites are functional important for patterning formation and also confer robustness against mutation (Segal et al., 2008).

A large fraction of TFBS in regulatory regions are overlapping. A mutation in overlapping binding sites is likely to affect binding of more than one transcription factors therefore it should be under stronger purifying selection. Some studies indeed report a stronger negative selection against mutations in overlapping (Kim et al., 2009; Mustonen et al., 2008). In some cases also closely located binding sites are found to be under stronger selective pressure perhaps indicating an significant effect of cooperativity (Kim et al., 2009).

Gene regulation is affected not only by the binding of transcription factors to the promoters but also by the chromatin structure related to the promoter. In particular the position of the nucleosome at the promoter region has been recently found to play an important role gene expression regulation (Field

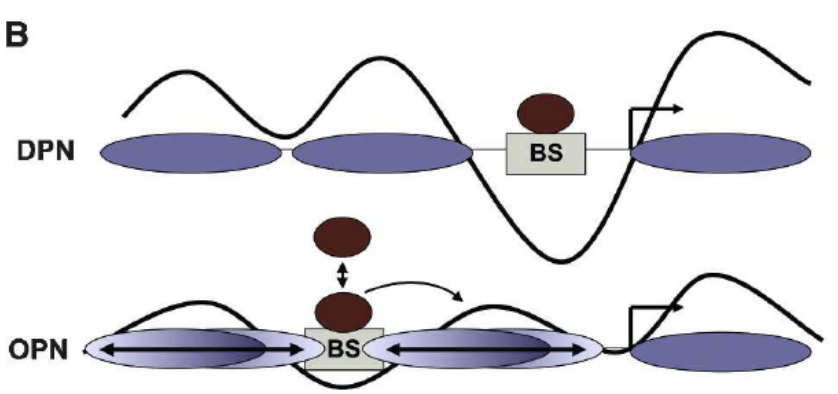

Figure 1.6: Two regulatory strategies for gene regulation taken from (Tirosh and Barkai, 2008)

et al., 2008; Raveh-Sadka et al., 2009; Tirosh and Barkai, 2008). Analyzing the nucleosome position across all yeast promoters two configurations appear to be the most common: in one configuration the promoter shows a large nucleosome free region immediatly upstream the transcription start site (up to -100 bp) and well positioned nucleosome further upstream (-150 ,-400 bp); this configuration has been defined depleted proximal nucleosome (DPN). The other configuration defined occupied proximal nucleosome (OPN), shows a less well positioned nucleosome that can be found all along the promoter also in the region proximal to the transcription start site. Binding sites are found to be more densely distributed in the nucleosome free region in DPN promoters while they are more uniformely distributed in OPN. Moreover binding site in OPN promoters are found to be subjected to 5-fold higher turnover rate. The two models can be seen in Figure 1.6.

Chapter 2

Aim of the work

Mutations can affect genes but also the interactions between genes. In transcriptional networks the interactions (or network edges) are defined by the binding of transcription factors to the cis regulatory regions of target genes determining their expression. Mutations in regulatory regions are probably one of the main source of phenotypic variation within and between species (Carroll, 2008; Wray, 2007). For example in humans it is likely that most disease-associated polymorphisms alter gene regulation rather than protein coding sequences (Hindorff et al., 2009). The effect of node perturbation in biological systems and the determinants of node importance have been widely investigated (Pál et al., 2006; Wall et al., 2005). In contrast, the importance of interactions between genes (Gao et al., 2004) and the effects of perturbing regulatory interactions(Isalan et al., 2008) are much less well understood.

The aim of this work is to systematically study edge perturbation in transcriptional regulatory network of budding yeast to uncover global determinants of edge importance, analyzing natural genetic variation within species (Liti et al., 2009) and between species (MacIsaac et al., 2006). Considering a prevalent role of purifying selection during cis-regulatory evolution (Ronald and Akey, 2007), binding sites and regulatory interactions that are conserved both within and between species will be enriched for the most functionally important sites and interactions.

In this study I consider several potential sources of differential conservation of regulatory edges at different levels. I first consider how the importance of the target gene influence edge conservation, but also the role of the importance of the regulator on edge importance. I analyze how the design properties of the promoter of the target gene influence binding site conservation. I also investigate the role of redundancy on relaxing evolutionary constraints in transcriptional networks, the role genomic position and finally if the global network properties can also account for differential binding site

and edge conservation.

The second, but not less important aim, is to combine all the data in an integrative model to verify to which extent it is possible to predict edge conservation at genomic scale.

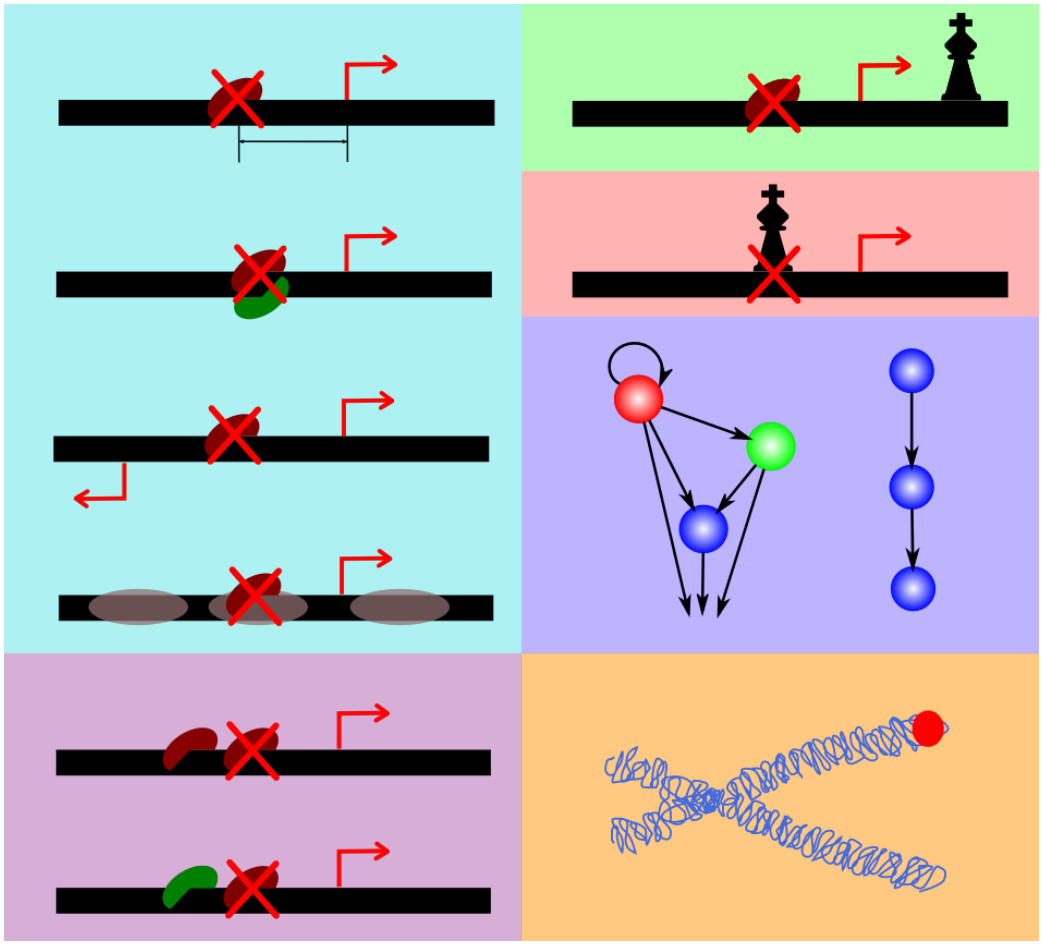

Figure 2.1: The putative determinants of transcription factor binding sites and edge importance that are considered in this study. On a light blue background architectural properties, distance from the start site overlapping binding sites divergent promoter and nucleosome occupancy of the promoter; on a violet background redundancy at level of binding site in regulatory edge and number of transcription factor targeting a gene; on a green background importance of the target gene; on red background importance of the regulator; on blue background network properties: hierachy of the regulator and edges that target regulators; on orange background genomic location

Chapter 3

Results

3.1 Defining TF binding site and transcriptional network edge conservation within and between species

To distinguish sets of functionally important binding sites and regulatory interactions I used an evolutionary approach. Binding sites and regulatory interactions that are conserved both within and between species will be enriched for the most functionally important sites and interactions. I focused my analysis on binding sites defined by direct experimental evidence from large-scale chromatin immunoprecipitation analysis in *Saccharomyces cerevisiae*. This dataset consists of 19671 sites within the promoter regions of 3832 genes and defines 12012 transcriptional interactions (or 'edges' in a network). Considering conservation across 3 closely related *sensu stricto Saccharomyces* species, 5719 of these sites are conserved in at least two species (MacIsaac et al., 2006). Considering transcriptional interactions, 5503 / 12012 edges are conserved in at least two species (ie when at least one instance of a particular TF's binding site is conserved in the promoter). To examine binding site conservation within a species, I used the complete genome sequences of 36 additional S. cerevisiae strains (Liti et al., 2009). 88.9% (17489/19671) of the binding sites under consideration are exactly conserved in sequence across all strains. Further, 92% of sites are considered as functionally conserved across all strains using the same criteria of the match between a binding site instance and a transcription factor's optimal binding site preference used in the map definition process (60% of the maximum possible score of the correspondent position specific scoring matrix , Harbison et al., 2004). Using absolute binding site sequence conser-

vation, 11303/12012 (94%) of transcriptional interactions are conserved in all strains, whereas 11567/12012 (96%) of edges are predicted to be functionally conserved. Finally, to control for any possible bias in the analysis derived from the use of PSSMs I also measured the number of sequence changes per base pair in each binding site across all 36 strains. By this criterion I find that 25% of binding sites that are not conserved in sequence show less than 0.125 SNPs per bp, 50% show less than 0.143 SNPs per bp, 75% show less than 0.167 SNPs per bp.

3.2 Binding sites and interactions that regulate more important genes are more conserved

To understand the properties that distinguish functionally important binding sites and transcriptional interactions I first considered how their conservation relates to the identity of the target genes. I find that binding sites and interactions targeting genes that are required for viability (Figure 3.1 A-D) or normal growth (Figure 3.1 E-H) are more conserved, suggesting that the maintenance of these sites is under stronger purifying selection. I also find some evidence that genes that are harmful when their expression is increased (Gelperin et al., 2005; Sopko et al., 2006) have more conserved binding sites and interactions (Figure 3.1 I-J) consistent with the generally tighter regulatory control of these genes (Vavouri et al., 2009). As for gene essentiality, controlling for other possible confounding factors confirms this result (supplementary figures 6.1 and 6.2).

3.3 Regulator importance influences edge conservation more than target importance

I next asked whether the binding sites and transcriptional interactions of transcription factors required for viability are more conserved than other sites. I find that they are. Considering binding site conservation within (figure 3.2 A) and between species (figure 3.2 B), as well as edge conservation within (figure 3.2 C) and between species (figure 3.2 D) shows that essential regulators have more conserved binding sites than other TFs. This is also true when controlling for the importance of the targeted gene (figure 3.2 E,F) and other potentially confounding factors (supplementary figures 6.3, 6.4 and).

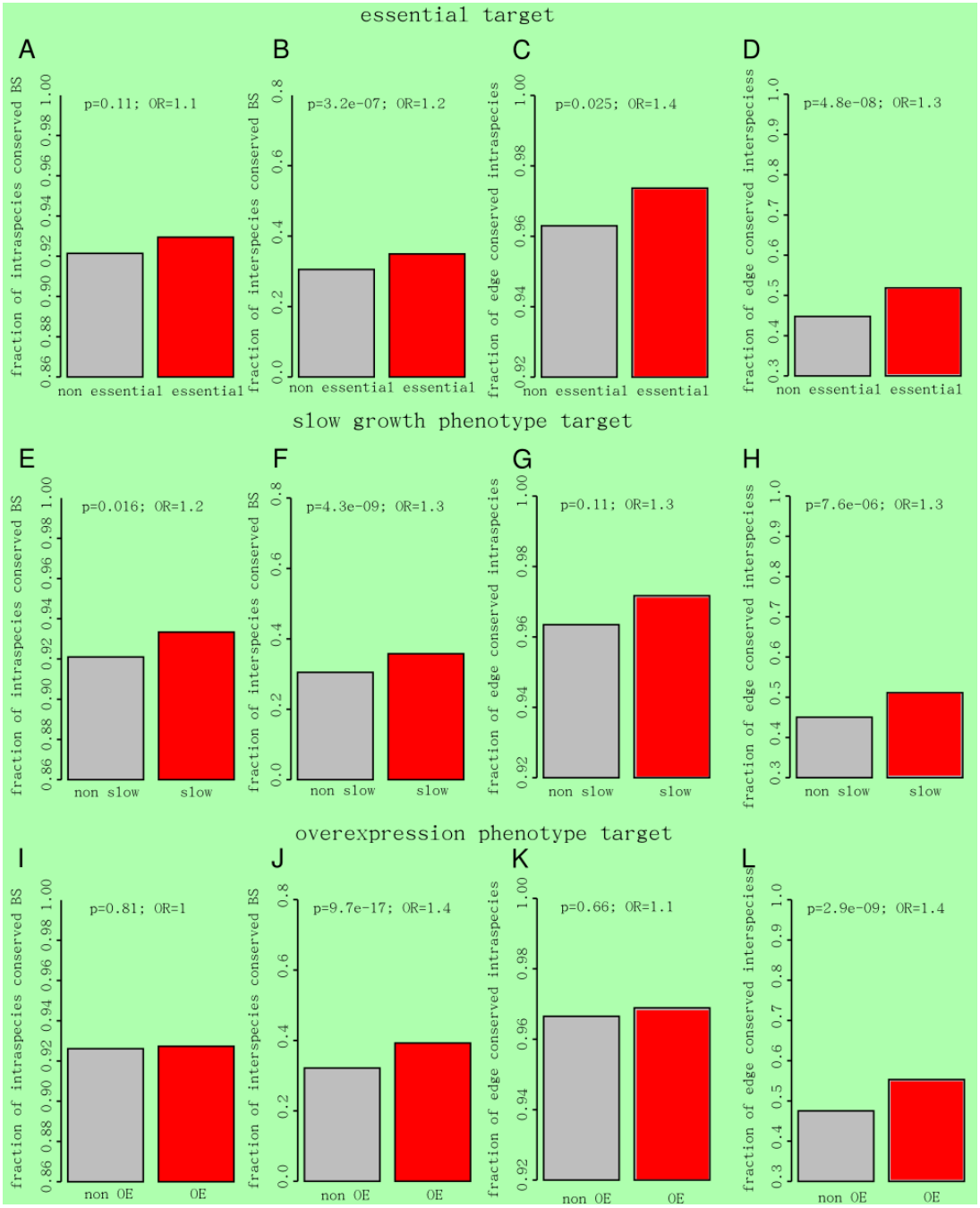

Figure 3.1: Binding sites and interactions that regulate more important genes are more conserved. Gene required for viability show more conserved binding sites within species(A), between species (B); they also show more conserved interactions within species (C) and between species (D); the same can be seen for target genes required for normal growth (E-H) and for target genes that are harmful when overexpressed (I-J); Chi square test was used to calculate p value; OR = odd ratio

Strikingly, I find that the conservation of binding sites and network edges are more strongly influenced by the importance of a regulator than of the targeted gene (compare figure 3.1 and figure 3.2). In short I can conclude that importance of an edge in a transcriptional network is related to the importance of both of the nodes that it connects, with the regulator having more of an influence than the target gene.

3.4 Design properties of the promoters alter the effects of regulatory mutations

Beyond the influence of the genes that an interaction connects, I may also expect design properties of the promoter in which a binding site is located to influence its importance. For example, considering the overall distribution of binding sites in both yeast and human shows that they are strongly biased towards the transcription initiation site (Harbison et al., 2004; Johnson et al., 2009; Tabach et al., 2007; Xie et al., 2005). This suggests that binding sites are likely of greater importance if they are nearer the start site. Consistent with this prediction, I find that binding sites and transcriptional interactions are indeed more conserved if they are located proximal to the transcription initiation site. The effect is quite strong (figure 3.3 A,B) and robust to possible confounders (supplementary figures 6.5, 6.6 and 6.7).

Promoters also differ in their nucleosome occupancy (Field et al., 2008; Tirosh and Barkai, 2008). Although considered as a whole I found no difference in the binding site conservation between promoters classified as containing a proximal nucleosome free region (depleted proximal nucleosome, DPN) and other promoters. I did notice an effect when specifically considering the nucleosome free regions of promoters. Previously, Tirosh and Barkai found that transcription factor binding sites are more conserved between species if they are found within nucleosome free regions. I confirmed this result also for intraspecies – sites located within a nucleosome free region are more conserved than other sites (figure 3.3 C). This is consistent with a model in which a nucleosome depleted region is defined specifically to facilitate functional TF-DNA interactions (Field et al., 2008; Segal et al., 2006).

Some binding sites have overlapping locations in a promoter In both flies and yeast, it has been previously reported that nucleotides located in overlapping binding sites are more conserved between species (Kim et al., 2009; Mustonen et al., 2008). I also see this effect in mine analysis, both for overlap between sites of the same binding site and for overlap between sites for different TFs, and both within and between species(figures 3.3 E,F,G,H).

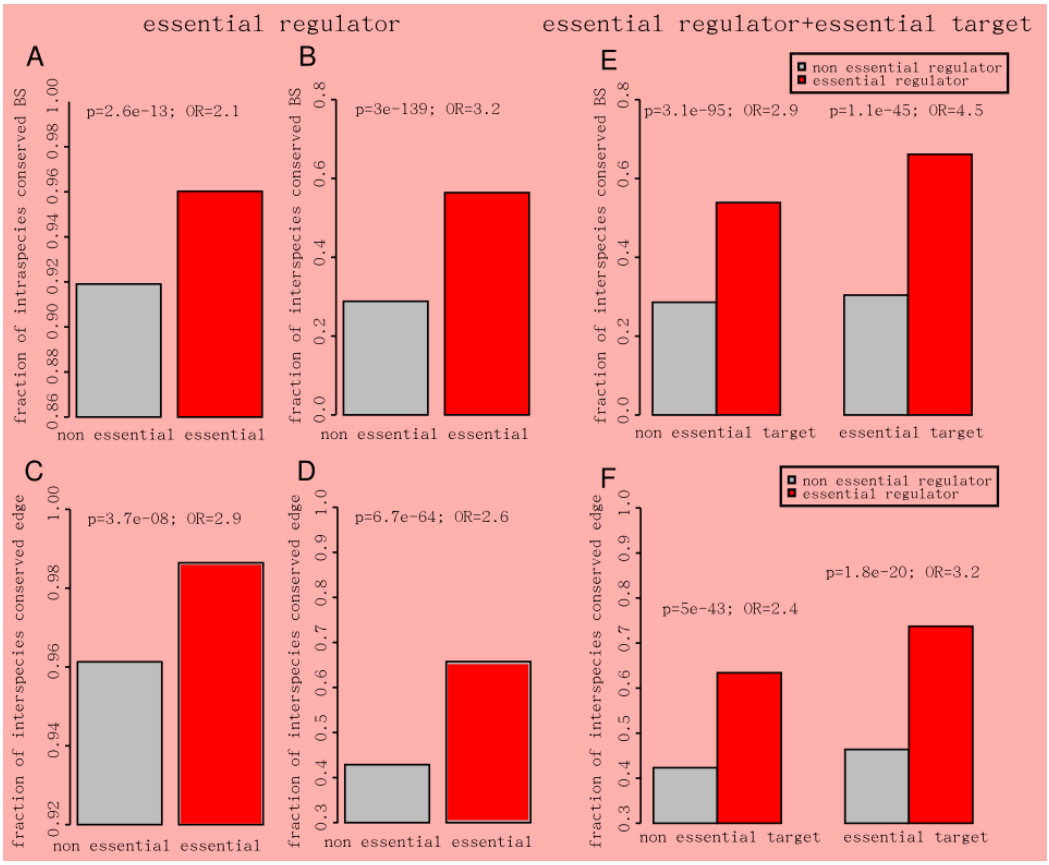

Figure 3.2: Regulator importance influences binding site importance more than target importance.Essential regulators binding sites are extremely more conserved than non essential regulator ones both intraspecies and interspecies (A,B) as well as essential regulator edges (C,D); controlling for target importance show that the effect is enhanced in essential target but is also strong for non essential target (E-F); OR = odds ratio; Fisher's test was used to test for independence.

Finally, I considered that binding sites with the potential to influence the expression of more than one gene might be more conserved than other sites. To test this, I considered the set of promoter regions located between two divergently transcribed genes – for these promoters, a binding site is likely to influence the expression of both genes. Consistent with expectation, binding sites targeting divergently transcribed promoters are more conserved within and between species (figure 3.3 I,J).

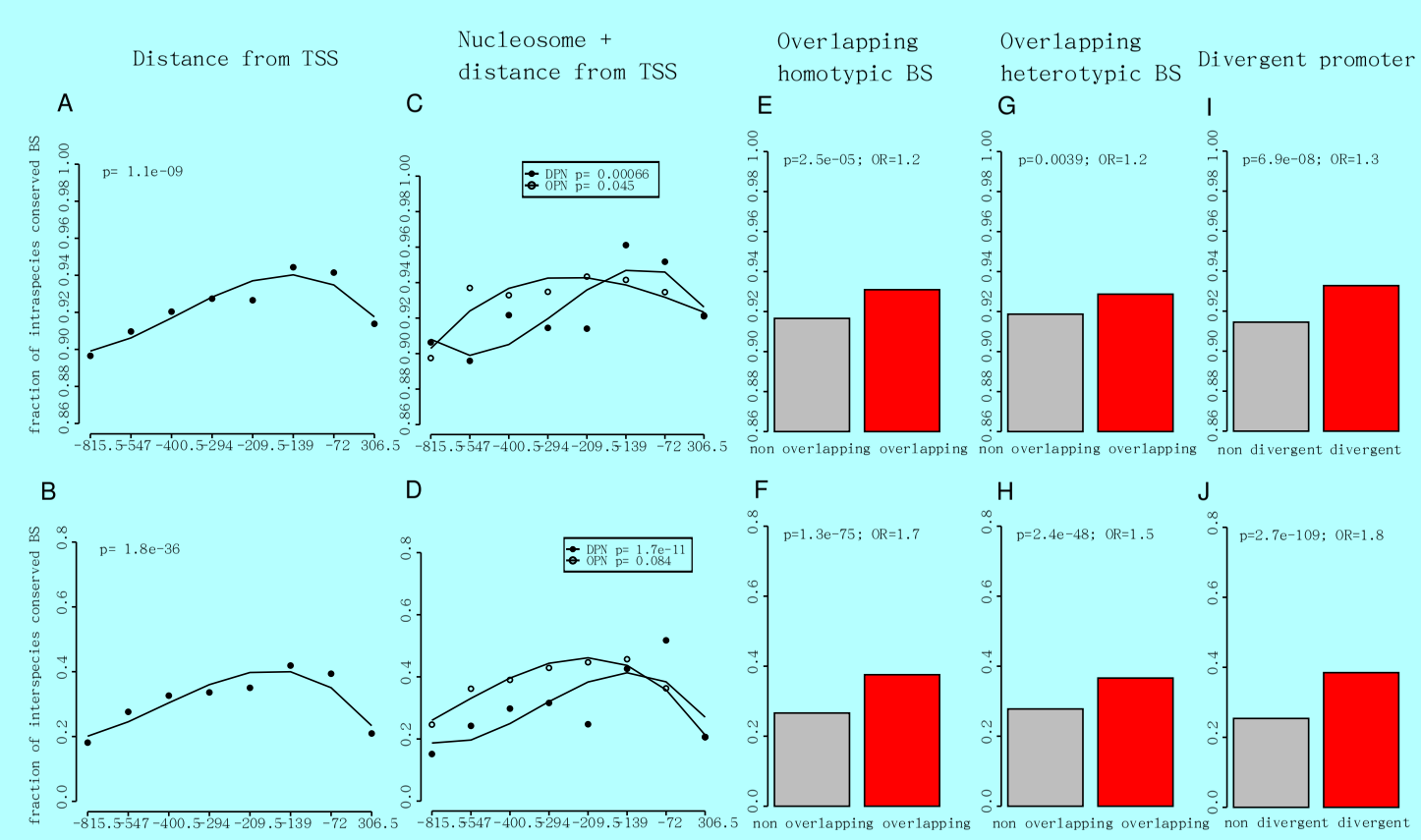

Figure 3.3: Design properties of the promoter influence binding site conservation. Binding sites closer to transcription start site are more conserved within (A) and between species (B); Promoter nucleosome occupancy also influence binding site conservation. Nucleosome free region of promoters depleted in proximal nucleosome (DPN) show higher conservation while binding sites further upstream are less conserved, while promoters without nucleosome free region (OPN) show more uniform and lower binding site conservation with respect to nucleosome free regions (C D). Homotypic overlapping binding sites show increased conservation within and between species (E-F) as well as heterotypic ones. Binding sites in promoters of divergently transcribed genes show increased conservation both inter and intraspecies (I-J). Chi square test was used to calculate p value; OR = odd ratio.

3.5 Redundancy in transcriptional networks

It is well established that gene duplication, by creating functional redundancy, relaxes selective constraints (Ohno et al., 1968). For regulatory interactions it is not clear that this should also be the case. Although multiple copies of a binding site in a promoter creates the potential for redundancy, multiple copies may also exist for functional reasons, e.g. to alter the sensitivity or co-operativity of a transcriptional response (Segal et al., 2008; Zeiser et al., 2006) or the dynamic range of the response (Giorgetti et al., 2010). Whereas redundancy predicts that multiple binding sites will be associated with lower conservation, a functional importance of multiple binding sites predicts no effect on conservation or indeed a higher conservation. Comparing cases where a single binding site for a particular TF is found in a promoter to cases where multiple sites are found, I find that multiple sites are associated with reduced conservation (figure 3.4 A,B). This result is upheld when accounting for variation in the total number of sites in a promoter up to about 6-12 binding sites (supplementary figure 6.8), the distance of these sites from the start site (supplementary 6.8), or TF importance (supplementary 6.8). This shows that for binding sites, just as for genes, redundancy tends to reduce the importance of individual sites, relaxing selective constraints.

I next considered the potential for redundancy at the level of regulation by different transcription factors. It is possible that when genes are regulated by multiple different TFs, some of this regulation may be (partially) redundant. This predicts that, across all genes, individual sites and regulatory interactions will be less conserved if a gene is regulated by multiple transcription factors. This is indeed the case. Both within (figure 3.4 C) and between (figure 3.4 D) species binding sites and interactions are less conserved if there is a potential for redundancy between TFs. Controlling for possible confounders upholds this result (supplementary figures 6.12, 6.13,6.14 and 6.14) Thus redundancy in transcription networks seems to exist both at the level of multiple binding sites for a particular TF, and at the level of compensation between different TFs. Similar to nodes, redundancy within and between edges in a network appears to influence their importance.

3.6 Binding sites in sub-telomeric regions show lower sequence conservation

Many TF binding sites in yeast, including very many sites located outside of promoter regions (@), are located in sub-telomeric regions. Previously it has been argued that binding in these regions may often be unlikely to occur

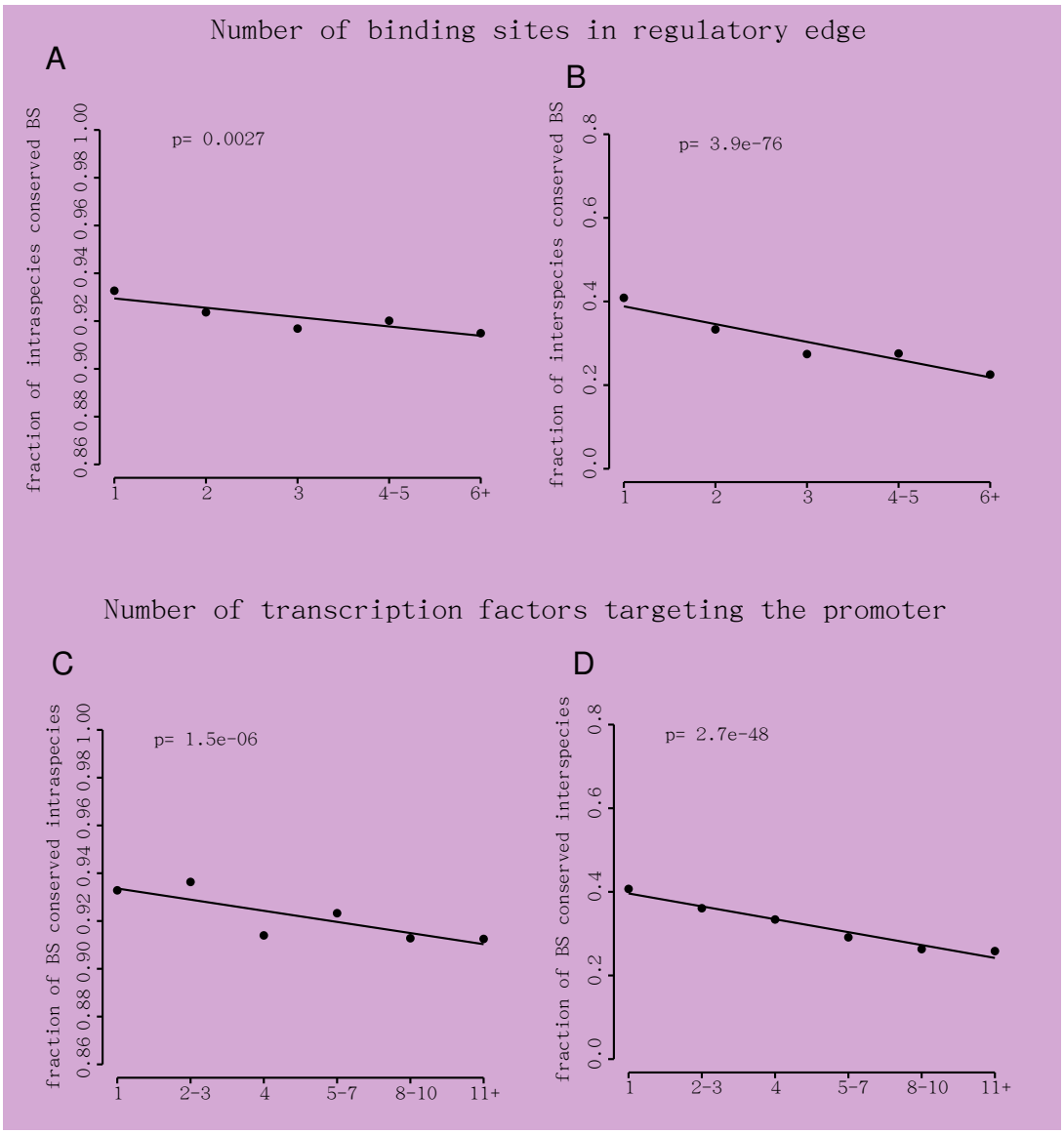

Figure 3.4: Redundancy in transcriptional networks. Binding site found in multiple copies in regulatory edges show decreased conservation supporting an important role of redundancy on relaxing evolutionary constraints on single binding sites (A-B). The role of redundancy can be seen also at the level of the number of transcription factors targeting a promoter (C-D). number of binding sites and transcription factors are binned in equal sized bins. P values are calculated using generalized linear model and analysis of deviance.

to regulate the expression of proximal genes (Balaji et al., 2006; Mak et al., 2009). Further, the sub-telomeric regions in yeast are known to have a higher mutation and rearrangement rate (Kellis et al., 2004; Teytelman et al., 2008). And are devoid of essential genes (only 2% of the subtelomeric genes compared to 18% on the total (Batada and Hurst, 2007). I therefore reasoned that binding sites in sub-telomeric regions may often be under reduced constraint compared to other sites. The number of sites targeted by essential TFs in sub-telomeric regions is very small (only 53 on 2229 (2.4%) compared to 1409 on 17442 (8.1%) in the rest of the genome OR 0.28 fisher test p ¡ 2.2e-16.). However these sites are only little less conserved than the targets of other essential TFs (Figure 3.5 A,B), showing that functionally important sites are still more conserved even if they are located in sub-telomeric regions. In contrast, I find that in general, binding sites and putative transcriptional interactions that target sub-telomeric regions are less conserved (Figure 3.5 C-F). Such a conclusions is reached even when correcting for other possible confounding factors (supplementary figures 6.16, 6.17, 6.18). Although this result may partially reflect the elevated mutation rate in sub-telomeric regions, the continued conservation of binding sites for essential genes suggests this also likely reflects that many sub-telomeric sites are of reduced functional importance.

3.7 Network characteristics influence binding site and interaction conservation

In addition to redundancy, the importance of the nodes that an interaction connects, and the architectural features of a promoter, I reasoned that features of the complete regulatory network may also influence the importance of interactions. I addressed two specific questions. First, whether the potential for errors to propagate in a network influences the conservation of binding sites and interactions. Second, transcription regulatory networks have a hierarchical structure (Jothi et al., 2009), and so the position of a TF in the hierarchy could influence the conservation of its binding sites. To address the first question, I compared the conservation of binding sites in the promoters of genes that themselves act as regulators (Segal et al., 2003). A mutation that alters the regulation of a regulator is likely to to affect not only target gene expression, but also, indirectly, multiple cellular processes that are controlled by the target gene. Consistent with this idea, I find that binding sites in the promoters of regulatory genes (both transcription factors and non non transcription factors) are more conserved within and between

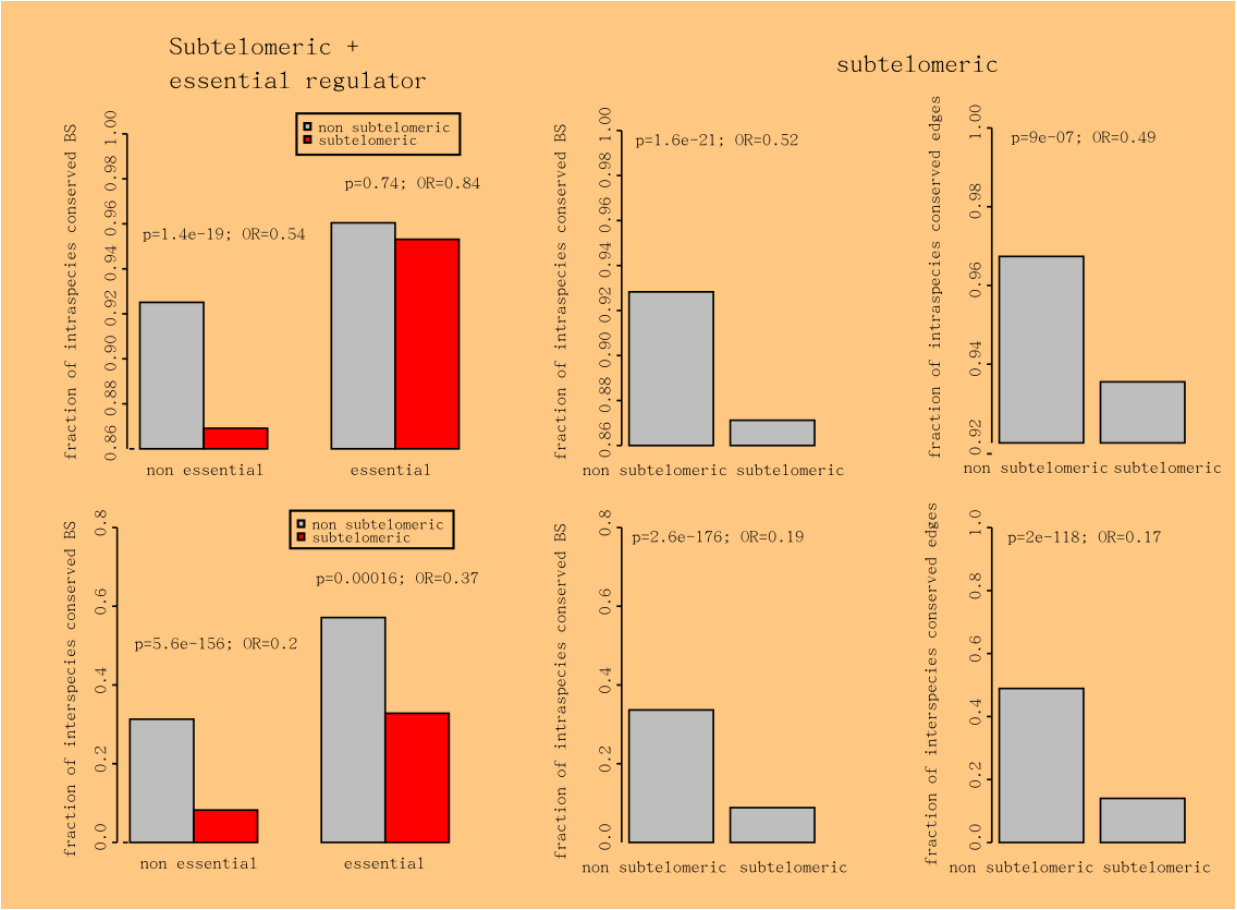

Figure 3.5: Binding sites in subtelomeric regions show lower sequence conservation In general binding sites and edges in sub-telomeric location are much less conserved than in th rest of the genome (A-D) Binding sites for essential regulators are less affected by sub-telomeric location than binding site for non essential regulators (E-F). Chi square test has been used to calculate p value. OR= odds ratio.

species (Figure 3.6 A-D). This result is upheld when accounting for other known influences, for example it is not dependent on the essentiality of the target gene or TF (supplementary figure 6.19). To address the second question, we compared the conservation of binding sites for TFs classified as in the top, middle, or bottom of the regulatory hierarchy (Jothi et al., 2009). Analyzing hierarchy effect separately for essential target and non essential target show an increased effect for the formers. Interestingly network hierarchy not only affect edge directed to regulator targets but also edge directed to non regulator target as can be seen analyzing them separately.

3.8 Predicting binding site importance across a genome

Taken together, I have identified a number of different determinants that influence the conservation of binding sites and transcription interactions within and between species. I next asked to what extent I could use this information to construct a model that predicts the conservation of binding sites across a genome. First I assessed the predictive power of each feature individually using a receiver-operating characteristic curve (ROC) analysis (fig 8, sf @). In this analysis the true positive rate is plotted against the false positive rate of the classifier at decreasing thresholds. The area under the curve (AUC) represents the probability of scoring a randomly selected mutated binding site higher than a non mutated one.

As expected due to the reduced number of mutations and the presence of segregating weakly detrimental mutations in a population, most features predict the between species conservation better than the within species conservation. However there is a general qualitative agreement between the two datasets. The most predictive determinants of binding site importance are relative to promoter architecture and redundancy, followed by regulator importance and network properties while the least predictive ones are relative to importance of the target gene.

To construct the integrated model I used a stepwise strategy: I started from a model including all the determinants found to have an effect plus their second order interactions and I excluded at each step the non significant terms starting from interactions. Categorical classification of promoters into Occupied proximal nucleosome (OPN) and Depleted Proximal nucleosome (DPN) according to nucleosome occupancy was excludes from the model because of low coverage. The final model that has been fitted to between species data, includes the significant interactions between essential regulators and over-

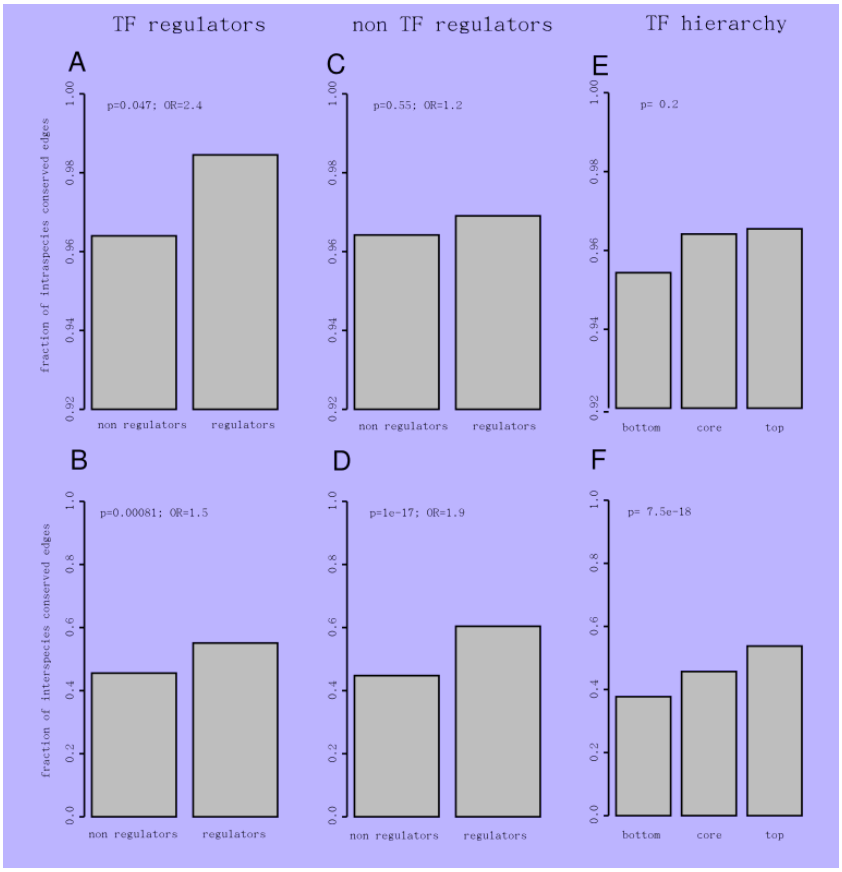

Figure 3.6: Network characteristics influence binding site and interaction conservation Edges that target regulators (both transcription factor and non transcription factor regulators) are more conserved suggesting that the potential for errors to propagate in a network increases selective pressure on the binding sites (A-D). Transcription factors higher in the regulatory network show increased conservation (E-F). Chi square test has been used to calculate p value. OR= odds ratio.

lapping binding sites and between essential regulators and the number of binding site in the regulatory edge. The predictive power of the final model trained on the interspecies data is measured using ROC AUC. The model gives an AUC of 0.685 +/- 0.005 when in prediction of interspecies binding site conservation and 0.593 +/- 0.005 when predicting intraspecies binding site conservation. Importantly training the model on intraspecies data and predicting interspecies data gives a ROC AUC performance of 0.672 +/- 0.004. This further emphasizes the robustness of our findings and the reliability of the determinants in predicting importance of binding sites.

This result show that, despite the complexity of evolutionary dynamics in regulatory regions, it is possible to predict binding sites perturbation during evolution analyzing a few simple properties that allow to understand the importance of binding sites and regulatory interactions in transcriptional networks.

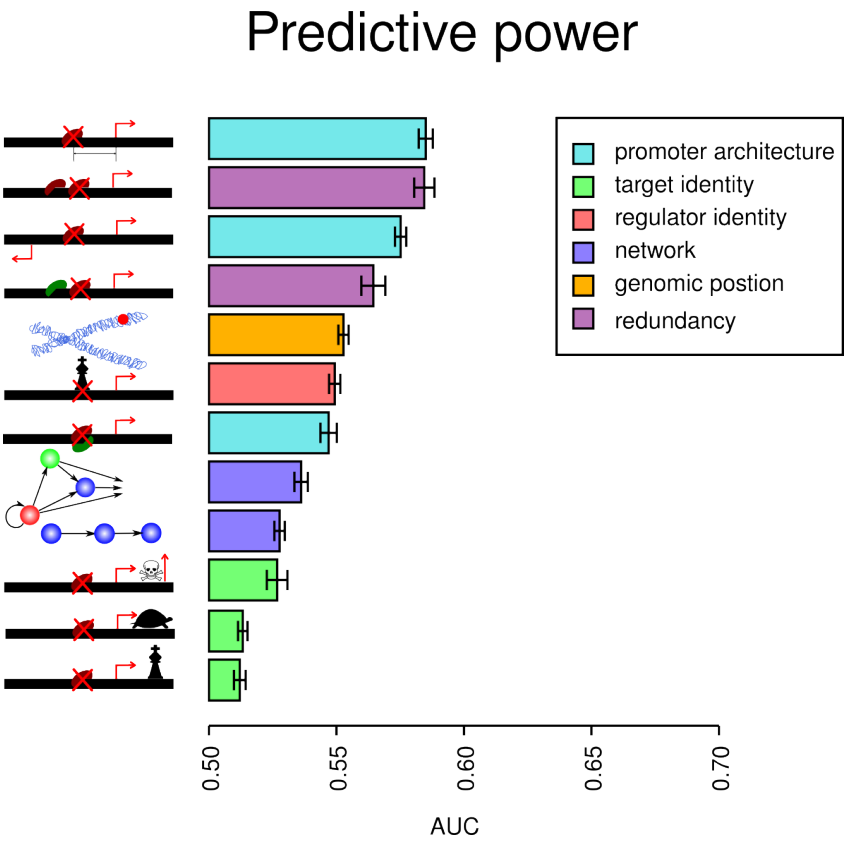

Figure 3.7: Predictive power of single determinants on binding site conservation interspecies. The most predictive determinants of binding site importance are relative to promoter architecture and redundancy while the last predictive ones are relative to target importance. Performance were assessed using 10-fold cross-validation and receiver operating characteristic (ROC) analysis. The plot represents mean and standard error of the area under the ROC curve. ROCAUC

Chapter 4

Discussion

With this systematic analysis on natural genetic perturbation in yeast, I give insight on which are the properties that determine selective pressure and evolutionary conservation of regulatory interactions and transcription factor binding sites in transcriptional regulatory networks. In particular I find that importance of regulatory edges depends on importance of target gene but unexpectedly importance of regulator affects edge importance more than the target gene. I show how design properties of the promoter strongly influence binding sites importance. In particular, binding site conservation strongly increases at decreasing distance from the transcription start site and it is increased in divergent promoters and for overlapping binding sites. As gene duplication provide functional redundancy and relax evolutionary constraint on the single genes, I show that redundancy both at level of binding site within a regulatory edge and at the level of transcription factors has a major role in relaxing evolutionary constraints on single binding sites. Genomic location location also have a profound effect: I find that binding sites in sub-telomeric regions are far less conserved reflecting both a lower functional constraint and increased mutation rate.

Importantly this analysis also underlines that as for network node also global properties of the regulatory network have to be taken into account to understand regulatory edges importance. First edges that potentially influence a wide portion of the regulatory network, such as edges whose target gene is itself a regulator, are more conserved. Second transcription factors higher in the regulatory network hierarchy also show more conserved edges.

With the integration of all these determinants together in a global model I show that is possible to predict binding site importance across an eukaryotic genome. The model performs well in predicting edge perturbations at inter-species level and show less performance when used to predict intraspecies

conservation. This is likely due to the short time scale which has two effects : first a reduced number of mutations also in neutrally evolving sites, second the persistence in a population of weakly deleterious mutations that have not been purged by purifying selection. However training the model on intraspecies data and predicting interspecies conservation show a performance level comparable to the model trained on interspecies data. This result demonstrates the true biological significance of the determinants we analyzed. An important point I want to underscore is that despite this analysis is focused on binding site conservation, the information I learnt can also be used to understand which newly acquired binding sites in promoters are more likely to be functional important. This model represents a first attempt predict binding site importance and some determinants are still probably lacking in this analysis. Nonetheless this work represents a proof of principle that it is possible to predict which binding sites are more likely to have a impact on organism fitness when lost or gained.

With the new sequencing technologies an enormous quantity of genomic data is going to be available at levels of single individuals in populations. The challenge is now to interpret all this variation and understand how it maps to phenotypic variation and ultimately to predict phenotype from sequence. The problem is challenging especially for common phenotype and diseases because their genetic background is probably complex can and involve several sequence variations and a possible role of interaction among them and with the environment. Traditional purely statistical approaches may not have sufficient power to detect low effect genetic variants. An approach that combines all available functional information can overcome such limitations. The integrated model presented here follow this functional approach and I believe it will be useful for prediction of phenotypic changes due to cis regulatory variation. Given that large fraction of genetic variation within a populations is likely to affect gene regulation rather than gene function a model that predicts which regulatory changes are causative of phenotype is likely to have a substantial impact on phenotype prediction from sequence. As recently shown a large fraction of disease associated SNPs in humans are found out side coding regions. I thus believe that integrated approaches like mine will be important for the discovery of disease causing SNPs also in humans.

Chapter 5

Materials and Methods

5.1 Transcriptional regulatory network

A comprehensive genome wide map of transcription factor binding site in S cerevisiae provided by (MacIsaac et al., 2006) was used. This map derives from an improved motif discovery analysis of the same genome-wide Chromatin immunoprecipitation data for 203 yeast transcriptional regulators used in a previously published regulatory map (Harbison et al., 2004). The dataset comprise the position of binding sites for the regulators for which a confident position specific matrix model is eiher previously know from literature or it has been derived with the two motif discovery algorithm used (MacIsaac et al., 2006). The dataset with binding confidence p = 0.005 and no conservation constraints across closely related species was considered for the analysis. Promoter region was defined to be 1000 bp uspstream the transcription start site. Position of transcription start site is calculated considering UTR definition taken from (Nagalakshmi et al., 2008) when available, otherwise start codon was used. Based on this definition, binding sites were then assigned to genes if their distance from transcription start site of the nearest genes is less than 1000 bp. We excluded from the analysis ORF classified as dubious according to (SGD) A binding site can also be assigned to two promoters whether it is found closer than 1000 bp from the transcription start site of two divergently transcribed genes. A regulatory edge is defined as including all the binding sites of a specific regulator in a promoter.

5.2 Natural genetic variation affecting binding sites within a species

A recent study on natural genetic variation in yeast population (Liti et al., 2009) have been used to analyze the perturbations that naturally occurr on S. cerevisiae regulatory interactions. In this study the genomes of 36 among wild and domestic S. cerevisiae strains have been sequenced and aligned. The data was mapped to the binding site location using genome annotation downloaded from Saccharomyces Genome Database (SGD) on October 10th 2007 as in (Liti et al., 2009). Only single nucleotide polymorphisms (SNPs) with a high sequence quality confidence level (p ¡ 10 e-30) were considered for the analysis. Deletions were excluded from analysis. Binding sites that do not show any SNP in any of the strains were defined as conserved at sequence level. For each binding site sequence divergence was calculated considering the maximum number of SNPs that affects the binding site across the strains, normalized by the total number of base pairs of the BS.

5.3 Evaluating within species binding site conservation

The affinity of a binding site to the correspondent transcription factor generally correlates with its similarity to the score given by position specific scoring matrix (PSSM) model (Berg and von Hippel, 1987; Stormo and Fields, 1998). For each BS that show at least one high quality SNP its score was calculated using log-likelihood position specific scoring matrices (PSSM) models provided by (MacIsaac et al., 2006) using custom routines written in Java (See chapter Java Code) and BioJava (Holland et al., 2008). A binding site is considered to be conserved if it scores at least 60% of the maximum possible score of its correspondent PSSM model as described in supplementary materials of (Harbison et al., 2004).

5.4 Evaluating between species binding site conservation

. Between species motif conservation was evaluated using the highest conservation cut-off in the classification of binding sites conservation of (Harbison et al., 2004) and (MacIsaac et al., 2006) that correspond to a binding site score conservation of at least 60% of maximum score of the correspondent

PSSM model in at least two out of other three *sensu stricto Saccharomyces* species closely related to S. cerevisiae.

5.5 Evaluating transcriptional edge conservation

A regulatory interaction between a transcription factor and its target gene (transcription network edge) is considered as conserved if at least one of the binding sites for the transcription factor is conserved (at least 60% of maximum score) in the promoter region of the target gene.

5.6 Gene importance

Essential genes were taken from Saccharomyces cerevisiae deletion project web page (YDP). Genes that cause a slow growth phenotype when deleted were taken from (Deutschbauer et al., 2005); Genes that cause an overexpression phenotype when deleted were taken from (Gelperin et al., 2005; Sopko et al., 2006);

5.7 Nucleosome occupancy

genes with two distinct patterns of nucleosome occupancy in their promoters, Occupied Proximal Nucleosome (OPN) and Depleted Proximal Nucleosome (DPN), were taken from (Tirosh and Barkai, 2008);

5.8 Network properties

. Transcriptional regulators were classified in three hierarchical levels, top, core and bottom, following (Jothi et al., 2009). We excluded from the classification regulators that couldn't be uniquely assigned to one of these three hierarchical layers. Node in degree is defined as the number of transcription factors regulating the target gene. Genes with transcriptional regulatory activity were taken from (Segal et al., 2003);

5.9 Statistical analysis

All statistical analyses were performed in R. Chi Square test or Fisher's test were used to test for independence with categorical data, binomial generalized linear model was used to test trend significance of continuous explanatory variables. For distance effect a third degree polynomial generalized linear model has been applied.

5.10 Integrative model

Binding site conservation within species and between species was predicted fitting a binomial generalized linear model to the data. The model was developed with a stepwise strategy starting from a model including all the determinants found to have an effect and including second order interaction and excluding at each step the non significant terms starting from interaction terms. Categorical classification of promoters into Occupied proximal nucleosome (OPN) and Depleted Proximal nucleosome (DPN) according to nucleosome occupancy was excludes from the model because of low coverage. 10-fold cross validation was used to validate model. Area under the Receiver Operating Characteristic curve (ROC AUC) was used to assess the performance of the model.

Chapter 6

Supplementary figures

These figures support the result chapter.

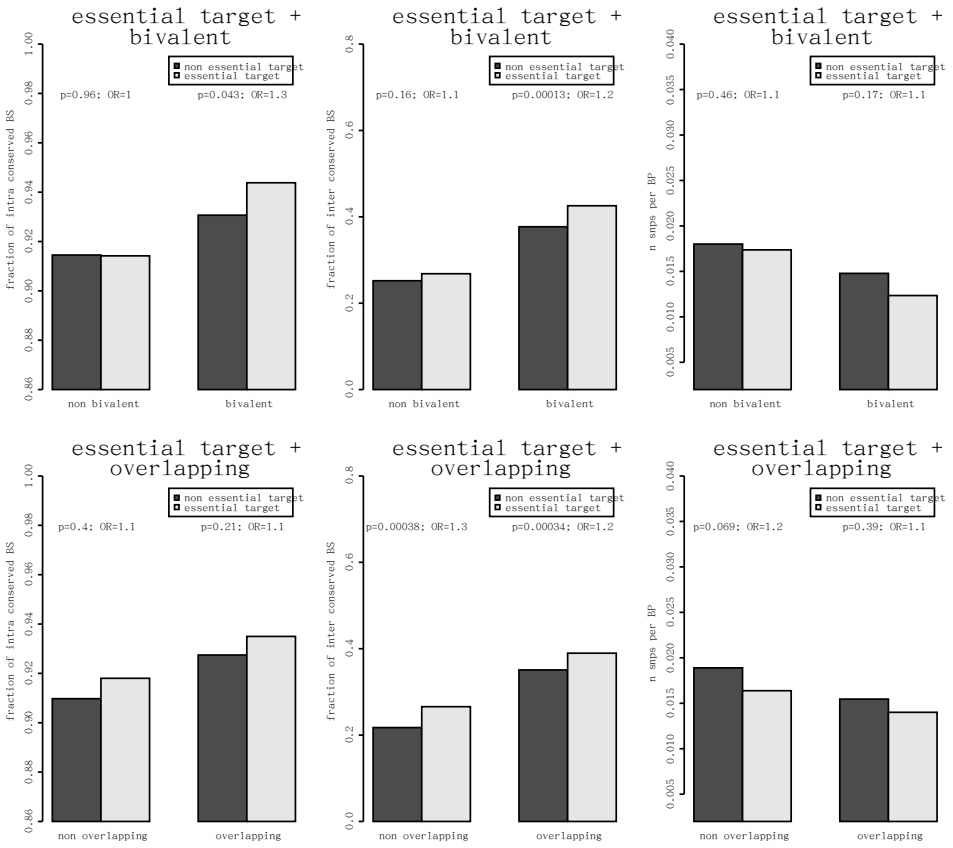

Figure 6.1: essential target interactions

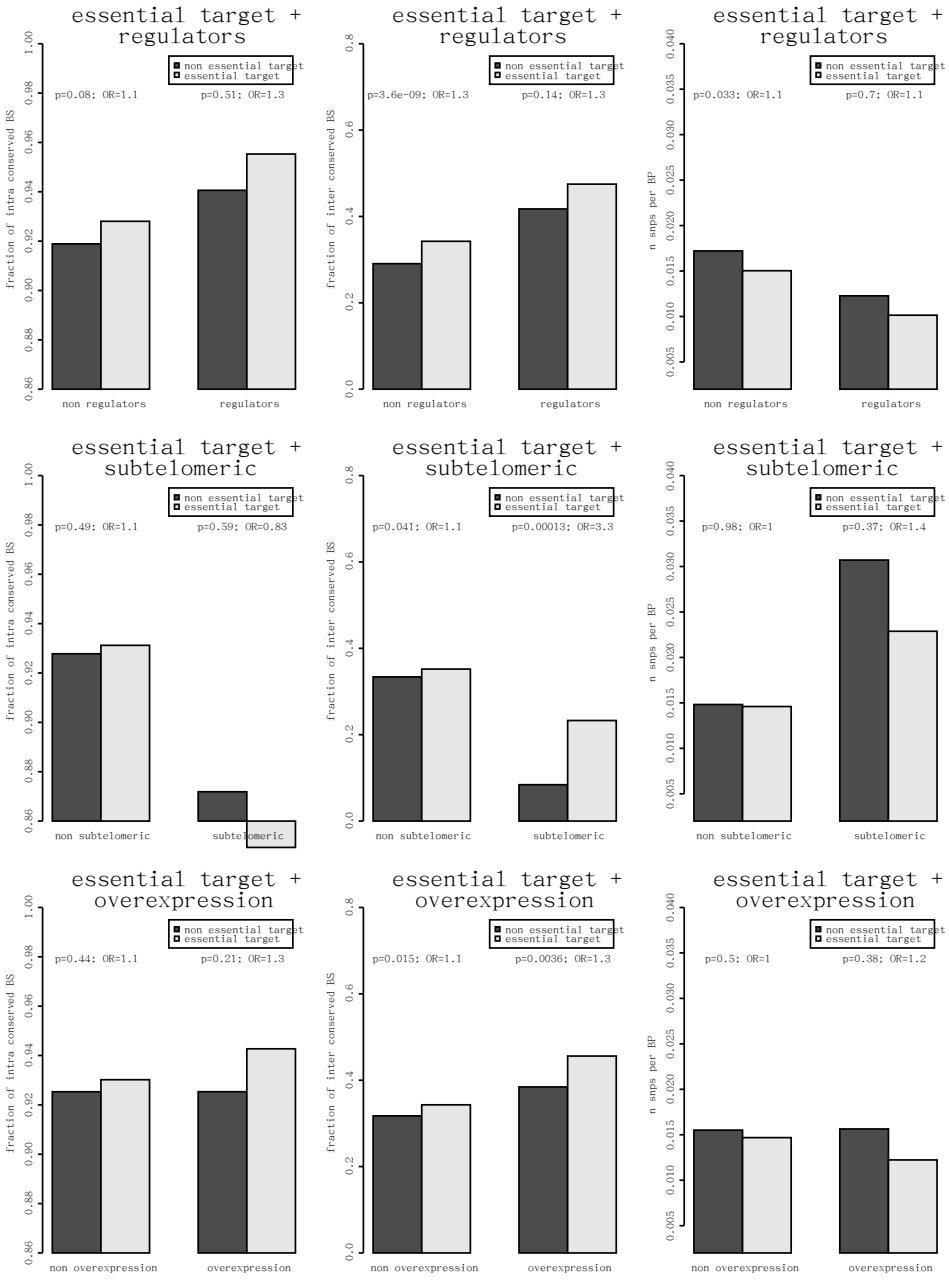

Figure 6.2: essential target interactions

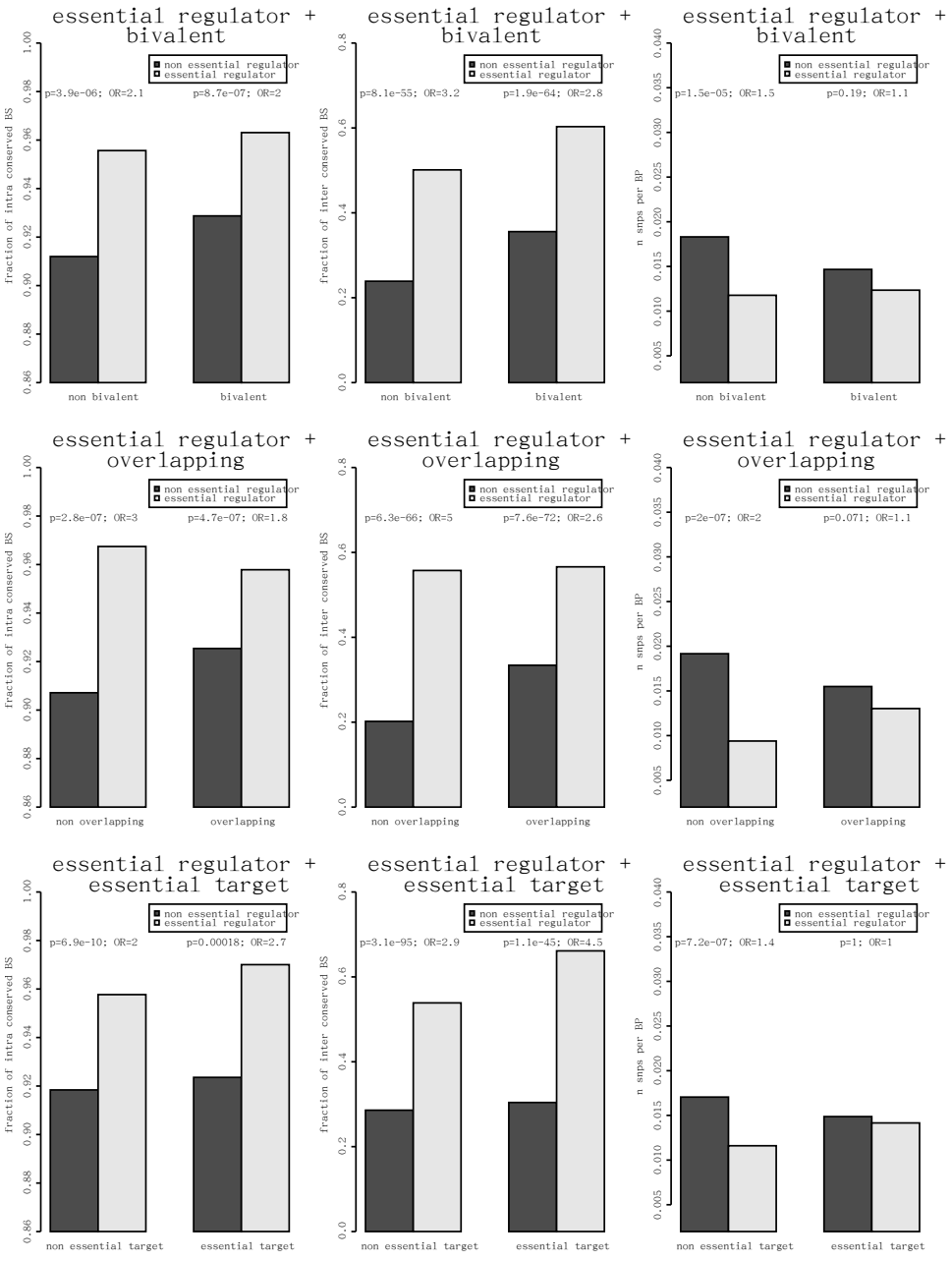

Figure 6.3: essential regulator interactions

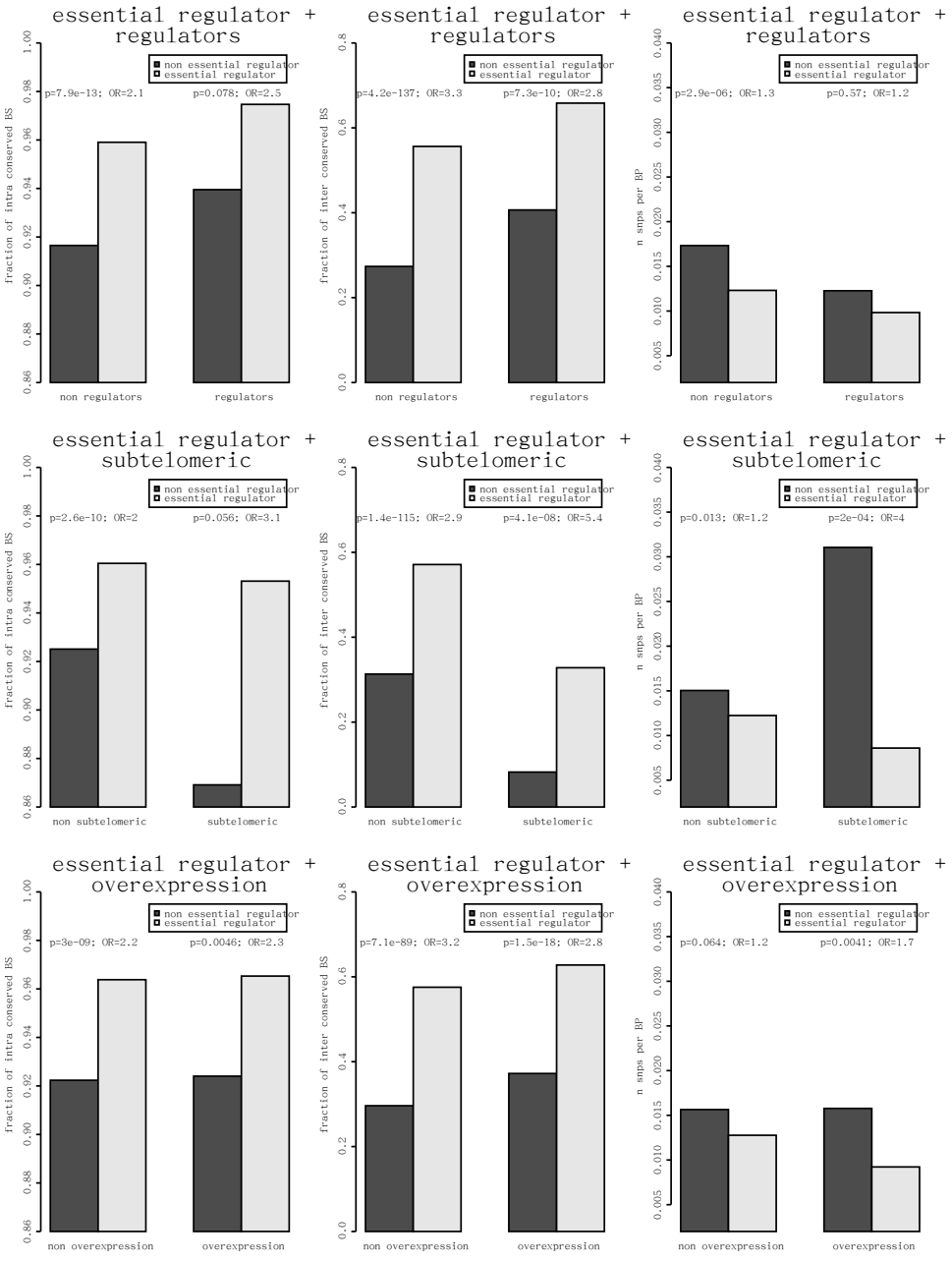

Figure 6.4: essential regulator interactions

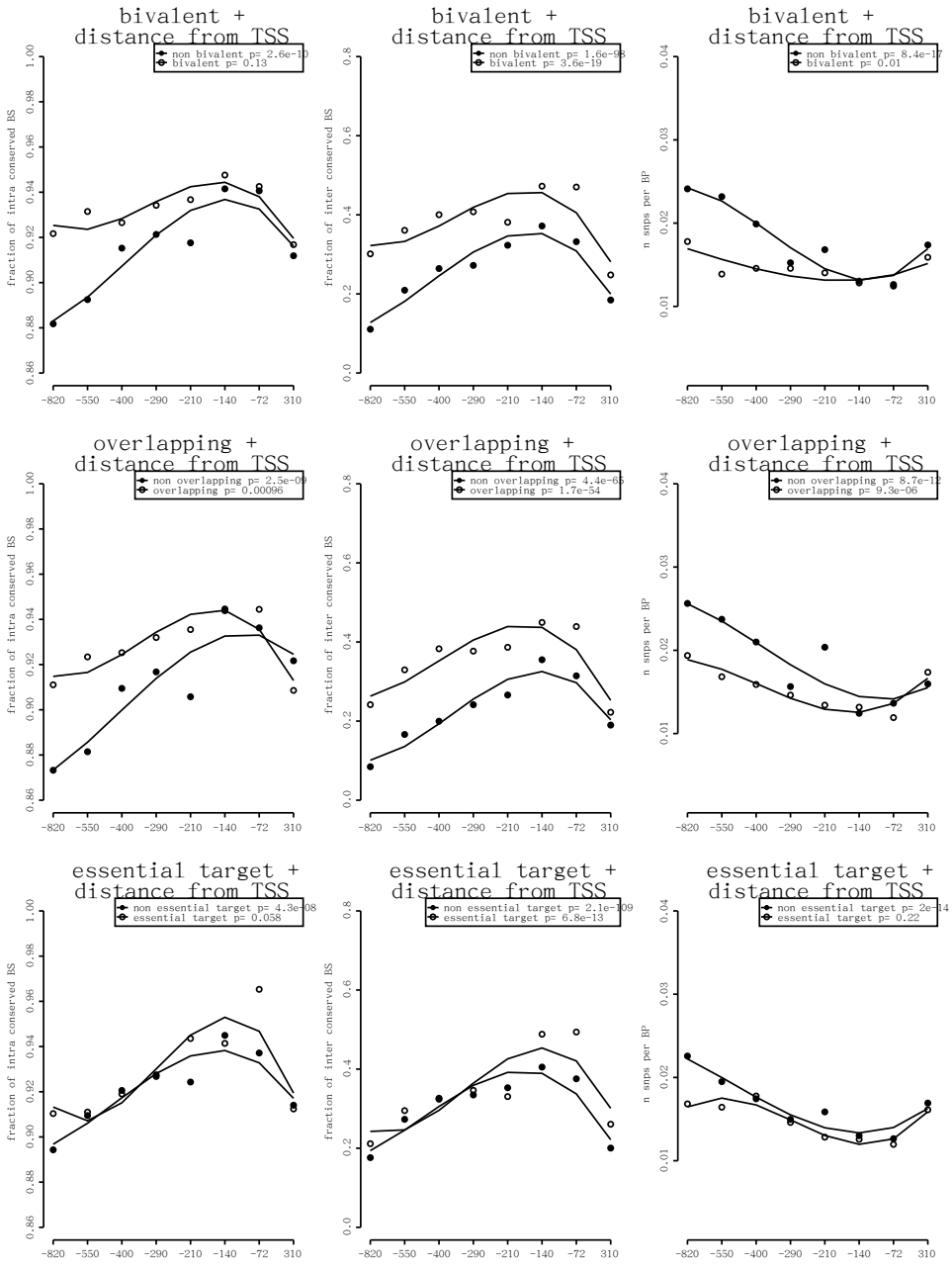

Figure 6.5: Distance interactions

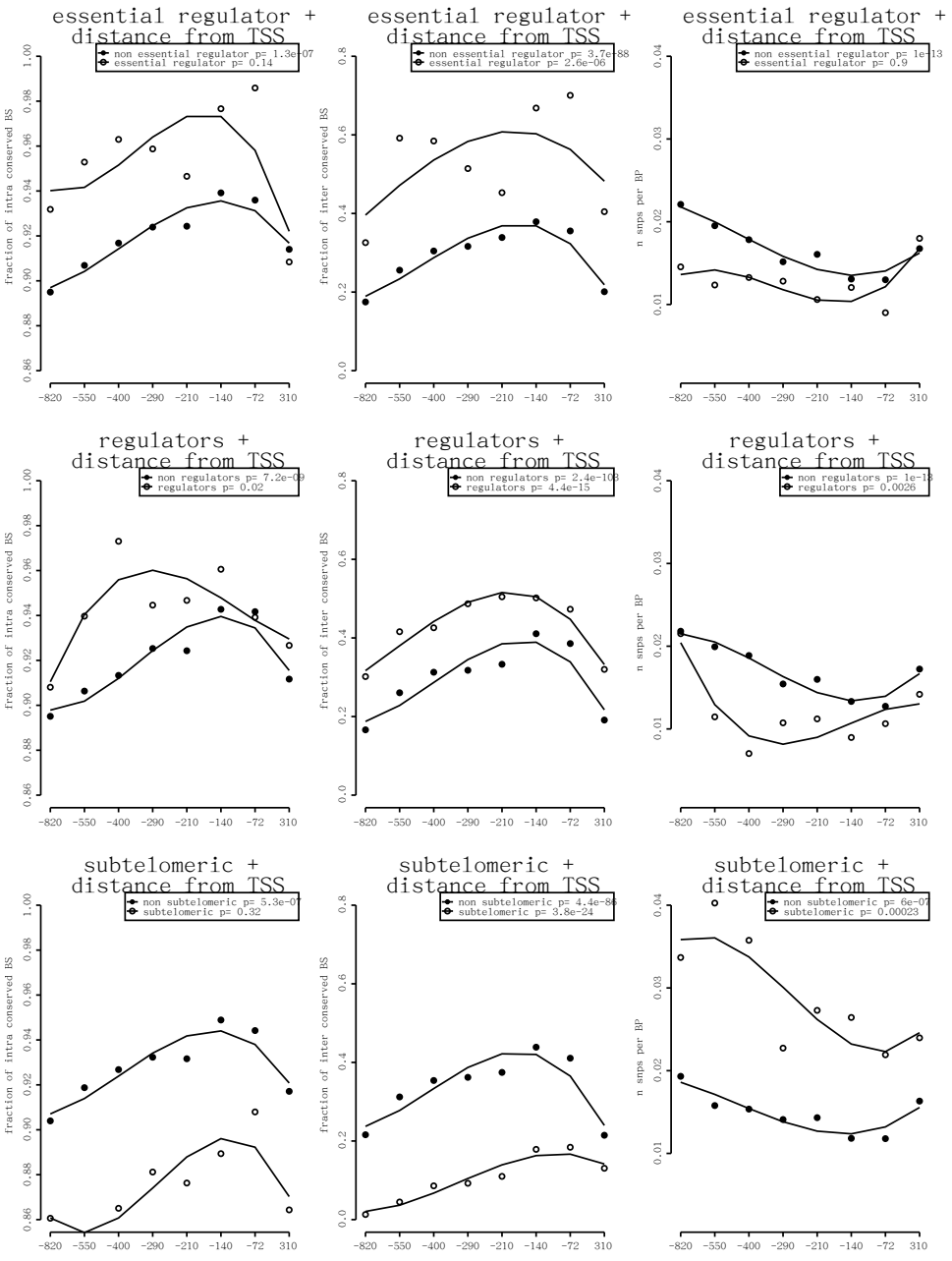

Figure 6.6: Distance interactions

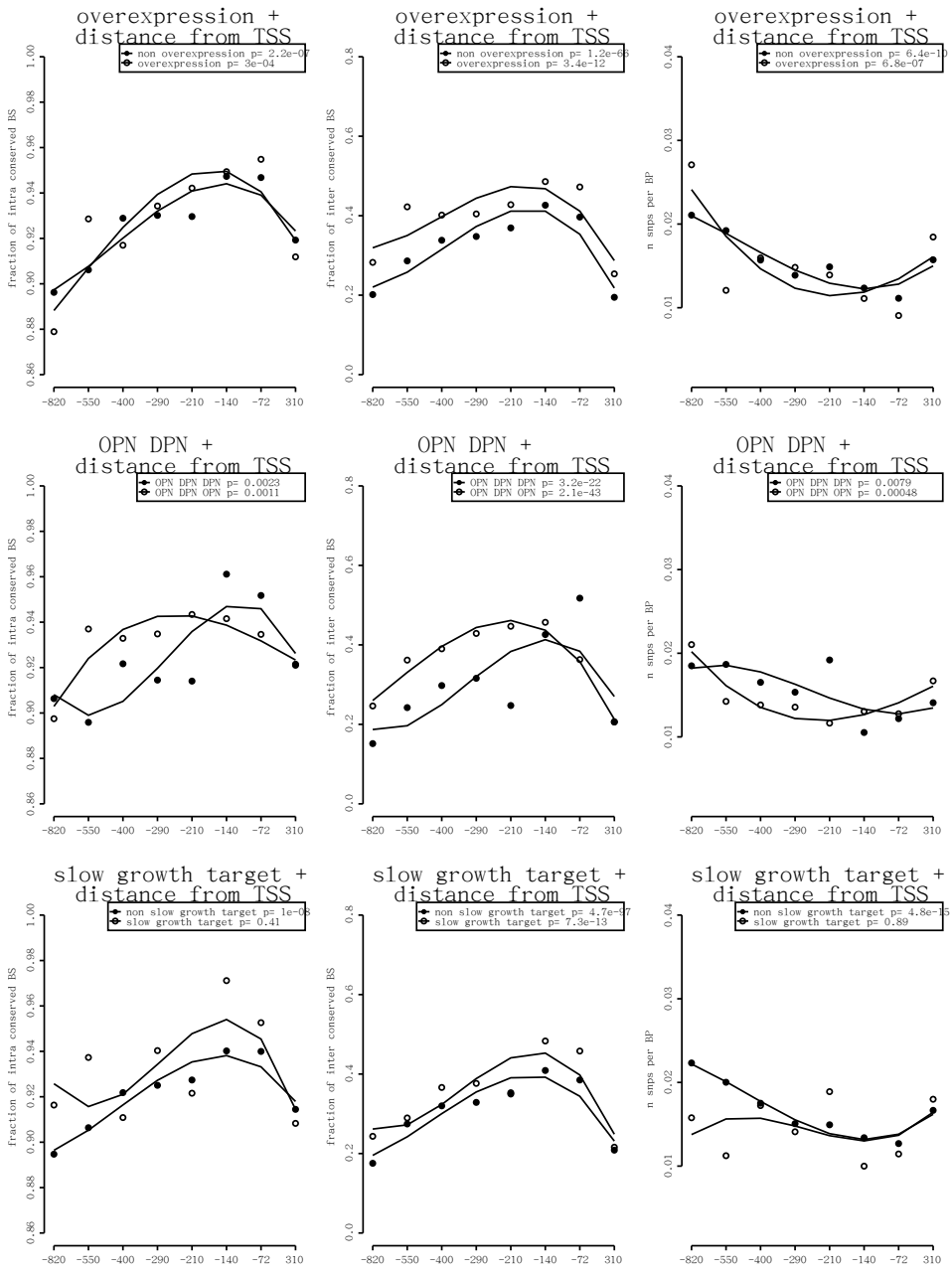

Figure 6.7: Distance interactions

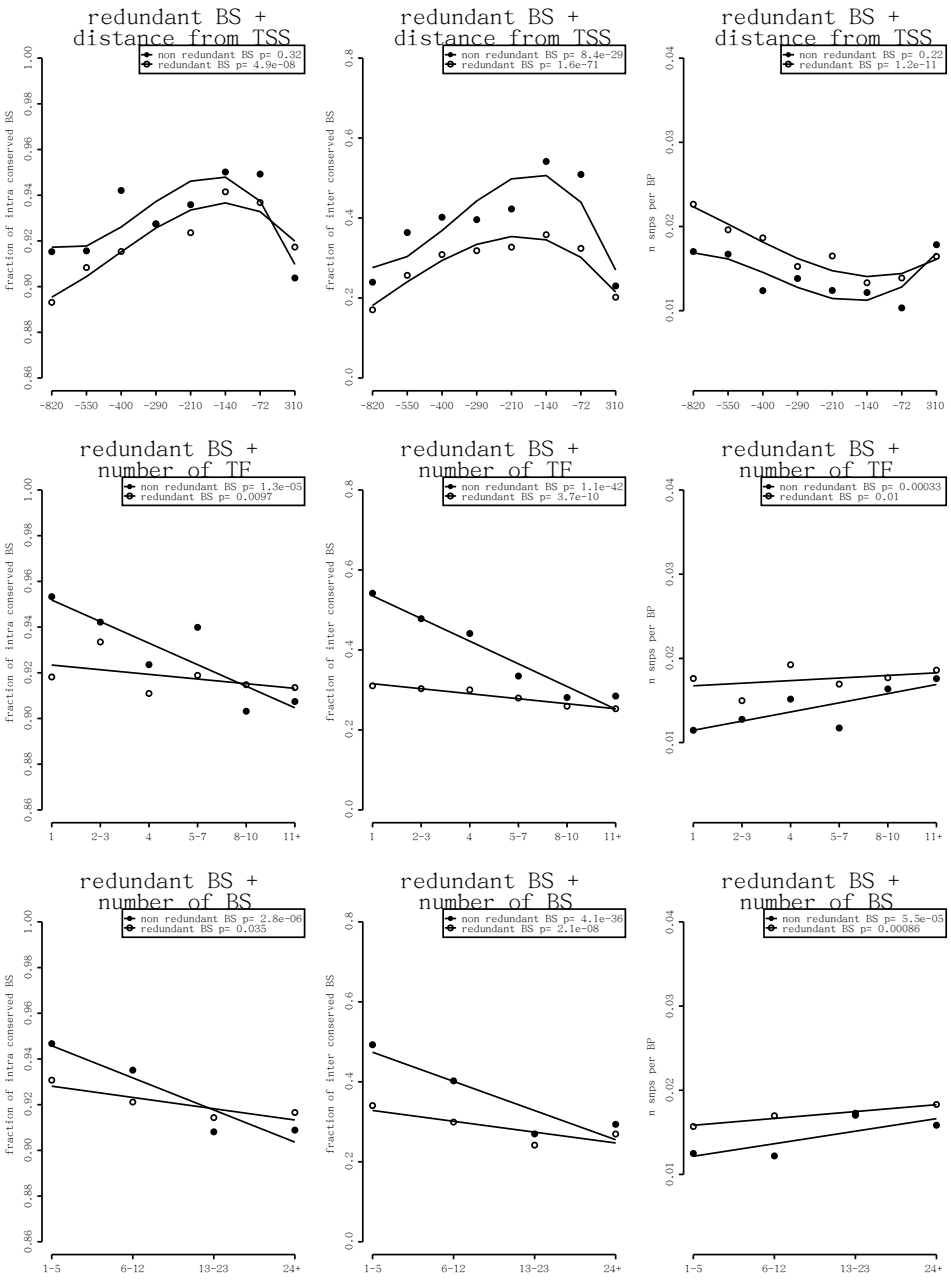

Figure 6.8: Binding site redundancy interactions

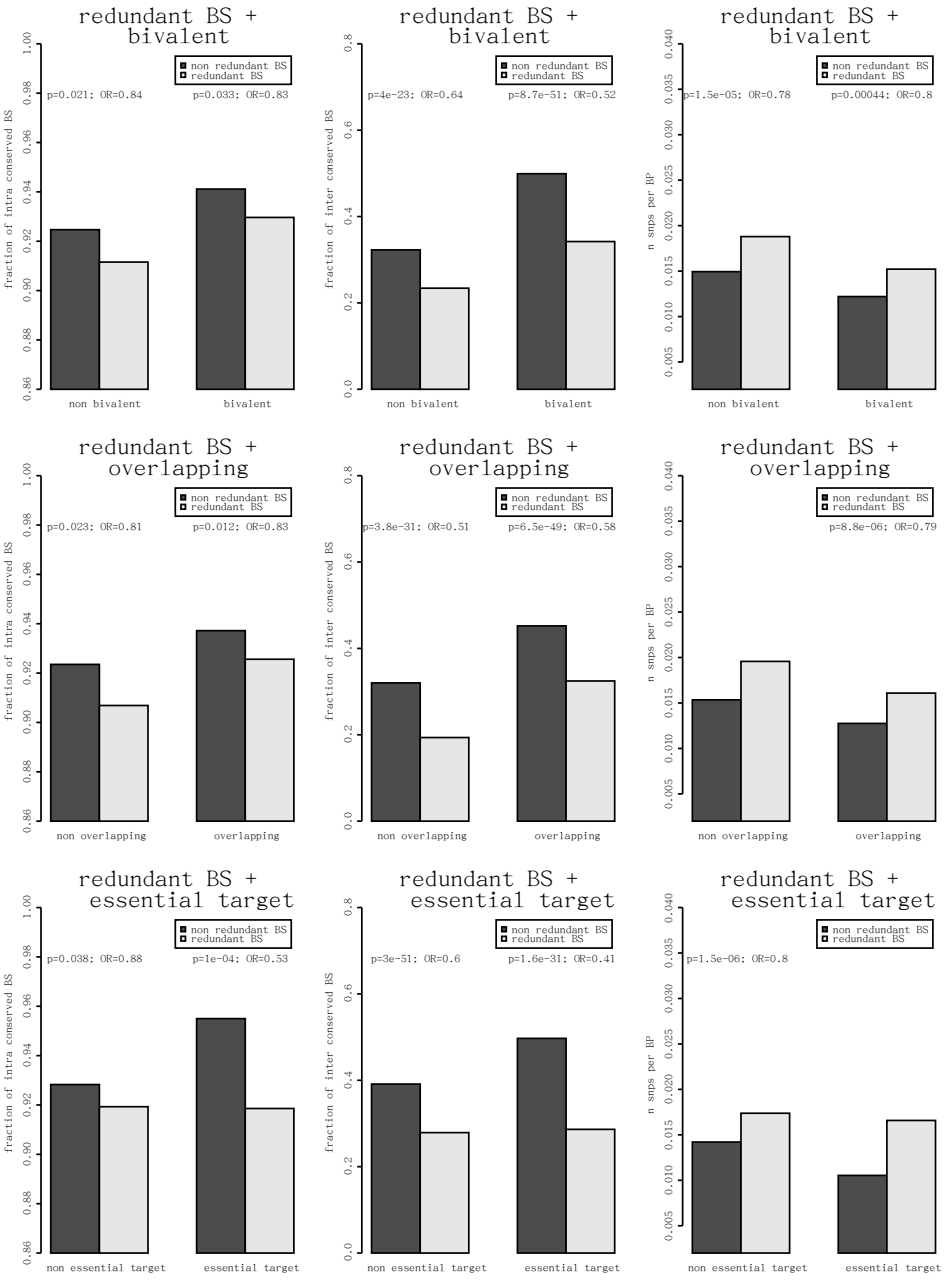

Figure 6.9: Binding site redundancy interactions

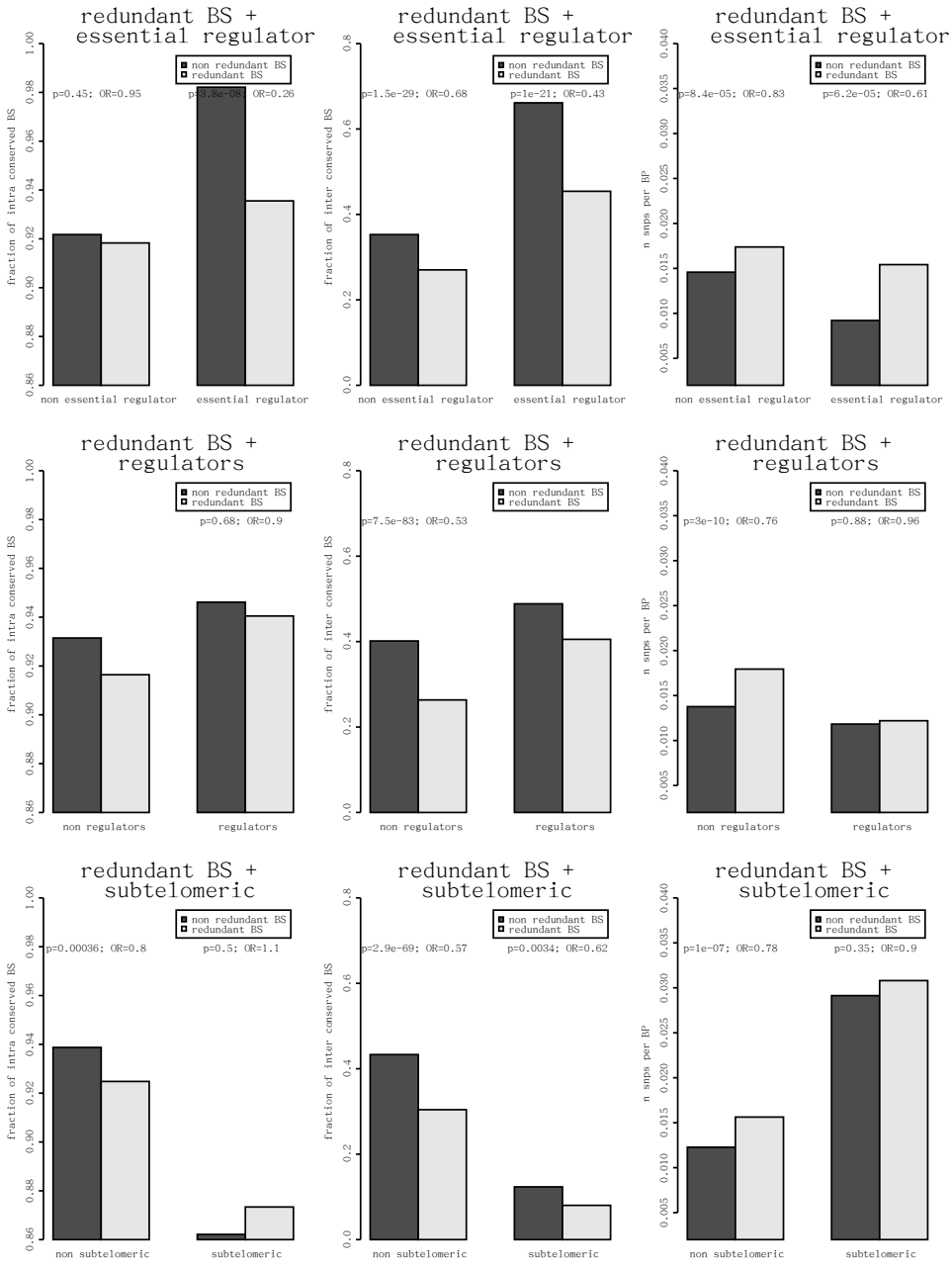

Figure 6.10: Binding site redundancy interactions

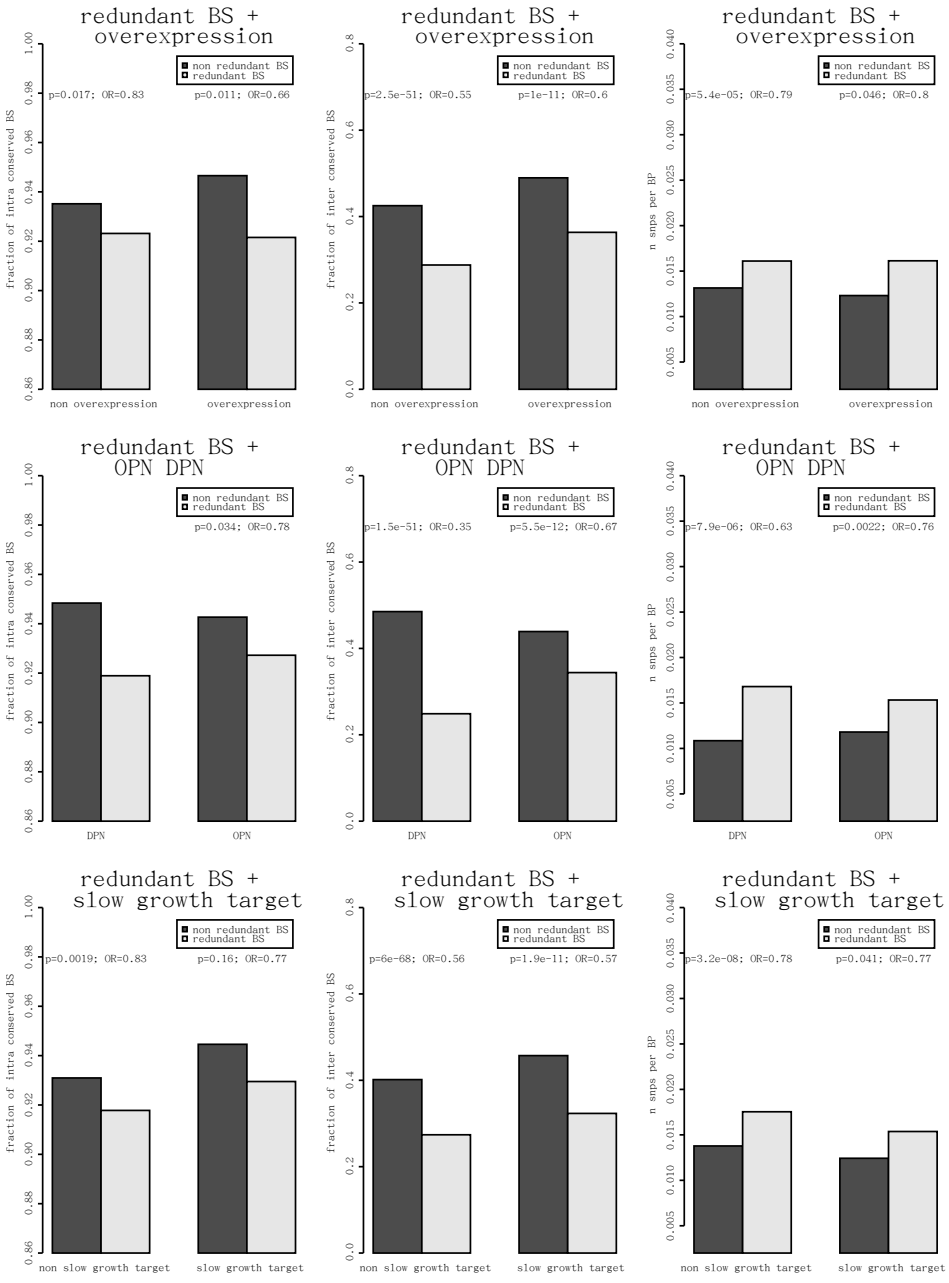

Figure 6.11: Binding site redundancy interactions

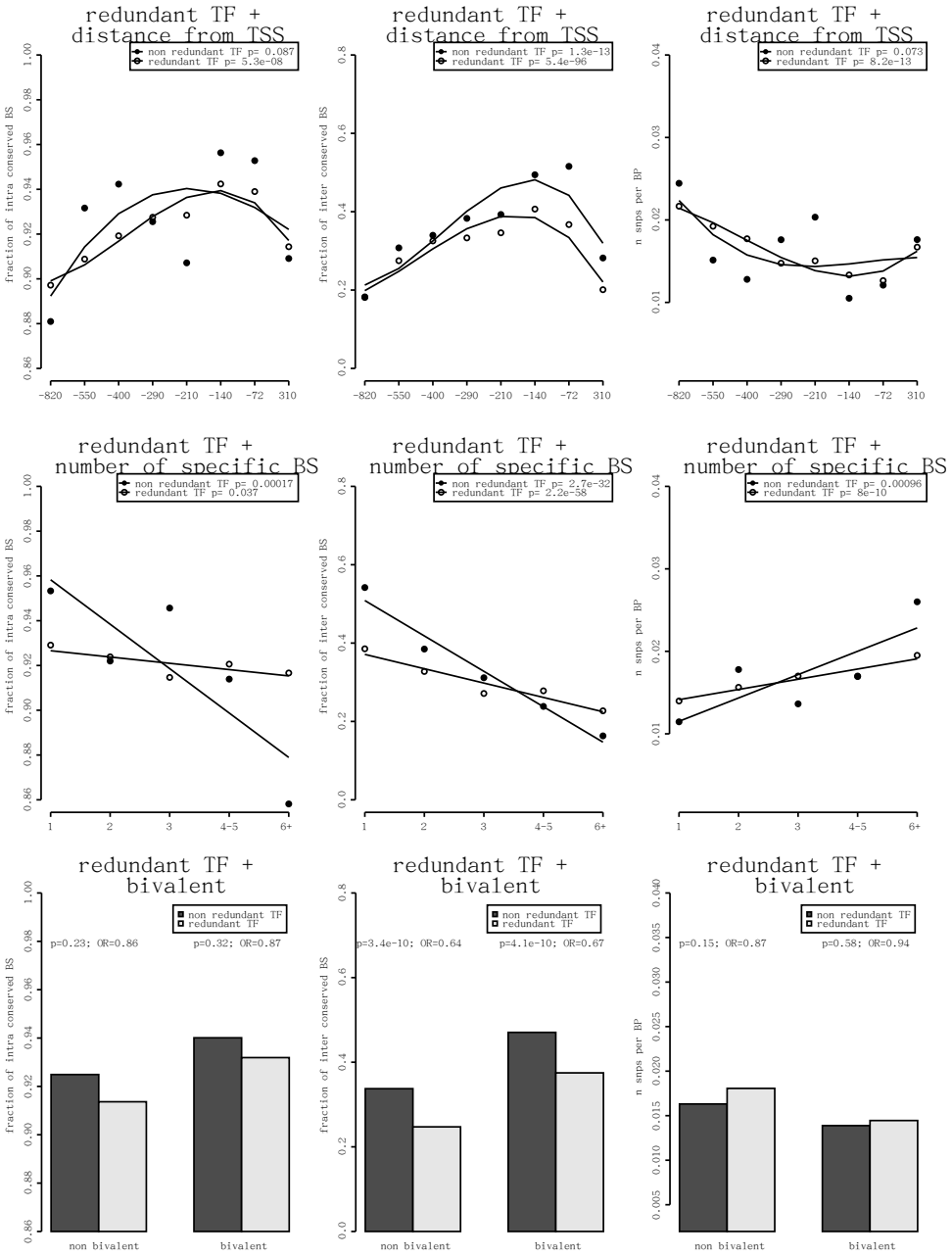

Figure 6.12: Transcription factor redundancy interactions

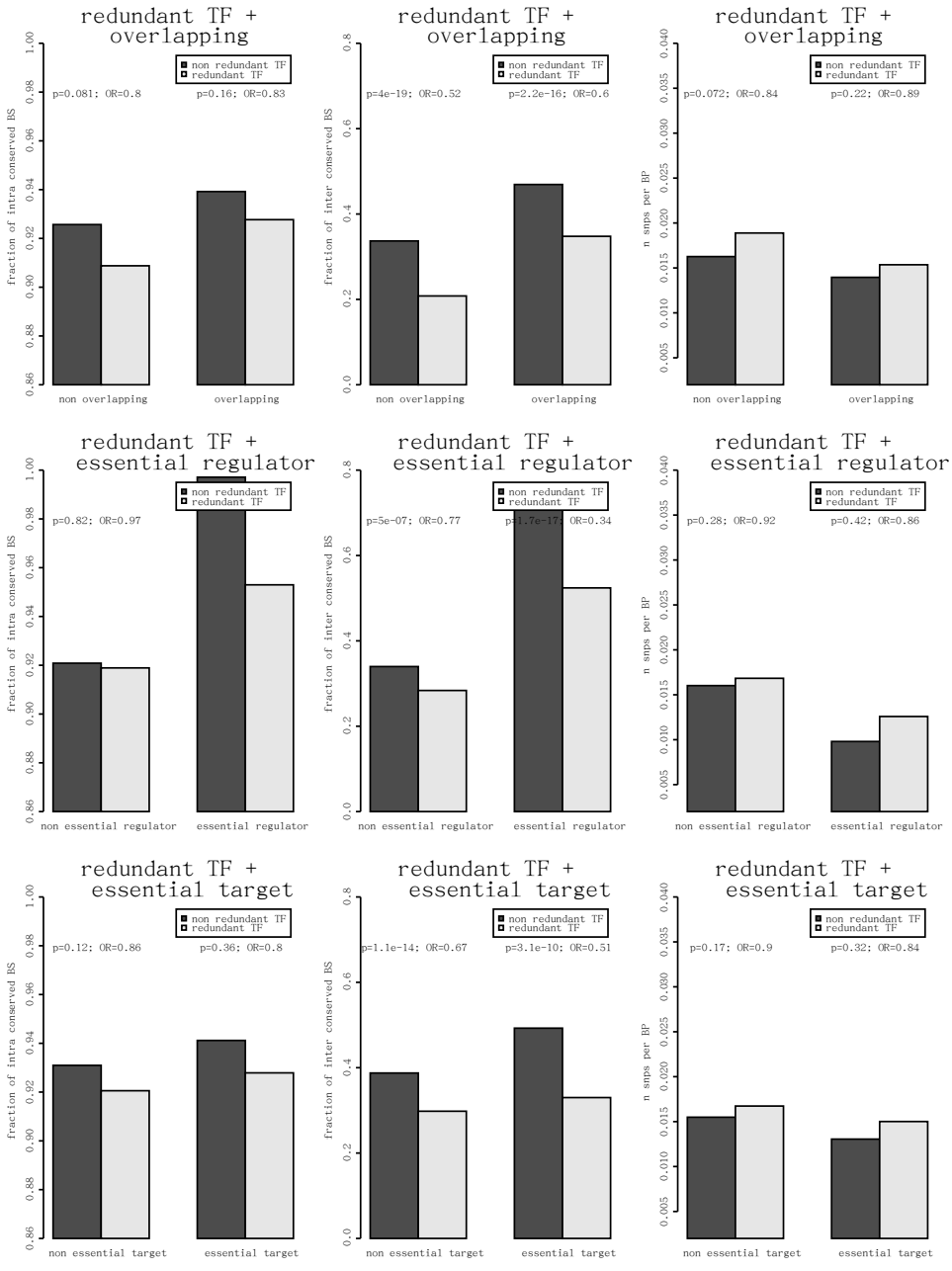

Figure 6.13: Transcription factor redundancy interactions

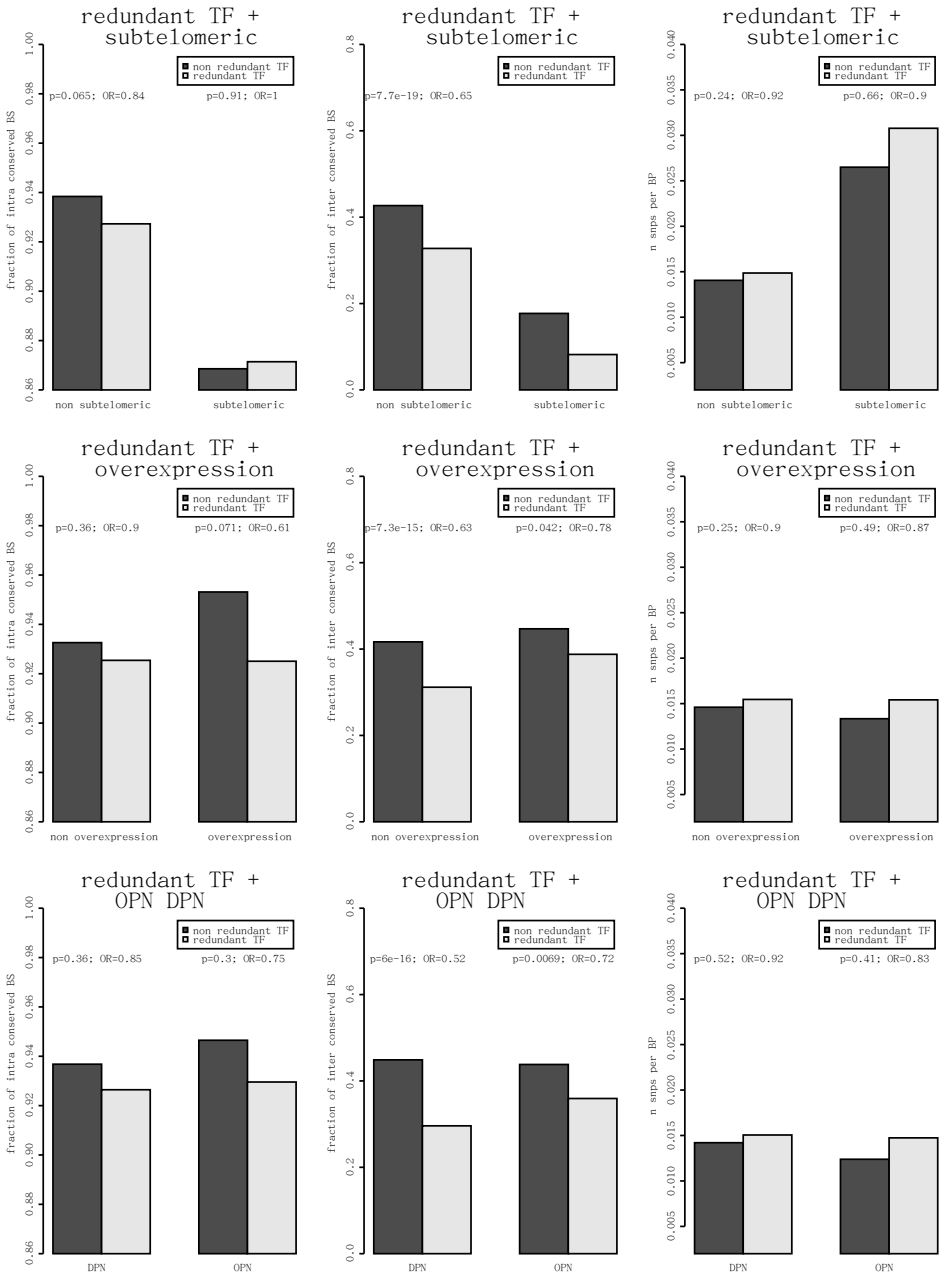

Figure 6.14: Transcription factor redundancy interactions

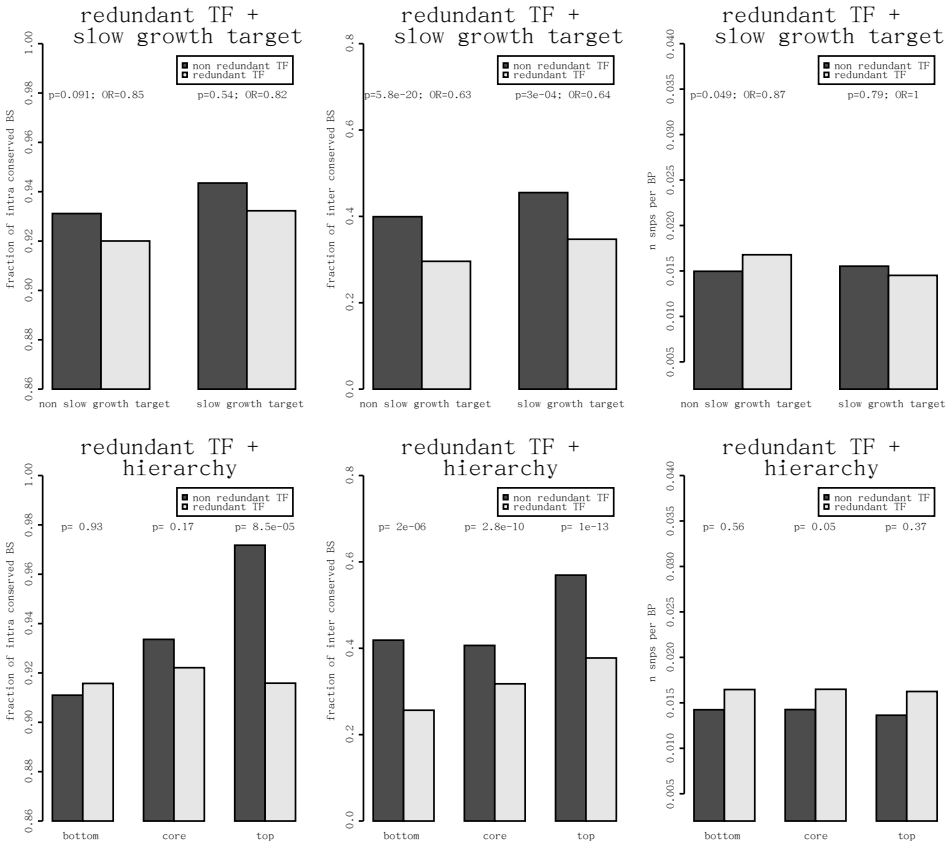

Figure 6.15: Transcription factors redundancy interactions

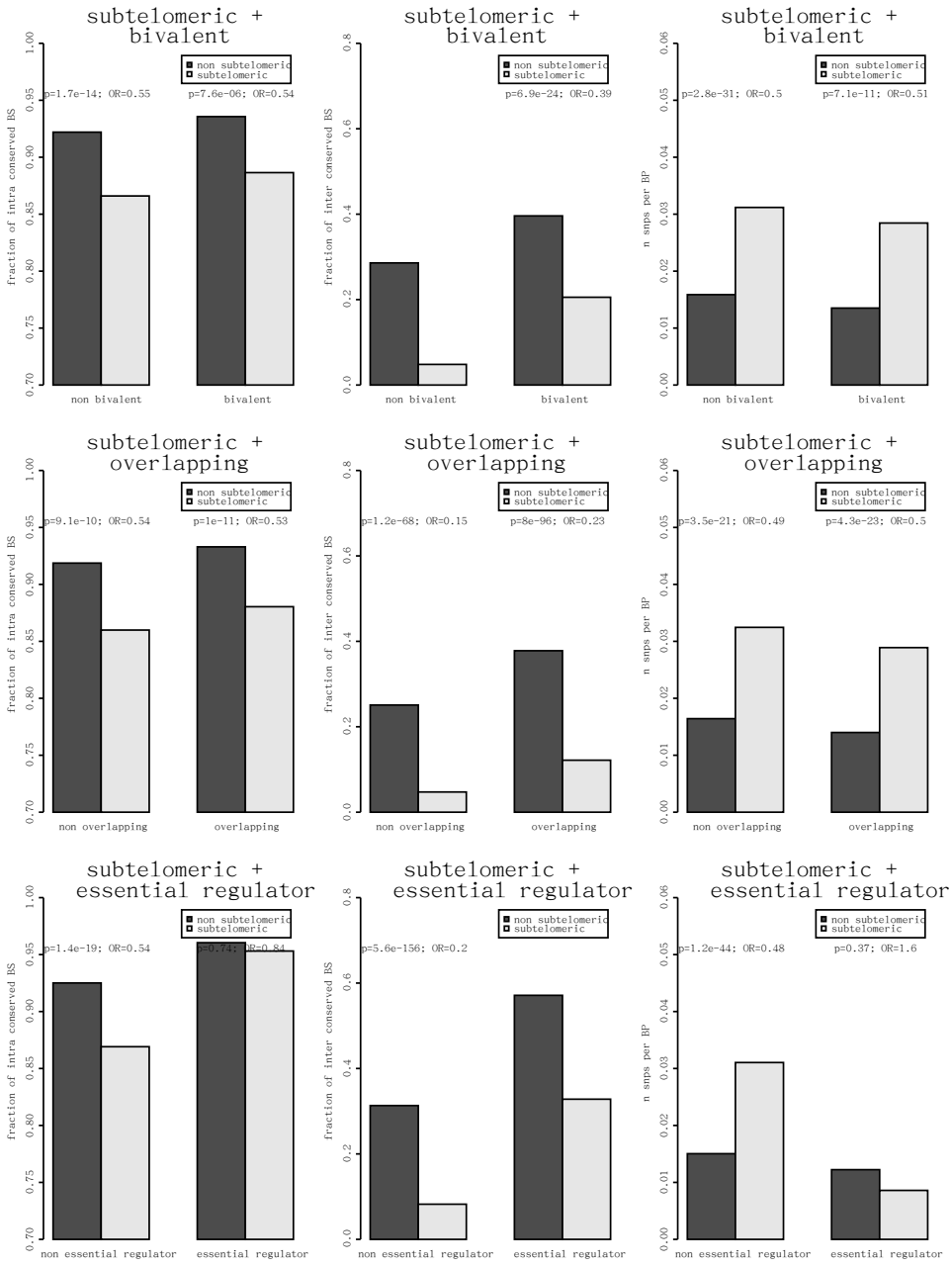

Figure 6.16: subtelomeric location interactions

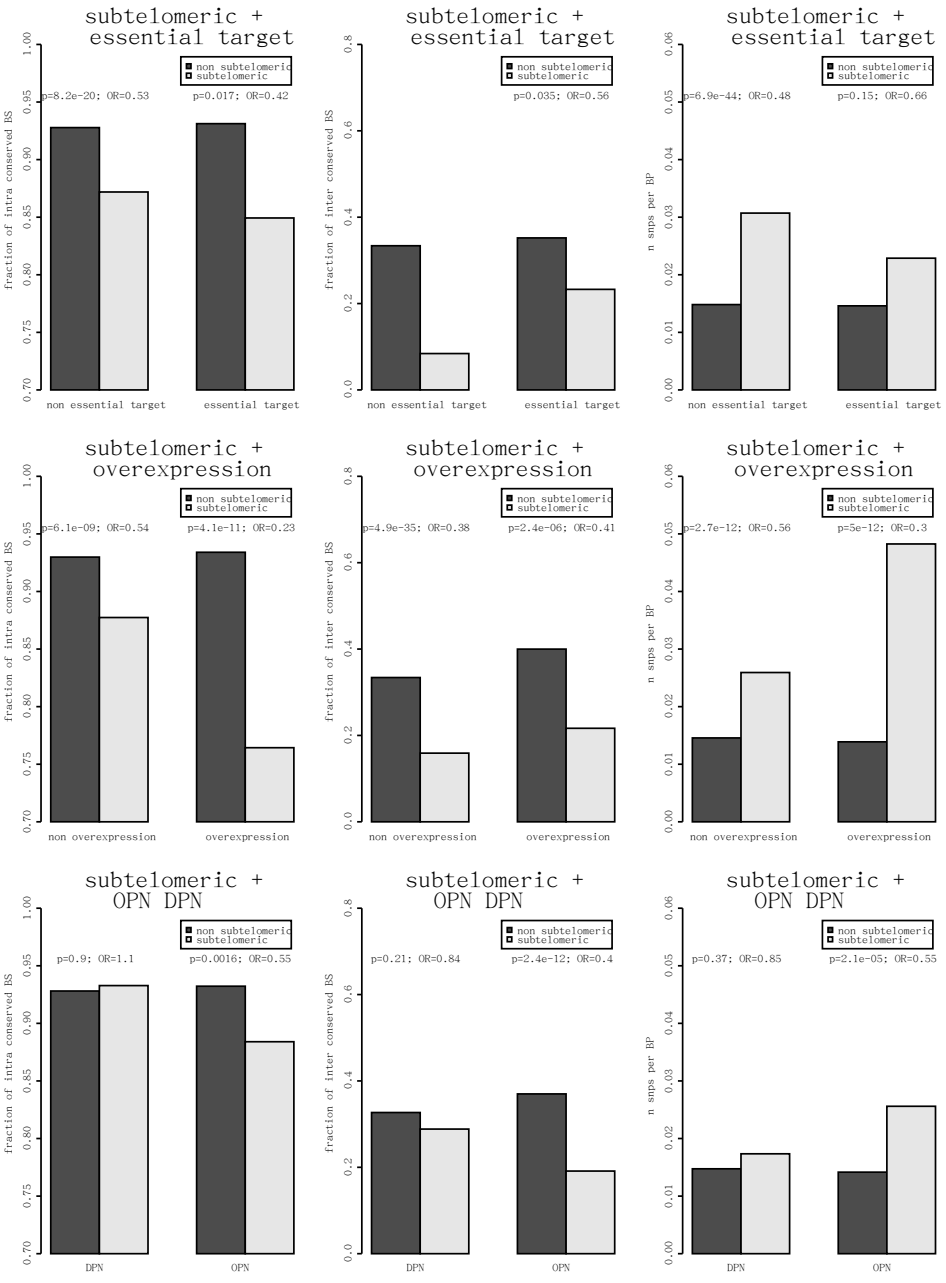

Figure 6.17: subtelomeric location interactions

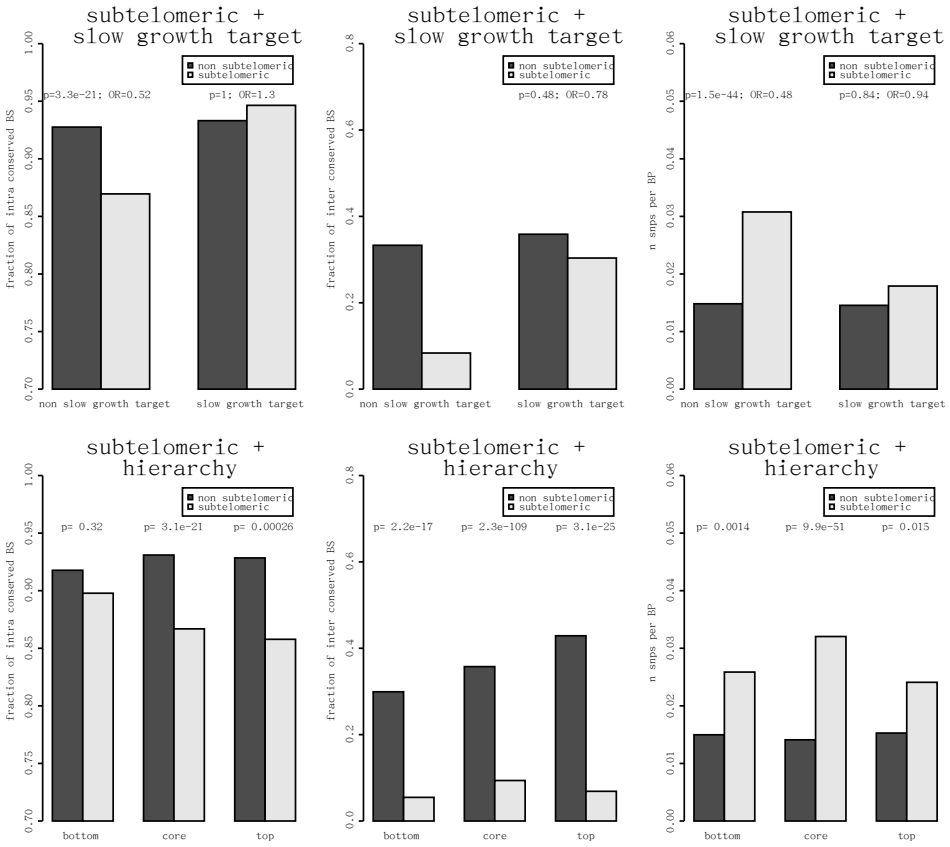

Figure 6.18: subtelomeric location interactions

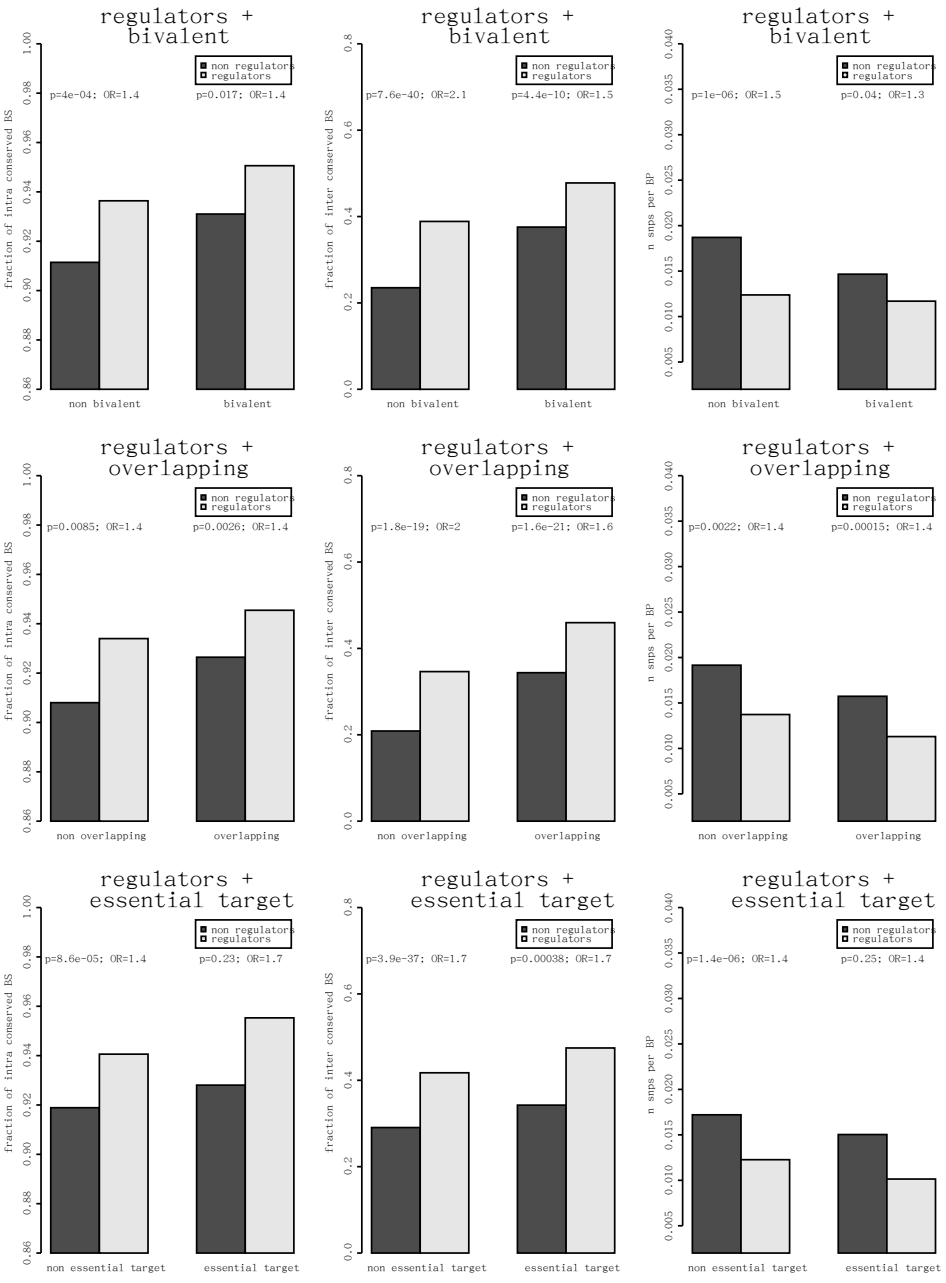

Figure 6.19: Regulators interactions

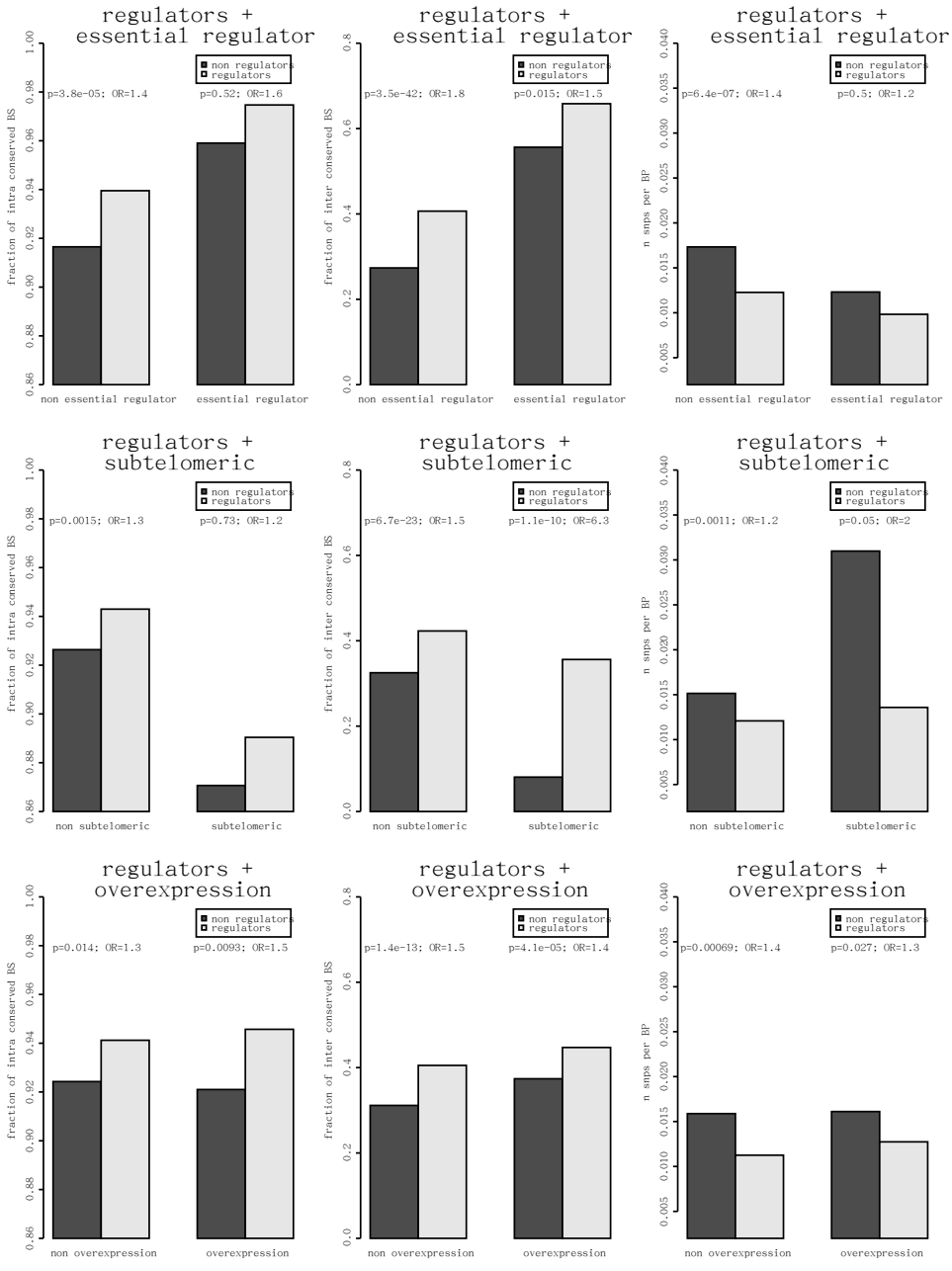

Figure 6.20: Regulators interactions

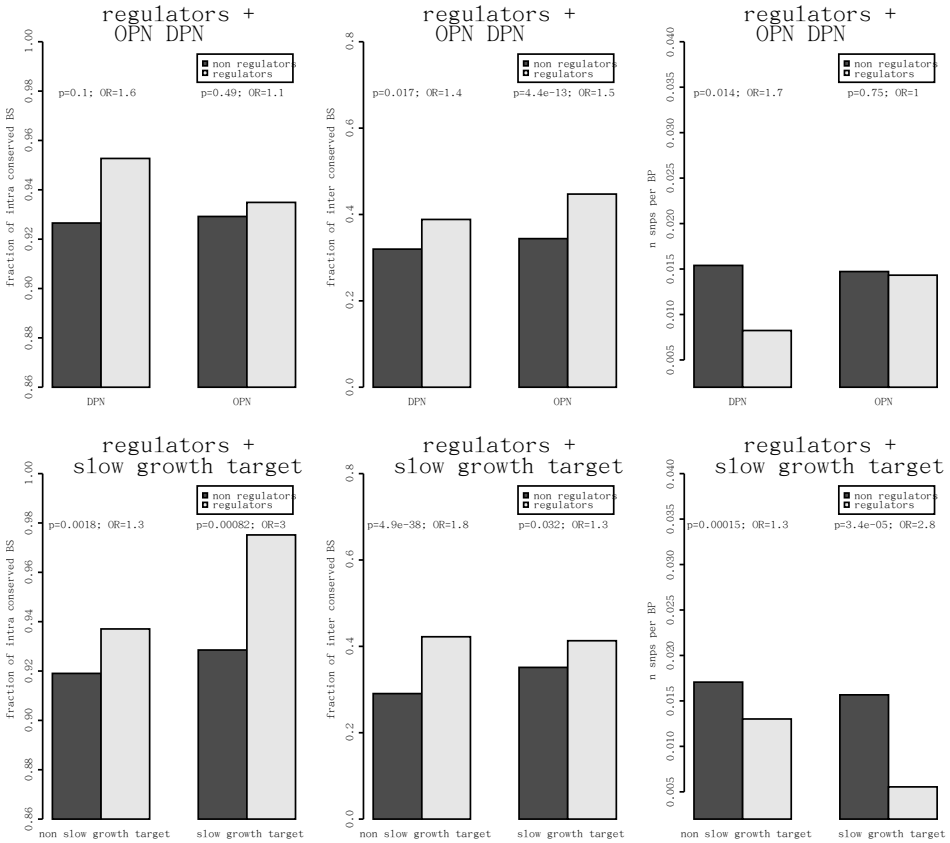

Figure 6.21: Regulators interactions

Chapter 7

Java Code

Code developed to map single nucleotide polymophisms to transcription facto binding site and to score the binding site has been written in Java language.

```
 package mirko;

import java.io.File;
import java.util.*;
import java.util.regex.*;

import org.biojava.bio.BioException;
import org.biojava.bio.seq.*;
import org.biojava.bio.dp.*;
import org.biojava.bio.symbol.*;
import org.biojava.utils.ChangeVetoException;
import org.cuc.SCresequencing.*;
import org.erasmusmc.utilities.TextFileUtilities;

public class WeightMatrixAnalysis {

String workDirectory = "/workdir/lehner/cere/match/";
String gffSet;
String outputDir = "/workdir/lehner/improved_map/";
Integer quality;
String gffFilename;
String type;
public WeightMatrixAnalysis(String gffSet,Integer quality,String type) {
this.gffSet = gffSet;
this.quality=quality;
```

```
this.type=type;
gffFilename = outputDir + gffSet + "_arab_corrected_new.gff";
}

public void writeMotifSequence() {

String snpFilename = outputDir + gffSet
+ "/snp_table_tot_imputed_new.txt";
String outputFile = outputDir + gffSet + "/motif_sequences_all_strains.txt";
List<GeneAnnotation> gff = new MotifGFFLoader2(gffFilename).getGas();
SequenceManager manager = new SequenceManager(workDirectory);
List<String> lines = new ArrayList<String>();

Map<String, Set<String>> strainsMap =
new SNPsStrainsLoader(snpFilename,quality).getStrains();

for (GeneAnnotation ga : gff) {
String start = ga.getFirst().toString();
String end = ga.getLast().toString();
String motif = ga.getName().toString();
String key = start + " " + end + " " + motif + ";";
ga.setFirst(ga.getFirst());
ga.setLast(ga.getLast());
Map<String, ORF> sq = manager.getORF(ga, quality,true);
if (strainsMap.keySet().contains(key)) {
Set<String> strains = strainsMap.get(key);
String strand = "";
System.out.print(ga.getStrand().toString());
if (ga.getStrand().toString().equals("NEGATIVE")) {
strand = "-";
} else {
strand = "+";
}
for (String strain : strains) {

String line = ga.getName() + "\t" + ga.getTemplateID()
+ "\t" + start + "\t" + end + "\t" + strand + "\t"
+ strain + "\t"
+ sq.get(strain).getCodingSequence() + "\t"
+ sq.get("ref").getCodingSequence();
lines.add(line);
```

```
}
}
}
TextFileUtilities.saveToFile(lines, outputFile);

}

public Map<String, Sequence> getMotifSequencesWT()
throws IllegalSymbolException {
String gffFilename = outputDir + gffSet + "_arab_corrected_new.gff";
List<GeneAnnotation> gff = new MotifGFFLoader2(gffFilename).getGas();
SequenceManager manager = new SequenceManager(workDirectory);
Map<String, Sequence> results = new HashMap<String, Sequence>();
for (GeneAnnotation ga : gff) {
String start = ga.getFirst().toString();
String end = ga.getLast().toString();
String motif = ga.getName().toString();
ga.setFirst(ga.getFirst());
ga.setLast(ga.getLast());
Map<String, ORF> sq = manager.getORF(ga, quality,true);
String strand = "";
System.out.print(ga.getStrand().toString());
if (ga.getStrand().toString().equals("NEGATIVE")) {
strand = "-";
} else {
strand = "+";
}
String key = ga.getName() + "\t" + ga.getTemplateID() + "\t"
+ start + "\t" + end + "\t" + strand;
String sequence = sq.get("ref").getCodingSequence();
Sequence seq = DNATools.createDNASequence(sequence, motif);
results.put(key, seq);
}
return results;
}

public void WriteMotifScoresWT()
throws ChangeVetoException, BioException {
String outFile;
if(type=="bulyk"){
outFile=outputDir+gffSet+"/scores_wt_bulyk.txt";
```

```
 }else{
outFile=outputDir+gffSet+"/scores_wt.txt";
 }
 Map<String ,WeightMatrix> matrices = new
 WeightMatrixLoader(type).getWeightMatrices();

 Map<String, Sequence> sequences = getMotifSequencesWTFromFile();
 Pattern s = Pattern.compile(" ");
 List<String> lines = new ArrayList<String>();
 String line = "";
 for (String key:sequences.keySet()){
 String[] keyfields =s.split(key);
 String motif = keyfields[0];
 String strand= keyfields[4];
 if(matrices.containsKey(motif)){
 WeightMatrix matrix = matrices.get(motif);
 WeightMatrixAnnotator wmaOdds = new WeightMatrixAnnotator(matrix,
 ScoreType.ODDS,0);

 Sequence seq = sequences.get(key);
 String name= seq.getName();

 seq =
 DNATools.createDNASequence(DNATools.reverseComplement(seq).seqString(),name);

 DNATools.createDNASequence(seq.seqString().toString(),"seq");
 System.out.println(seq.seqString());
 Sequence seqodds = wmaOdds.annotate(seq);
 int j=0;
 float oddscore =0;
 for (Iterator<List> it = seqodds.features(); it.hasNext() ;) {
 Feature f_odds = (Feature)it.next();
 System.out.println(strand);
 System.out.println(j);
 if (oddscore >
 Float.parseFloat(f_odds.getAnnotation().getProperty("score").toString())){

 }else{
 oddscore =
 Float.parseFloat(f_odds.getAnnotation().getProperty("score").toString());
 ++j;
```

```
 }
 }
 line= key+"\t"+oddscore+"\t"+seq.seqString();
 lines.add(line);
 line="";
 }
 }
 TextFileUtilities.saveToFile(lines, outFile);

 }
```

```
public Map<String, Sequence> getMotifSequencesWTFromFile()
throws IllegalSymbolException {
String filename;
if(type=="bulyk"){
filename=outputDir + gffSet + "/motif_sequences_bulyk.txt";
}else{
filename=outputDir + gffSet + "/motif_sequences.txt";
}
List<String> lines;
File file = new File(filename);
String refKey = "ref";
lines = TextFileUtilities.loadFromFile(file.getAbsolutePath());
Pattern tab = Pattern.compile("\t");
Map<String, Sequence> results = new HashMap<String, Sequence>();
for (String line : lines) {
String[] fields = tab.split(line);
String[] keyArray = new String[5];
for (int i = 0; i < keyArray.length; ++i) {
keyArray[i] = fields[i];
}
String motifKey = join(keyArray, " ");
lines = TextFileUtilities.loadFromFile(file.getAbsolutePath());
Sequence seq = DNATools.createDNASequence(fields[5], refKey);

results.put(motifKey, seq);
}
return results;
```

```
}

public void writeMotifSequenceWT() {
String outputFile;
if(type=="bulyk"){
outputFile= outputDir + gffSet + "/motif_sequences_bulyk.txt";
}else{
outputFile= outputDir + gffSet + "/motif_sequences.txt";
}
String gffFilename = outputDir + gffSet + "_arab_corrected_new.gff";
List<GeneAnnotation> gff = new MotifGFFLoader2(gffFilename).getGas();
SequenceManager manager = new SequenceManager(workDirectory);
List<String> lines = new ArrayList<String>();

for (GeneAnnotation ga : gff) {
if(type=="bulyk"){
ga.setFirst(ga.getFirst()-10);
ga.setLast(ga.getLast()+10);
}
String start = ga.getFirst().toString();
String end = ga.getLast().toString();
String motif = ga.getName().toString();
Map<String, ORF> sq = manager.getORF(ga, 0);
String strand = "";
System.out.print(ga.getStrand().toString());
if (ga.getStrand().toString().equals("NEGATIVE")) {
strand = "-";
} else {
strand = "+";
}
String line = motif + "\t" + ga.getTemplateID() + "\t"
+ start + "\t" + end + "\t" + strand + "\t"
+ sq.get("ref").getCodingSequence();
System.out.println(line);
lines.add(line);
}
TextFileUtilities.saveToFile(lines, outputFile);

}
```

```
public Map<GeneAnnotation, List<Sequence>> getMotifSequences()
throws IllegalSymbolException {
String gffFilename = outputDir + gffSet + "_arab_corrected_new.gff";
String snpFilename;
if(type=="bulyk"){snpFilename = outputDir + gffSet
+ "/SNPFromSequence_bulyk.txt";
}else{
snpFilename = outputDir + gffSet
+ "/SNPFromSequence.txt";
}
List<GeneAnnotation> gff = new MotifGFFLoader2(gffFilename).getGas();
SequenceManager manager = new SequenceManager(workDirectory);
Map<String, Set<String>> strainsMap = new SNPsStrainsLoader(snpFilename,quality)
.getStrains();
String refKey = "ref";
Map<GeneAnnotation, List<Sequence>> results =
new HashMap<GeneAnnotation, List<Sequence>>();
for (GeneAnnotation ga : gff) {
if(type=="bulyk"){
ga.setFirst(ga.getFirst()-10);
ga.setLast(ga.getLast()+10);
}
String start = ga.getFirst().toString();
String end = ga.getLast().toString();
String motif = ga.getName().toString();
String key = start + " " + end + " " + motif;
Map<String, ORF> sq = manager.getORF(ga, quality,true);
List<Sequence> sequences = new ArrayList<Sequence>();
if (strainsMap.keySet().contains(key)) {
Set<String> strains = strainsMap.get(key);
Sequence refSeq = DNATools.createDNASequence(sq.get("ref")
.getCodingSequence(), refKey);
sequences.add(refSeq);
for (String strain : strains) {
strain=strain.replace('_', '.');
strain=strain.toUpperCase();
Sequence strainSeq = DNATools.createDNASequence(sq.get(
strain).getCodingSequence(), strain);
sequences.add(strainSeq);
```

```
}
results.put(ga, sequences);
}

}
return results;
}


public Map<GeneAnnotation, Map<String, Float>> getMotifScores()
throws ChangeVetoException, BioException {
Map<String, WeightMatrix> matrices =
new WeightMatrixLoader(type).getWeightMatrices();
Map<GeneAnnotation, Map<String, Float>> results =
new HashMap<GeneAnnotation, Map<String, Float>>();
Map<GeneAnnotation, List<Sequence>> sequences = getMotifSequences();
for (GeneAnnotation ga : sequences.keySet()) {
System.out. print(ga.getName());
if(matrices.containsKey(ga.getName())){
WeightMatrix matrix = matrices.get(ga.getName());
WeightMatrixAnnotator wmaOdds =
new WeightMatrixAnnotator(matrix,ScoreType.ODDS, 0);

List<Sequence> seqlist = sequences.get(ga);
Map<String, Float> scores=calculateScore(wmaOdds, seqlist);

results.put(ga,scores);
}
}
return results;

}

public void writeMotifScores(){
String outFile;
if(type=="bulyk"){
outFile=outputDir+gffSet+"/scores_from_sequences_bulyk.txt";
}else{
outFile=outputDir+gffSet+"/scores_from_sequences.txt";
}
```

```
List<String> lines = new ArrayList<String>();
Map<GeneAnnotation, Map<String, Float>> scores = null;
try {
scores = getMotifScores();
} catch (ChangeVetoException e) {
// TODO Auto-generated catch block
e.printStackTrace();
} catch (BioException e) {
// TODO Auto-generated catch block
e.printStackTrace();
}
for (GeneAnnotation ga:scores.keySet()){
String annot= new AnnotationToString(ga).get();
for (String strain:scores.get(ga).keySet()){
String score = Float.toString(scores.get(ga).get(strain));
String add= "\t"+strain+"\t"+score;
String line=annot+add;
lines.add(line);
}
}
TextFileUtilities.saveToFile(lines, outFile);
}

public Map<String, Float> calculateScore(WeightMatrixAnnotator wma ,List<Sequenc
throws ChangeVetoException, BioException{
Map<String, Float> results = new HashMap<String, Float>();
for (ListIterator I = seqlist.listIterator(); I.hasNext();) {

Sequence seq = (Sequence) I.next();
String name = seq.getName()+"\t"+seq.seqString();

seq = DNATools.createDNASequence(DNATools
.reverseComplement(seq).seqString(), name);

System.out.println(seq.seqString());

Float score = new Float(0);
Location loc = null;
for (Iterator<List> it = seq.features(); it.hasNext();) {

Feature f = (Feature) it.next();
```

```
if (score > Float.parseFloat(f.getAnnotation()
.getProperty("score").toString())) {

} else {
score = Float.parseFloat(f.getAnnotation()
.getProperty("score").toString());
loc = f.getLocation();
}

}
String max = "0";
String min = "0";
if(score > 0){
max = Integer.toString(loc.getMax());
min = Integer.toString(loc.getMin());
}
name = name + "\t"+ min +"\t"+max+"\t";
results.put(name,score);
}
return results;
}

public Map<String, Float> calculateRatios(Map<String, Float> scores){
Map<String, Float> ratios= new HashMap<String, Float>();
Float refscore= scores.get("ref");
for (String strain:scores.keySet()){
Float ratio = scores.get(strain)/refscore;
ratios.put(strain, ratio);
}
return ratios;

}

public Map<String, Map<String, Float>> getMotifScoresFromFile()
throws ChangeVetoException, BioException {
List<String> lines;
String filename= outputDir + gffSet + "/scores.txt";
File file = new File(filename);
lines = TextFileUtilities.loadFromFile(file.getAbsolutePath());
```

```
Pattern tab = Pattern.compile("\t");
Map<String, Map<String,Float>> results =
new HashMap<String, Map<String,Float>>();
for (String line : lines) {
String[] fields = tab.split(line);
String key = fields[0]+"\t"+fields[1]+"\t"+fields[2]+"\t"+fields[3]+
"\t"+fields[4];
String strain = fields[5];
Float score = Float.valueOf(fields[6]);
if (!results.containsKey(key)){
Map<String, Float> scores = new HashMap<String, Float>();
results.put(key, scores);
}
results.get(key).put(strain, score);
}
return results;
}
```

```
public Map<String, Map<String,Float>>
getModifiedPromotersScores(String scoreType,String filt,int thresh) {
PromoterBuilder PB = new PromoterBuilder(gffSet,thresh,filt);
Map<String, List<String>> promoters = PB.getPromotersString();
PB.writePromoters();
Map<String, Map<String, Float>> Scores = null;
Map<String, Map<String,Float>> results =
new HashMap<String, Map<String,Float>>();
try {
Scores = getMotifScoresFromFile();
} catch (ChangeVetoException e) {
// TODO Auto-generated catch block
e.printStackTrace();
} catch (BioException e) {
// TODO Auto-generated catch block
e.printStackTrace();
}
for (String ga:promoters.keySet()){
```

```
List<String> motifs = promoters.get(ga);
Map<String,Float>  mutatedMotifs = new HashMap<String, Float>();
for (Iterator<String> i =motifs.iterator();i.hasNext();){
String motif = i.next();
if (Scores.keySet().contains(motif)){
Map<String, Float> strains = Scores.get(motif);
if (scoreType.equals("ratios")){
strains=calculateRatios(strains);
}
for(String strain : strains.keySet()){
String key = motif+"\t"+strain;
mutatedMotifs.put(key, strains.get(strain));
}
}
}
if (!mutatedMotifs.isEmpty()){
results.put(ga, mutatedMotifs);
}
}

return results;

}



public void writePromoterScores(String scoreType,String filt,int thresh){
Map<String, Map<String,Float>> scores =
getModifiedPromotersScores(scoreType,filt,thresh);
String filename= outputDir+gffSet+"/"+filt+"/"+"promoter_"+scoreType+".txt";
List<String> lines =  new ArrayList<String>();
for (String ga:scores.keySet()){
for (String strain:scores.get(ga).keySet()){
String score = Float.toString(scores.get(ga).get(strain));
String add= "\t"+strain+"\t"+score;
String line=ga+add;
lines.add(line);
}
}
TextFileUtilities.saveToFile(lines, filename);
}
```

```
public void writeMotifSnps(){
List<GeneAnnotation> gas = new MotifGFFLoader2(gffFilename).getGas();
SequenceManager manager = new SequenceManager(workDirectory);
String filename;
if(type=="bulyk"){
filename= outputDir+gffSet+"/SNPFromSequence_bulyk.txt";
}else{
filename= outputDir+gffSet+"/SNPFromSequence.txt";
}
List<String> lines =  new ArrayList<String>();
for (GeneAnnotation ga:gas){
if(type=="bulyk"){
ga.setFirst(ga.getFirst()-10);
ga.setLast(ga.getLast()+10);
}
Map<String,List<Morph>> allMorphs =
manager.getMorphsForGene(ga,quality,false);
if(!allMorphs.isEmpty()){
for (String  strain : allMorphs.keySet()){
List<Morph> morphs=allMorphs.get(strain);
if(!morphs.isEmpty()){
for(Morph morph:morphs){
if (morph instanceof SNP){
SNP snp = (SNP) morph;
String chr = snp.getTemplate();
String pos = Integer.toString(snp.getStartPos());
Character wt = snp.getChange().object1;
Character mut =snp.getChange().object2;
String change = wt.toString()+">"+mut.toString();
String errRate= snp.getErrorProbs().toString();
String motifStart = Integer.toString(ga.getFirst());
String motifEnd= Integer.toString(ga.getLast());
String motif = ga.getName();
String strand = "";
if (ga.getStrand().toString().equals("NEGATIVE")) {
strand = "-";
} else {
strand = "+";
}
String line= chr+"\t"+pos+"\t"+pos+"\t"+strain+
"\t"+change+"\t"+errRate+"\t"+motifStart
```

```
+"\t"+motifEnd+"\t"+strand+"\t"+motif;
lines.add(line);
}
}
}
}
}
}
TextFileUtilities.saveToFile(lines, filename);
return;
}




static String join(String[] array, String delim) {
StringBuffer sb = join(array, delim, new StringBuffer());
return sb.toString();
}

static StringBuffer join(String[] array, String delim, StringBuffer sb) {
for (int i = 0; i < array.length; i++) {
if (i != 0)
sb.append(delim);
sb.append(array[i]);
}
return sb;
}



}
```

```
package mirko;

import org.biojava.bio.BioException;
import org.biojava.utils.ChangeVetoException;
```

```
public class MotifsSNPSAnalyzer {


public static void main(String[] args)
throws ChangeVetoException, BioException {

String [] genesets= {"all_genes"};
String [] gffsets= {"p005_c1"};
Integer quality = 30;

int thresh=1000;
new WeightMatrixLoader("bulyk").WriteMaxScores();
for (String gffset:gffsets){
WeightMatrixAnalysis analysis =
new WeightMatrixAnalysis(gffset,quality,"bulyk");
analysis.writeMotifSnps();
analysis.writeMotifSequenceWT();
analysis.WriteMotifScoresWT();
analysis.writeMotifScores();


for (String geneset:genesets){
new PromoterBuilder(gffset,1000,geneset).writePromoters();
analysis.writePromoterScores("odds",geneset,thresh);
analysis.writePromoterScores("ratios",geneset,thresh);
}
}
}

}
```

Bibliography

Uri Alon. Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8(6):450–461, Jun 2007. doi: 10.1038/nrg2102. URL http://dx.doi.org/10.1038/nrg2102.

Oscar Aparicio, Joseph V Geisberg, Edward Sekinger, Annie Yang, Zarmik Moqtaderi, and Kevin Struhl. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr Protoc Mol Biol*, Chapter 21: Unit 21.3, Feb 2005. doi: 10.1002/0471142727.mb2103s69. URL http://dx.doi.org/10.1002/0471142727.mb2103s69.

M. Madan Babu, Nicholas M Luscombe, L. Aravind, Mark Gerstein, and Sarah A Teichmann. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, 14 (3):283–291, Jun 2004. doi: 10.1016/j.sbi.2004.05.004. URL http://dx.doi.org/10.1016/j.sbi.2004.05.004.

S. Balaji, M. Madan Babu, Lakshminarayan M Iyer, Nicholas M Luscombe, and L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol*, 360(1):213–227, Jun 2006. doi: 10.1016/j.jmb.2006.04.029. URL http://dx.doi.org/10.1016/j.jmb.2006.04.029.

Mukesh Bansal, Vincenzo Belcastro, Alberto Ambesi-Impiombato, and Diego di Bernardo. How to infer gene networks from expression profiles. *Mol Syst Biol*, 3:78, 2007. doi: 10.1038/msb4100120. URL http://dx.doi.org/10.1038/msb4100120.

Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, Feb 2004. doi: 10.1038/nrg1272. URL http://dx.doi.org/10.1038/nrg1272.

Nizar N Batada and Laurence D Hurst. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet*, 39(8):945–949, Aug 2007. doi: 10.1038/ng2071. URL http://dx.doi.org/10.1038/ng2071.

O. G. Berg and P. H. von Hippel. Selection of dna binding sites by regulatory proteins. statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193(4):723–750, Feb 1987.

Jesse D Bloom, D. Allan Drummond, Frances H Arnold, and Claus O Wilke. Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol*, 23(9):1751–1761, Sep 2006. doi: 10.1093/molbev/msl040. URL http://dx.doi.org/10.1093/molbev/msl040.

Dario Boffelli, Jon McAuliffe, Dmitriy Ovcharenko, Keith D Lewis, Ivan Ovcharenko, Lior Pachter, and Edward M Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, Feb 2003. doi: 10.1126/science.1081331. URL http://dx.doi.org/10.1126/science.1081331.

Dario Boffelli, Claire V Weer, Li Weng, Keith D Lewis, Malak I Shoukry, Lior Pachter, David N Keys, and Edward M Rubin. Intraspecies sequence comparisons for annotating genomes. *Genome Res*, 14(12):2406–2411, Dec 2004. doi: 10.1101/gr.3199704. URL http://dx.doi.org/10.1101/gr.3199704.

Richard Bonneau, David J Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S Baliga, and Vesteinn Thorsson. The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol*, 7(5):R36, 2006. doi: 10.1186/gb-2006-7-5-r36. URL http://dx.doi.org/10.1186/gb-2006-7-5-r36.

Richard Bonneau, Marc T Facciotti, David J Reiss, Amy K Schmid, Min Pan, Amardeep Kaur, Vesteinn Thorsson, Paul Shannon, Michael H Johnson, J. Christopher Bare, William Longabaugh, Madhavi Vuthoori, Kenia Whitehead, Aviv Madar, Lena Suzuki, Tetsuya Mori, Dong-Eun Chang, Jocelyne Diruggiero, Carl H Johnson, Leroy Hood, and Nitin S Baliga. A predictive model for transcriptional control of physiology in a free living cell. *Cell*, 131(7):1354–1365, Dec 2007. doi: 10.1016/j.cell.2007.10.053. URL http://dx.doi.org/10.1016/j.cell.2007.10.053.

Alice Bossi and Ben Lehner. Tissue specificity and the human protein interaction network. *Mol Syst Biol*, 5:260, 2009. doi: 10.1038/msb.2009.17. URL http://dx.doi.org/10.1038/msb.2009.17.

Michael Boutros, Amy A Kiger, Susan Armknecht, Kim Kerr, Marc Hild, Britta Koch, Stefan A Haas, Renato Paro, Norbert Perrimon, and Heidelberg Fly Array Consortium. Genome-wide rnai analysis of growth and viability in drosophila cells. *Science*, 303(5659):832–835, Feb 2004.

Laurie A Boyer, Tong Ihn Lee, Megan F Cole, Sarah E Johnstone, Stuart S Levine, Jacob P Zucker, Matthew G Guenther, Roshan M Kumar, Heather L Murray, Richard G Jenner, David K Gifford, Douglas A Melton, Rudolf Jaenisch, and Richard A Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956, Sep 2005. doi: 10.1016/j.cell.2005.08.020. URL `http://dx.doi.org/10.1016/j.cell.2005.08.020`.

Rachel B Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A*, 102(5):1572–1577, Feb 2005. doi: 10.1073/pnas.0408709102. URL `http://dx.doi.org/10.1073/pnas.0408709102`.

Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, Apr 2002. doi: 10.1126/science.1069516. URL `http://dx.doi.org/10.1126/science.1069516`.

Jason S Carroll, Clifford A Meyer, Jun Song, Wei Li, Timothy R Geistlinger, Jérôme Eeckhoute, Alexander S Brodsky, Erika Krasnickas Keeton, Kirsten C Fertuck, Giles F Hall, Qianben Wang, Stefan Bekiranov, Victor Sementchenko, Edward A Fox, Pamela A Silver, Thomas R Gingeras, X. Shirley Liu, and Myles Brown. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet*, 38(11):1289–1297, Nov 2006. doi: 10.1038/ng1901. URL `http://dx.doi.org/10.1038/ng1901`.

S. B. Carroll. Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, 101(6):577–580, Jun 2000.

Sean B Carroll. Evolution at two levels: on genes and form. *PLoS Biol*, 3(7):e245, Jul 2005. doi: 10.1371/journal.pbio.0030245. URL `http://dx.doi.org/10.1371/journal.pbio.0030245`.

Sean B Carroll. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134 (1):25–36, Jul 2008. doi: 10.1016/j.cell.2008.06.030. URL `http://dx.doi.org/10.1016/j.cell.2008.06.030`.

K. Cartharius, K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein, and T. Werner. Matinspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 21(13):2933–2942, Jul 2005. doi: 10.1093/bioinformatics/bti473. URL `http://dx.doi.org/10.1093/bioinformatics/bti473`.

Esther T Chan, Gerald T Quon, Gordon Chua, Tomas Babak, Miles Trochesset, Ralph A Zirngibl, Jane Aubin, Michael J H Ratcliffe, Andrew Wilde, Michael Brudno, Quaid D Morris, and Timothy R Hughes. Conservation of core gene expression in vertebrate tissues. *J Biol*, 8(3):33, 2009. doi: 10.1186/jbiol130. URL `http://dx.doi.org/10.1186/jbiol130`.

A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res*, 13(9):3021–3030, May 1985.

D. Corà, C. Herrmann, C. Dieterich, F. Di Cunto, P. Provero, and M. Caselle. Ab initio identification of putative human transcription factor binding sites by comparative genomics. *BMC Bioinformatics*, 6:110, 2005. doi: 10.1186/1471-2105-6-110. URL `http://dx.doi.org/10.1186/1471-2105-6-110`.

Davide Corà, Ferdinando Di Cunto, Paolo Provero, Lorenzo Silengo, and Michele Caselle. Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs. *BMC Bioinformatics*, 5:57, May 2004. doi: 10.1186/1471-2105-5-57. URL `http://dx.doi.org/10.1186/1471-2105-5-57`.

Emmanouil T Dermitzakis and Andrew G Clark. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol*, 19(7):1114–1121, Jul 2002.

Adam M Deutschbauer, Daniel F Jaramillo, Michael Proctor, Jochen Kumm, Maureen E Hillenmeyer, Ronald W Davis, Corey Nislow, and Guri Giaever. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics*, 169(4): 1915–1925, Apr 2005. doi: 10.1534/genetics.104.036871. URL `http://dx.doi.org/10.1534/genetics.104.036871`.

Scott W Doniger and Justin C Fay. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol*, 3(5):e99, May 2007. doi: 10.1371/journal.pcbi.0030099. URL `http://dx.doi.org/10.1371/journal.pcbi.0030099`.

Yair Field, Noam Kaplan, Yvonne Fondufe-Mittendorf, Irene K Moore, Eilon Sharon, Yaniv Lubling, Jonathan Widom, and Eran Segal. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol*, 4 (11):e1000216, Nov 2008. doi: 10.1371/journal.pcbi.1000216. URL http://dx.doi.org/10.1371/journal.pcbi.1000216.

Hunter B Fraser, Aaron E Hirsh, Guri Giaever, Jochen Kumm, and Michael B Eisen. Noise minimization in eukaryotic gene expression. *PLoS Biol*, 2(6):e137, Jun 2004. doi: 10.1371/journal.pbio.0020137. URL http://dx.doi.org/10.1371/journal.pbio.0020137.

Feng Gao, Barrett C Foat, and Harmen J Bussemaker. Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data. *BMC Bioinformatics*, 5:31, Mar 2004. doi: 10.1186/1471-2105-5-31. URL http://dx.doi.org/10.1186/1471-2105-5-31.

Timothy S Gardner, Diego di Bernardo, David Lorenz, and James J Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301 (5629):102–105, Jul 2003. doi: 10.1126/science.1081900. URL http://dx.doi.org/10.1126/science.1081900.

Daniel M Gelperin, Michael A White, Martha L Wilkinson, Yoshiko Kon, Li A Kung, Kevin J Wise, Nelson Lopez-Hoyo, Lixia Jiang, Stacy Piccirillo, Haiyuan Yu, Mark Gerstein, Mark E Dumont, Eric M Phizicky, Michael Snyder, and Elizabeth J Grayhack. Biochemical and genetic analysis of the yeast proteome with a movable orf collection. *Genes Dev*, 19(23):2816–2826, Dec 2005. doi: 10.1101/gad.1362105. URL http://dx.doi.org/10.1101/gad.1362105.

Justin Gerke, Kim Lorenz, and Barak Cohen. Genetic interactions between transcription factors cause natural variation in yeast. *Science*, 323(5913):498–501, Jan 2009. doi: 10.1126/science.1166426. URL http://dx.doi.org/10.1126/science.1166426.

Guri Giaever, Angela M Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Véronneau, Sally Dow, Ankuta Lucau-Danila, Keith Anderson, Bruno André, Adam P Arkin, Anna Astromoff, Mohamed El-Bakkoury, Rhonda Bangham, Rocio Benito, Sophie Brachat, Stefano Campanaro, Matt Curtiss, Karen Davis, Adam Deutschbauer, Karl-Dieter Entian, Patrick Flaherty, Francoise Foury, David J Garfinkel, Mark Gerstein, Deanna Gotte,

Ulrich Güldener, Johannes H Hegemann, Svenja Hempel, Zelek Herman, Daniel F Jaramillo, Diane E Kelly, Steven L Kelly, Peter Kötter, Darlene LaBonte, David C Lamb, Ning Lan, Hong Liang, Hong Liao, Lucy Liu, Chuanyun Luo, Marc Lussier, Rong Mao, Patrice Menard, Siew Loon Ooi, Jose L Revuelta, Christopher J Roberts, Matthias Rose, Petra Ross-Macdonald, Bart Scherens, Greg Schimmack, Brenda Shafer, Daniel D Shoemaker, Sharon Sookhai-Mahadeo, Reginald K Storms, Jeffrey N Strathern, Giorgio Valle, Marleen Voet, Guido Volckaert, Ching yun Wang, Teresa R Ward, Julie Wilhelmy, Elizabeth A Winzeler, Yonghong Yang, Grace Yen, Elaine Youngman, Kexin Yu, Howard Bussey, Jef D Boeke, Michael Snyder, Peter Philippsen, Ronald W Davis, and Mark Johnston. Functional profiling of the saccharomyces cerevisiae genome. *Nature*, 418(6896):387–391, Jul 2002. doi: 10.1038/nature00935. URL `http://dx.doi.org/10.1038/nature00935`.

Luca Giorgetti, Trevor Siggers, Guido Tiana, Greta Caprara, Samuele Notarbartolo, Teresa Corona, Manolis Pasparakis, Paolo Milani, Martha L Bulyk, and Gioacchino Natoli. Noncooperative interactions between transcription factors and clustered dna binding sites enable graded transcriptional responses to environmental inputs. *Mol Cell*, 37 (3):418–428, Feb 2010. doi: 10.1016/j.molcel.2010.01.016. URL `http://dx.doi.org/10.1016/j.molcel.2010.01.016`.

Nicolas Gompel, Benjamin Prud'homme, Patricia J Wittkopp, Victoria A Kassner, and Sean B Carroll. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in drosophila. *Nature*, 433(7025):481–487, Feb 2005. doi: 10.1038/nature03235. URL `http://dx.doi.org/10.1038/nature03235`.

Christopher T Harbison, D. Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-Bosco Tagne, David B Reynolds, Jane Yoo, Ezra G Jennings, Julia Zeitlinger, Dmitry K Pokholok, Manolis Kellis, P. Alex Rolfe, Ken T Takusagawa, Eric S Lander, David K Gifford, Ernest Fraenkel, and Richard A Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, Sep 2004. doi: 10.1038/nature02800. URL `http://dx.doi.org/10.1038/nature02800`.

Maureen E Hillenmeyer, Eula Fung, Jan Wildenhain, Sarah E Pierce, Shawn Hoon, William Lee, Michael Proctor, Robert P St Onge, Mike Tyers, Daphne Koller, Russ B Altman, Ronald W Davis, Corey Nislow, and Guri Giaever. The chemical genomic portrait

of yeast: uncovering a phenotype for all genes. *Science*, 320 (5874):362–365, Apr 2008. doi: 10.1126/science.1150021. URL http://dx.doi.org/10.1126/science.1150021.

Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–9367, Jun 2009. doi: 10.1073/pnas.0903103106. URL http://dx.doi.org/10.1073/pnas.0903103106.

Hopi E Hoekstra and Jerry A Coyne. The locus of evolution: evo devo and the genetics of adaptation. *Evolution*, 61(5):995–1016, May 2007. doi: 10.1111/j.1558-5646.2007.00105.x. URL http://dx.doi.org/10.1111/j.1558-5646.2007.00105.x.

R. C G Holland, T. A. Down, M. Pocock, A. Prlić, D. Huen, K. James, S. Foisy, A. Dräger, A. Yates, M. Heuer, and M. J. Schreiber. Biojava: an open-source framework for bioinformatics. *Bioinformatics*, 24 (18):2096–2097, Sep 2008. doi: 10.1093/bioinformatics/btn397. URL http://dx.doi.org/10.1093/bioinformatics/btn397.

T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518): 929–934, May 2001. doi: 10.1126/science.292.5518.929. URL http://dx.doi.org/10.1126/science.292.5518.929.

Jan Ihmels, Sven Bergmann, Maryam Gerami-Nejad, Itai Yanai, Mark McClellan, Judith Berman, and Naama Barkai. Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science*, 309(5736):938–940, Aug 2005. doi: 10.1126/science.1113833. URL http://dx.doi.org/10.1126/science.1113833.

Mark Isalan, Caroline Lemerle, Konstantinos Michalodimitrakis, Carsten Horn, Pedro Beltrao, Emanuele Raineri, Mireia Garriga-Canut, and Luis Serrano. Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, 452(7189):840–845, Apr 2008. doi: 10.1038/nature06847. URL http://dx.doi.org/10.1038/nature06847.

F. JACOB and J. MONOD. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356, Jun 1961.

H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001. doi: 10.1038/35075138. URL http://dx.doi.org/10.1038/35075138.

David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, Jun 2007. doi: 10.1126/science.1141319. URL http://dx.doi.org/10.1126/science.1141319.

Rory Johnson, John Samuel, Calista Keow Leng Ng, Ralf Jauch, Lawrence W Stanton, and Ian C Wood. Evolution of the vertebrate gene regulatory network controlled by the transcriptional repressor rest. *Mol Biol Evol*, 26(7):1491–1507, Jul 2009. doi: 10.1093/molbev/msp058. URL http://dx.doi.org/10.1093/molbev/msp058.

Raja Jothi, S. Balaji, Arthur Wuster, Joshua A Grochow, Jörg Gsponer, Teresa M Przytycka, L. Aravind, and M. Madan Babu. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Mol Syst Biol*, 5:294, 2009. doi: 10.1038/msb.2009.52. URL http://dx.doi.org/10.1038/msb.2009.52.

Ravi S Kamath, Andrew G Fraser, Yan Dong, Gino Poulin, Richard Durbin, Monica Gotta, Alexander Kanapin, Nathalie Le Bot, Sergio Moreno, Marc Sohrmann, David P Welchman, Peder Zipperlen, and Julie Ahringer. Systematic functional analysis of the caenorhabditis elegans genome using rnai. *Nature*, 421(6920):231–237, Jan 2003. doi: 10.1038/nature01278. URL http://dx.doi.org/10.1038/nature01278.

Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, May 2003. doi: 10.1038/nature01644. URL http://dx.doi.org/10.1038/nature01644.

Manolis Kellis, Bruce W Birren, and Eric S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617–624, Apr 2004. doi: 10.1038/nature02424. URL http://dx.doi.org/10.1038/nature02424.

Philipp Khaitovich, Wolfgang Enard, Michael Lachmann, and Svante Pääbo. Evolution of primate gene expression. *Nat Rev Genet*, 7(9):693–702, Sep 2006. doi: 10.1038/nrg1940. URL http://dx.doi.org/10.1038/nrg1940.

Jaebum Kim, Xin He, and Saurabh Sinha. Evolution of regulatory sequences in 12 drosophila species. *PLoS Genet*, 5(1): e1000330, Jan 2009. doi: 10.1371/journal.pgen.1000330. URL http://dx.doi.org/10.1371/journal.pgen.1000330.

M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, Apr 1975.

C. E. Lawrence and A. A. Reilly. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1):41–51, 1990. doi: 10.1002/prot.340070105. URL http://dx.doi.org/10.1002/prot.340070105.

C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, Oct 1993.

Insuk Lee, Shailesh V Date, Alex T Adai, and Edward M Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–1558, Nov 2004. doi: 10.1126/science.1099511. URL http://dx.doi.org/10.1126/science.1099511.

Insuk Lee, Ben Lehner, Catriona Crombie, Wendy Wong, Andrew G Fraser, and Edward M Marcotte. A single gene network accurately predicts phenotypic effects of gene perturbation in caenorhabditis elegans. *Nat Genet*, 40(2):181–188, Feb 2008. doi: 10.1038/ng.2007.70. URL http://dx.doi.org/10.1038/ng.2007.70.

Tong Ihn Lee, Nicola J Rinaldi, François Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, Julia Zeitlinger, Ezra G Jennings, Heather L Murray, D. Benjamin Gordon, Bing Ren, John J Wyrick, Jean-Bosco Tagne, Thomas L Volkert, Ernest Fraenkel, David K Gifford, and Richard A Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298 (5594):799–804, Oct 2002. doi: 10.1126/science.1075090. URL http://dx.doi.org/10.1126/science.1075090.

Ben Lehner. Modelling genotype-phenotype relationships and human disease with genetic interaction networks. *J Exp Biol*, 210 (Pt 9):1559–1566, May 2007. doi: 10.1242/jeb.002311. URL http://dx.doi.org/10.1242/jeb.002311.

Ben Lehner. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol Syst Biol*, 4:170, 2008. doi: 10.1038/msb.2008.11. URL http://dx.doi.org/10.1038/msb.2008.11.

Ben Lehner, Catriona Crombie, Julia Tischler, Angelo Fortunato, and Andrew G Fraser. Systematic mapping of genetic interactions in caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet*, 38(8):896–903, Aug 2006. doi: 10.1038/ng1844. URL http://dx.doi.org/10.1038/ng1844.

Bernardo Lemos, Luciana O Araripe, Pierre Fontanillas, and Daniel L Hartl. Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. *Proc Natl Acad Sci U S A*, 105 (38):14471–14476, Sep 2008. doi: 10.1073/pnas.0805160105. URL http://dx.doi.org/10.1073/pnas.0805160105.

Gianni Liti, David M Carter, Alan M Moses, Jonas Warringer, Leopold Parts, Stephen A James, Robert P Davey, Ian N Roberts, Austin Burt, Vassiliki Koufopanou, Isheng J Tsai, Casey M Bergman, Douda Bensasson, Michael J T O'Kelly, Alexander van Oudenaarden, David B H Barton, Elizabeth Bailes, Alex N Nguyen, Matthew Jones, Michael A Quail, Ian Goodhead, Sarah Sims, Frances Smith, Anders Blomberg, Richard Durbin, and Edward J Louis. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–341, Mar 2009. doi: 10.1038/nature07743. URL http://dx.doi.org/10.1038/nature07743.

Gabriela G Loots, Ivan Ovcharenko, Lior Pachter, Inna Dubchak, and Edward M Rubin. rvista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res*, 12(5):832–839, May 2002. doi: 10.1101/gr.225502. Article published online before print in April 2002. URL http://dx.doi.org/10.1101/gr.225502. Article published online before print in April 2002.

Peng Lu, Christine Vogel, Rong Wang, Xin Yao, and Edward M Marcotte. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, 25(1):117–124, Jan 2007. doi: 10.1038/nbt1270. URL http://dx.doi.org/10.1038/nbt1270.

Nicholas M Luscombe, M. Madan Babu, Haiyuan Yu, Michael Snyder, Sarah A Teichmann, and Mark Gerstein. Genomic analysis of

regulatory network dynamics reveals large topological changes. *Nature*, 431(7006):308–312, Sep 2004. doi: 10.1038/nature02782. URL http://dx.doi.org/10.1038/nature02782.

Stewart MacArthur, Xiao-Yong Li, Jingyi Li, James B Brown, Hou Cheng Chu, Lucy Zeng, Brandi P Grondona, Aaron Hechmer, Lisa Simirenko, Soile V E Keränen, David W Knowles, Mark Stapleton, Peter Bickel, Mark D Biggin, and Michael B Eisen. 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*, 10(7):R80, 2009. doi: 10.1186/gb-2009-10-7-r80. URL http://dx.doi.org/10.1186/gb-2009-10-7-r80.

Kenzie D MacIsaac, Ting Wang, D. Benjamin Gordon, David K Gifford, Gary D Stormo, and Ernest Fraenkel. An improved map of conserved regulatory sites for saccharomyces cerevisiae. *BMC Bioinformatics*, 7:113, 2006. doi: 10.1186/1471-2105-7-113. URL http://dx.doi.org/10.1186/1471-2105-7-113.

Brendan Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21, Nov 2008. doi: 10.1038/456018a. URL http://dx.doi.org/10.1038/456018a.

H. Craig Mak, Lorraine Pillus, and Trey Ideker. Dynamic reprogramming of transcription factors to and from the subtelomere. *Genome Res*, 19(6):1014–1025, Jun 2009. doi: 10.1101/gr.084178.108. URL http://dx.doi.org/10.1101/gr.084178.108.

Kriston L McGary, Insuk Lee, and Edward M Marcotte. Broad network-based predictability of saccharomyces cerevisiae gene loss-of-function phenotypes. *Genome Biol*, 8(12):R258, 2007. doi: 10.1186/gb-2007-8-12-r258. URL http://dx.doi.org/10.1186/gb-2007-8-12-r258.

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, Oct 2002. doi: 10.1126/science.298.5594.824. URL http://dx.doi.org/10.1126/science.298.5594.824.

J. MONOD and F. JACOB. Teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb Symp Quant Biol*, 26:389–401, 1961.

Alan M Moses, Daniel A Pollard, David A Nix, Venky N Iyer, Xiao-Yong Li, Mark D Biggin, and Michael B Eisen. Large-scale turnover of functional transcription factor binding sites in drosophila. *PLoS Comput Biol*, 2(10):e130, Oct 2006. doi: 10.1371/journal.pcbi.0020130. URL http://dx.doi.org/10.1371/journal.pcbi.0020130.

Ville Mustonen, Justin Kinney, Curtis G Callan, and Michael Lässig. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci U S A*, 105(34):12376–12381, Aug 2008. doi: 10.1073/pnas.0805909105. URL http://dx.doi.org/10.1073/pnas.0805909105.

Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 320(5881):1344–1349, Jun 2008. doi: 10.1126/science.1158441. URL http://dx.doi.org/10.1126/science.1158441.

Lee A Newberg, William A Thompson, Sean Conlan, Thomas M Smith, Lee Ann McCue, and Charles E Lawrence. A phylogenetic gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics*, 23(14):1718–1727, Jul 2007. doi: 10.1093/bioinformatics/btm241. URL http://dx.doi.org/10.1093/bioinformatics/btm241.

Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–3814, Jul 2003.

Duncan T Odom, Robin D Dowell, Elizabeth S Jacobsen, William Gordon, Timothy W Danford, Kenzie D MacIsaac, P. Alexander Rolfe, Caitlin M Conboy, David K Gifford, and Ernest Fraenkel. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet*, 39(6):730–732, Jun 2007. doi: 10.1038/ng2047. URL http://dx.doi.org/10.1038/ng2047.

S. Ohno, U. Wolf, and N. B. Atkin. Evolution from fish to mammals by gene duplication. *Hereditas*, 59(1):169–187, 1968.

Zhengqing Ouyang, Qing Zhou, and Wing Hung Wong. Chip-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A*, 106 (51):21521–21526, Dec 2009. doi: 10.1073/pnas.0904863106. URL http://dx.doi.org/10.1073/pnas.0904863106.

Ivan Ovcharenko, Dario Boffelli, and Gabriela G Loots. eshadow: a tool for comparing closely related sequences. *Genome Res*, 14(6):1191–1198, Jun 2004. doi: 10.1101/gr.1773104. URL http://dx.doi.org/10.1101/gr.1773104.

Balázs Papp, Csaba Pál, and Laurence D Hurst. Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–197, Jul 2003. doi: 10.1038/nature01771. URL http://dx.doi.org/10.1038/nature01771.

Lourdes Peña-Castillo, Murat Tasan, Chad L Myers, Hyunju Lee, Trupti Joshi, Chao Zhang, Yuanfang Guan, Michele Leone, Andrea Pagnani, Wan Kyu Kim, Chase Krumpelman, Weidong Tian, Guillaume Obozinski, Yanjun Qi, Sara Mostafavi, Guan Ning Lin, Gabriel F Berriz, Francis D Gibbons, Gert Lanckriet, Jian Qiu, Charles Grant, Zafer Barutcuoglu, David P Hill, David Warde-Farley, Chris Grouios, Debajyoti Ray, Judith A Blake, Minghua Deng, Michael I Jordan, William S Noble, Quaid Morris, Judith Klein-Seetharaman, Ziv Bar-Joseph, Ting Chen, Fengzhu Sun, Olga G Troyanskaya, Edward M Marcotte, Dong Xu, Timothy R Hughes, and Frederick P Roth. A critical assessment of mus musculus gene function prediction using integrated genomic evidence. *Genome Biol*, 9 Suppl 1:S2, 2008. doi: 10.1186/gb-2008-9-s1-s2. URL http://dx.doi.org/10.1186/gb-2008-9-s1-s2.

Benjamin Prud'homme, Nicolas Gompel, Antonis Rokas, Victoria A Kassner, Thomas M Williams, Shu-Dan Yeh, John R True, and Sean B Carroll. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, 440(7087):1050–1053, Apr 2006. doi: 10.1038/nature04597. URL http://dx.doi.org/10.1038/nature04597.

Benjamin Prud'homme, Nicolas Gompel, and Sean B Carroll. Emerging principles of regulatory evolution. *Proc Natl Acad Sci U S A*, 104 Suppl 1:8605–8612, May 2007. doi: 10.1073/pnas.0700488104. URL http://dx.doi.org/10.1073/pnas.0700488104.

Csaba Pál, Balázs Papp, and Martin J Lercher. An integrated view of protein evolution. *Nat Rev Genet*, 7(5):337–348, May 2006. doi: 10.1038/nrg1838. URL http://dx.doi.org/10.1038/nrg1838.

Daniela Raijman, Ron Shamir, and Amos Tanay. Evolution and selection in yeast promoters: analyzing the combined effect of diverse transcription factor binding sites. *PLoS Comput Biol*, 4

(1):e7, Jan 2008. doi: 10.1371/journal.pcbi.0040007. URL http://dx.doi.org/10.1371/journal.pcbi.0040007.

Vasily Ramensky, Peer Bork, and Shamil Sunyaev. Human non-synonymous snps: server and survey. *Nucleic Acids Res*, 30(17):3894–3900, Sep 2002.

E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, Aug 2002. doi: 10.1126/science.1073374. URL http://dx.doi.org/10.1126/science.1073374.

Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 67(2 Pt 2): 026112, Feb 2003.

Tali Raveh-Sadka, Michal Levo, and Eran Segal. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res*, 19(8):1480–1496, Aug 2009. doi: 10.1101/gr.088260.108. URL http://dx.doi.org/10.1101/gr.088260.108.

James Ronald and Joshua M Akey. The evolution of gene expression qtl in saccharomyces cerevisiae. *PLoS One*, 2 (7):e678, 2007. doi: 10.1371/journal.pone.0000678. URL http://dx.doi.org/10.1371/journal.pone.0000678.

James Ronald, Rachel B Brem, Jacqueline Whittle, and Leonid Kruglyak. Local regulatory variation in saccharomyces cerevisiae. *PLoS Genet*, 1(2):e25, Aug 2005. doi: 10.1371/journal.pgen.0010025. URL http://dx.doi.org/10.1371/journal.pgen.0010025.

T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–6100, Oct 1990.

Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–176, Jun 2003. doi: 10.1038/ng1165. URL http://dx.doi.org/10.1038/ng1165.

Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thåström, Yair Field, Irene K Moore, Ji-Ping Z Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, Aug 2006. doi: 10.1038/nature04979. URL http://dx.doi.org/10.1038/nature04979.

Eran Segal, Tali Raveh-Sadka, Mark Schroeder, Ulrich Unnerstall, and Ulrike Gaul. Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, 451(7178):535–540, Jan 2008. doi: 10.1038/nature06496. URL http://dx.doi.org/10.1038/nature06496.

SGD. Saccharomyces genome database. http://www.yeastgenome.org/.

Rahul Siddharthan, Eric D Siggia, and Erik van Nimwegen. Phylogibbs: a gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, Dec 2005. doi: 10.1371/journal.pcbi.0010067. URL http://dx.doi.org/10.1371/journal.pcbi.0010067.

Saurabh Sinha and Martin Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, 30(24): 5549–5560, Dec 2002.

Saurabh Sinha and Martin Tompa. Ymf: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, 31(13):3586–3588, Jul 2003.

Richelle Sopko, Dongqing Huang, Nicolle Preston, Gordon Chua, Balázs Papp, Kimberly Kafadar, Mike Snyder, Stephen G Oliver, Martha Cyert, Timothy R Hughes, Charles Boone, and Brenda Andrews. Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell*, 21(3):319–330, Feb 2006. doi: 10.1016/j.molcel.2005.12.011. URL http://dx.doi.org/10.1016/j.molcel.2005.12.011.

Jörg Stelling, Steffen Klamt, Katja Bettenbrock, Stefan Schuster, and Ernst Dieter Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193, Nov 2002. doi: 10.1038/nature01166. URL http://dx.doi.org/10.1038/nature01166.

G. D. Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, Jan 2000.

G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-dna interactions. *Trends Biochem Sci*, 23(3):109–113, Mar 1998.

Huang-Mo Sung, Tzi-Yuan Wang, Daryi Wang, Yu-Shan Huang, Jen-Pey Wu, Huai-Kuang Tsai, Jengnan Tzeng, Chih-Jen Huang, Yi-Chen Lee, Peggy Yang, Joyce Hsu, Tiffany Chang, Chung-Yi Cho, Li-Chuan Weng, Tso-Ching Lee, Tien-Hsien Chang, Wen-Hsiung Li, and Ming-Che Shih. Roles of trans and cis variation in yeast intraspecies evolution of gene

expression. *Mol Biol Evol*, 26(11):2533–2538, Nov 2009. doi: 10.1093/molbev/msp171. URL http://dx.doi.org/10.1093/molbev/msp171.

Yuval Tabach, Ran Brosh, Yossi Buganim, Anat Reiner, Or Zuk, Assif Yitzhaky, Mark Koudritsky, Varda Rotter, and Eytan Domany. Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. *PLoS One*, 2(8):e807, 2007. doi: 10.1371/journal.pone.0000807. URL http://dx.doi.org/10.1371/journal.pone.0000807.

Amos Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res*, 16(8):962–972, Aug 2006. doi: 10.1101/gr.5113606. URL http://dx.doi.org/10.1101/gr.5113606.

Leonid Teytelman, Michael B Eisen, and Jasper Rine. Silent but not static: accelerated base-pair substitution in silenced chromatin of budding yeasts. *PLoS Genet*, 4(11): e1000247, Nov 2008. doi: 10.1371/journal.pgen.1000247. URL http://dx.doi.org/10.1371/journal.pgen.1000247.

D. Thieffry, A. M. Huerta, E. Pérez-Rueda, and J. Collado-Vides. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in escherichia coli. *Bioessays*, 20(5):433–440, May 1998. doi: 3.0.CO;2-2. URL http://dx.doi.org/3.0.CO;2-2.

Gert Thijs, Kathleen Marchal, Magali Lescot, Stephane Rombauts, Bart De Moor, Pierre Rouzé, and Yves Moreau. A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, 9(2):447–464, 2002. doi: 10.1089/10665270252935566. URL http://dx.doi.org/10.1089/10665270252935566.

Dawn Anne Thompson and Aviv Regev. Fungal regulatory evolution: cis and trans in the balance. *FEBS Lett*, 583(24): 3959–3965, Dec 2009. doi: 10.1016/j.febslet.2009.11.032. URL http://dx.doi.org/10.1016/j.febslet.2009.11.032.

Itay Tirosh and Naama Barkai. Two strategies for gene regulation by promoter nucleosomes. *Genome Res*, 18(7): 1084–1091, Jul 2008. doi: 10.1101/gr.076059.108. URL http://dx.doi.org/10.1101/gr.076059.108.

Itay Tirosh, Adina Weinberger, Dana Bezalel, Mark Kaganovich, and Naama Barkai. On the relation between promoter divergence and gene expression

evolution. *Mol Syst Biol*, 4:159, 2008. doi: 10.1038/msb4100198. URL http://dx.doi.org/10.1038/msb4100198.

Itay Tirosh, Sharon Reikhav, Avraham A Levy, and Naama Barkai. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*, 324(5927):659–662, May 2009. doi: 10.1126/science.1169766. URL http://dx.doi.org/10.1126/science.1169766.

Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yutao Fu, W. James Kent, Vsevolod J Makeev, Andrei A Mironov, William Stafford Noble, Giulio Pavesi, Graziano Pesole, Mireille Régnier, Nicolas Simonis, Saurabh Sinha, Gert Thijs, Jacques van Helden, Mathias Vandenbogaert, Zhiping Weng, Christopher Workman, Chun Ye, and Zhou Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–144, Jan 2005. doi: 10.1038/nbt1053. URL http://dx.doi.org/10.1038/nbt1053.

Amy Hin Yan Tong, Guillaume Lesage, Gary D Bader, Huiming Ding, Hong Xu, Xiaofeng Xin, James Young, Gabriel F Berriz, Renee L Brost, Michael Chang, YiQun Chen, Xin Cheng, Gordon Chua, Helena Friesen, Debra S Goldberg, Jennifer Haynes, Christine Humphries, Grace He, Shamiza Hussein, Lizhu Ke, Nevan Krogan, Zhijian Li, Joshua N Levinson, Hong Lu, Patrice Ménard, Christella Munyana, Ainslie B Parsons, Owen Ryan, Raffi Tonikian, Tania Roberts, Anne-Marie Sdicu, Jesse Shapiro, Bilal Sheikh, Bernhard Suter, Sharyl L Wong, Lan V Zhang, Hongwei Zhu, Christopher G Burd, Sean Munro, Chris Sander, Jasper Rine, Jack Greenblatt, Matthias Peter, Anthony Bretscher, Graham Bell, Frederick P Roth, Grant W Brown, Brenda Andrews, Howard Bussey, and Charles Boone. Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–813, Feb 2004. doi: 10.1126/science.1091317. URL http://dx.doi.org/10.1126/science.1091317.

Tanya Vavouri, Jennifer I Semple, Rosa Garcia-Verdugo, and Ben Lehner. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell*, 138 (1):198–208, Jul 2009. doi: 10.1016/j.cell.2009.04.029. URL http://dx.doi.org/10.1016/j.cell.2009.04.029.

Dennis P Wall, Aaron E Hirsh, Hunter B Fraser, Jochen Kumm, Guri Giaever, Michael B Eisen, and Marcus W Feldman. Functional genomic analysis of the rates of protein evolution. *Proc Natl Acad Sci U S*

A, 102(15):5483–5488, Apr 2005. doi: 10.1073/pnas.0501761102. URL http://dx.doi.org/10.1073/pnas.0501761102.

Daryi Wang, Huang-Mo Sung, Tzi-Yuan Wang, Chih-Jen Huang, Peggy Yang, Tiffany Chang, Yang-Chao Wang, Da-Lun Tseng, Jen-Pey Wu, Tso-Ching Lee, Ming-Che Shih, and Wen-Hsiung Li. Expression evolution in yeast genes of single-input modules is mainly due to changes in trans-acting factors. *Genome Res*, 17(8):1161–1169, Aug 2007. doi: 10.1101/gr.6328907. URL http://dx.doi.org/10.1101/gr.6328907.

Ting Wang and Gary D Stormo. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–2380, Dec 2003.

Matthew T Weirauch and Timothy R Hughes. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet*, 26 (2):66–74, Feb 2010. doi: 10.1016/j.tig.2009.12.002. URL http://dx.doi.org/10.1016/j.tig.2009.12.002.

Andrew Whitehead and Douglas L Crawford. Variation within and among species in gene expression: raw material for evolution. *Mol Ecol*, 15 (5):1197–1211, Apr 2006. doi: 10.1111/j.1365-294X.2006.02868.x. URL http://dx.doi.org/10.1111/j.1365-294X.2006.02868.x.

E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Véronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston, and R. W. Davis. Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429):901–906, Aug 1999.

Gregory A Wray. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*, 8(3):206–216, Mar 2007. doi: 10.1038/nrg2063. URL http://dx.doi.org/10.1038/nrg2063.

Zeba Wunderlich and Leonid A Mirny. Using the topology of metabolic networks to predict viability of mutant strains. *Biophys J*, 91 (6):2304–2311, Sep 2006. doi: 10.1529/biophysj.105.080572. URL http://dx.doi.org/10.1529/biophysj.105.080572.

Zeba Wunderlich and Leonid A Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet*, 25(10):434–440, Oct 2009. doi: 10.1016/j.tig.2009.08.003. URL http://dx.doi.org/10.1016/j.tig.2009.08.003.

Xiaohui Xie, Jun Lu, E. J. Kulbokas, Todd R Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals. *Nature*, 434(7031):338–345, Mar 2005. doi: 10.1038/nature03441. URL http://dx.doi.org/10.1038/nature03441.

Xiao yong Li, Stewart MacArthur, Richard Bourgon, David Nix, Daniel A Pollard, Venky N Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton, Cris L Luengo Hendriks, Hou Cheng Chu, Nobuo Ogawa, William Inwood, Victor Sementchenko, Amy Beaton, Richard Weiszmann, Susan E Celniker, David W Knowles, Tom Gingeras, Terence P Speed, Michael B Eisen, and Mark D Biggin. Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. *PLoS Biol*, 6(2):e27, Feb 2008. doi: 10.1371/journal.pbio.0060027. URL http://dx.doi.org/10.1371/journal.pbio.0060027.

Gaël Yvert, Rachel B Brem, Jacqueline Whittle, Joshua M Akey, Eric Foss, Erin N Smith, Rachel Mackelprang, and Leonid Kruglyak. Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nat Genet*, 35(1):57–64, Sep 2003. doi: 10.1038/ng1222. URL http://dx.doi.org/10.1038/ng1222.

Stefan Zeiser, H. Volkmar Liebscher, Hendrik Tiedemann, Isabel Rubio-Aliaga, Gerhard K H Przemeck, Martin Hrabé de Angelis, and Gerhard Winkler. Number of active transcription factor binding sites is essential for the hes7 oscillator. *Theor Biol Med Model*, 3:11, 2006. doi: 10.1186/1742-4682-3-11. URL http://dx.doi.org/10.1186/1742-4682-3-11.

Xinmin Zhang, Duncan T Odom, Seung-Hoi Koo, Michael D Conkright, Gianluca Canettieri, Jennifer Best, Huaming Chen, Richard Jenner, Elizabeth Herbolsheimer, Elizabeth Jacobsen, Shilpa Kadam, Joseph R Ecker, Beverly Emerson, John B Hogenesch, Terry Unterman, Richard A Young, and Marc Montminy. Genome-wide analysis of camp-response

element binding protein occupancy, phosphorylation, and target gene activation in human tissues. *Proc Natl Acad Sci U S A*, 102 (12):4459–4464, Mar 2005. doi: 10.1073/pnas.0501076102. URL http://dx.doi.org/10.1073/pnas.0501076102.