Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

| | |
|---|---|
| Master of Information Technology | Anna-Lena Kranz |
| born in: | Bielefeld |
| Oral-examination: | 28.03.2011 |

# Analyzing combinatorial regulation of transcription in mammalian cells

Referees: PD Dr. Rainer König
PD Dr. Stefan Wiemann

# Abstract

## Analyzing combinatorial regulation of transcription in mammalian cells

During development and differentiation of an organism, accurate gene regulation is central for cells to maintain and balance their differentiation processes. Transcriptional interactions between cis-acting DNA-elements such as promoters and enhancers are the basis for precise and balanced transcriptional regulation. In this thesis, proximal and distal regulatory regions upstream of all transcription start sites were considered *in silico* to identify regulatory modules consisting of combinations of transcription factors (TFs) with binding sites at promoters and enhancers. Applying these modules to a broad variety of gene expression profiles demonstrated that the identified modules regulate gene expression during mouse embryonic development and human stem cell differentiation in a tissue- and temporal-specific manner. Whereas tissue-specific regulation is mainly controlled by combinations of TFs binding at promoters, the combination of TFs binding at promoters together with TFs binding at the respective enhancers determines the regulation of temporal progression during development. The identified regulatory modules showed considerably good predictive power to discriminate genes being differentially regulated at a specific time interval. In addition, TF binding sites are immanently different for promoter and enhancer regions.

One example for combinatorial regulation of transcription in mammals is cholesterol biosynthesis. Cholesterol biosynthesis is regulated by the family of sterol regulatory element binding proteins (SREBPs) that control the expression of genes involved in the uptake and synthesis of cholesterol and lipids. However, SREBPs are weak transcriptional activators themselves and have been shown to work in co-operation with other transcription factors such as Sp1 transcription factor (SP1) and nuclear transcription factor Y (NF-Y). Although the metabolism for cholesterol biosynthesis is well described, it is assumed that many other proteins contribute to cholesterol homeostasis and cholesterol mediated homeostasis of the cell. In this thesis, an integrative approach was applied that allowed systematic identification of potential SREBP target genes. Candidate genes were identified by gene expression profiling of sterol-depleted cells and *in silico* prediction of SREBP, SP1, and NF-Y binding sites. With this, 99 putative SREBP target genes were identified among which a major portion of genes (21 genes) known to regulate cholesterol biosynthesis and 78 novel potential SREBP target genes were retrieved. Ten of the putative novel 78 SREBP target genes were selected for experimental validation and *slc2a6*, *c17orf59*, *hes6*, and *tmem55b* showed reduced mRNA expression after SREBP knockdown, indicating a regulatory role by SREBP in combination with SP1 and NF-Y.

Combinations of transcription factors are substantial to the understanding of regulation of transcription and enhancer function, can yield generic insights into tissue- and temporal regulation of gene expression, and can elucidate novel target genes involved in a specific pathway.

# Zusammenfassung

## Analyse kombinatorischer Regulation der Transkription in Säugetierzellen

Präzise Genregulation ist während der Entwicklung und Differenzierung eines Organismus äußerst wichtig, um die notwendige Homeostase während der Zellentwicklung und -differenzierung zu ermöglichen. Dabei bilden Interaktionen zwischen *cis*-wirkenden DNA-Elementen wie Promotern und Enhancern die Basis für eine abgestimmte Regulation der Transkription. In dieser Arbeit wurden proximale und weiter entfernte Regionen stromaufwärts aller Transkriptionsstartpunkte *in silico* betrachtet, um regulatorische Module vorherzusagen, die aus Transkriptionsfaktorkombinationen mit Bindestellen an Promotern und Enhancern bestehen. Die Anwendung auf verschiedene Genexpressionsprofile zeigte eine gewebe- und zeitspezifische Regulation der identifizierten Module in der embryonischen Entwicklung der Maus und in der menschlichen Stammzelldifferenzierung. Zusätzlich zur gewebespezifischen Regulation von Transkriptionsfaktorkombinationen am Promoter bestimmen Kombinationen von Transkriptionsfaktoren an Promotern und Enhancern zeitspezifische Regulation während des Entwicklungsprozesses. Die identifizierten regulatorischen Module zeigten eine gute Vorhersagefähigkeit, differenziell exprimierte Gene unterschiedlicher Zeitpunkte zu unterscheiden. Außerdem wurde gezeigt, dass Transkriptionsfaktorbindestellen unterschiedliche Eigenschaften an Promoter- und Enhancerregionen aufzeigen.

Ein Beispiel für kombinatorische Regulation der Transkription in Säugetierzellen ist die Cholesterinbiosynthese. Die Cholesterinbiosynthese wird durch die SREBP (*sterol regulatory element binding protein*) Proteinfamilie kontrolliert, die die Expression von Genen regulieren, die in der Aufnahme und Synthese von Cholesterin und Lipiden involviert sind. SREBPs sind nur schwache transkriptionelle Aktivatoren und kooperieren mit anderen Transkriptionsfaktoren wie SP1 (Sp1 transcription factor) und NF-Y (nuclear transcription factor Y). Obwohl der Metabolismus der Cholesterinbiosynthese gut beschrieben ist, wird angenommen, dass viele weitere noch unbekannte Proteine an der Cholesterinhomeostase der Zelle beteiligt sind. Daher wurde in dieser Arbeit ein integrativer Ansatz verfolgt, um neue Zielgene von SREBP zu identifizieren. Dazu wurden Genexpressionsprofile von sterol-depletierten Zellen mit *in silico* Vorhersagen von SREBP, SP1, und NF-Y Bindestellen kombiniert. Insgesamt wurden 99 mögliche SREBP Zielgene identifiziert, von denen 21 Gene bereits im Zusammenhang mit Cholesterin beschrieben wurden und 78 Gene potentiell neue SREBP Zielgene darstellen. Zehn der potenziell neuen Zielgene wurden für eine experimentelle Valdierung ausgewählt, wovon *slc2a6*, *c17orf59*, *hes6*, and *tmem55b* niedrigere mRNA Expression nach SREBP Knockdowns zeigten und damit potentiell regulatorisch von SREBP abhängig sind.

Kombinationen von Transkriptionsfaktoren sind äußerst wichtig, um sowohl Regulationsmechanismen der Transkription als auch die Funktion von Enhancern zu verstehen. Sie können neue Erkenntnisse über die gewebe- und zeitspezifische Regulation der Genexpression bringen und die Identifizierung neuer Zielgene in bestimmten Prozessen ermöglichen.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

There exist about 20,000-25,000 protein-coding genes in human [66]. However, the number of DNA-binding factors is estimated to be ~1850 [179]. Multiple regulatory elements within promoters allow combinatorial control of transcriptional regulation, which increases the potential number of expression profiles notably [179]. Transcription factors (TFs) cooperate with other regulatory co-factors and the complex combinations of multiple cooperative interactions give the necessary specificity for spatio-temporal transcriptional regulation [174]. The combinatorial and temporal binding of TFs is crucial for metazoan development [310] and for the establishment of tissue specific gene expression [229].

Specifically in higher organisms, proximal versus distal regulation needs to be well balanced [147]. Whereas promoters are proximal to transcription start sites (TSS), enhancers can be quite distant from their target genes. Transcription factors bound at an enhancer interact with co-activators and transcription factors bound at the promoter. Hence, they increase the concentration of activators at promoters. The large distance between long-range enhancers and proximal promoters can be overcome by chromatin loops, bringing these elements in close proximity [118, 196]. Thus, enhancers can increase the activity of a promoter considerably, even when located several kilo bases away.

Promoter-enhancer interactions depend on regulatory factors binding at promoter-proximal regions. In turn, these factors may recruit specific distal enhancers [49, 267] depending on the combination of regulatory factors at the proximal promoter [164]. Levine and Tjian [164] suggested that a combination of different complexes are needed for a temporal- and tissue-specific regulation of *cis*-DNA elements, allowing a vast variety of distinct gene expression patterns. One of the main challenges is now to understand how different combinations of transcription factors establish gene expression under specific conditions [179].

One example of combinatorial regulation in mammals is the transcriptional regulation of cholesterol biosynthesis. Cholesterol biosynthesis in mammals is regulated by the family of sterol regulatory element binding proteins (SREBPs) that control the expression of genes involved in the uptake and synthesis of cholesterol and lipids. However, SREBPs are weak transcriptional activators themselves and work in co-operation with other transcription factors such as Sp1 transcription factor (SP1) and nuclear transcription factor Y (NF-Y) [72, 134, 234, 241]. Although the metabolism for cholesterol biosynthesis is well described, it is assumed that many other proteins contribute to cholesterol homeostasis and cholesterol mediated homeostasis of the cell [121]. In addition, dysfunction of cholesterol has been implicated in a number of diseases such as cardiac diseases, dementia, diabetes, and cancer [127, 185]. Furthermore, diseases caused by dysfunctional cholesterol uptake and synthesis like Familial Hypercholesterolemia and Niemann-Pick Disease Type C are poorly understood and open questions remain concerning the regulation of cholesterol metabolism and molecular interactions of known cholesterol-regulating factors [22].

The focus of this thesis is twofold. First, combinations of transcription factors at promoter and enhancer regions were analyzed and immanent differences of enhancers and promoters affecting the regulation of genes during critical developmental stages of different tissues and cell types identified. Differentially expressed genes at specific time intervals of the development of each analyzed tissue were associated to their regulating TFs. The systematic comparison of temporal and tissue specificity of TFs and combinations of TFs binding at promoters and enhancers revealed a major role of the combinations of TFs in promoter regions together with TFs in enhancer regions for temporal specificity in development and differentiation. Second, an integrated approach was applied to identify novel putative SREBP target genes by combining gene expression profiling and *in silico* predictions of binding sites for SREBP and its known co-factors SP1, NF-Y, and LXR. This approach allowed the identification of 78 genes that have not yet been described to be involved in cholesterol biosynthesis. We picked ten genes for experimental validation out of which four genes showed lower mRNA expression after knockdown of SREBP, indicating their regulatory role depending on SREBP.

### 1.1.1 Publications

The main work presented in this thesis is currently in review and the manuscript is entitled “Enhancers regulate progression of development in mammalian cells”. A manuscript covering the second part of the thesis (“Identification of novel putative SREBP target genes”) is currently in preparation. I was also involved in a publication “PathWave: discovering patterns of differentially regulated enzymes in metabolic pathways” published in Bioinformatics [247].

### 1.1.2 Thesis outline

The first chapter introduces the biological background, existing computational methods, and basic topics covered in this thesis. The focus lies on transcriptional regulation in eukaryotes and the computational identification of transcription factor binding sites. Additionally, basic concepts in machine learning and network analysis are presented. In chapter 2, the methods applied in this thesis are presented, specifically the developed approach to identify regulatory modules and successive analyses as well as the integrated approach for the identification of novel putative SREBP target genes. Chapter 3 depicts the results of this thesis and presents regulatory modules that control gene expression during development and differentiation in a time- and tissue-specific manner as well as potential new SREBP target genes. The results are discussed and an outlook is given in chapter 4.

## 1.2 Organization and regulation of eukaryotic genomes

The genome, the complete set of information in the DNA of a cell or an organism, not only encodes all genes, but also contains the information to express these genes in a spatio-temporal manner [260]. The central dogma in science regarding gene expression is that genes encode mRNA, which in turn encodes proteins [285]. The genetic information is encoded in its deoxyribonucleic acid (DNA) sequence. DNA is a large polymer composed of four different nucleotide subunits, each consisting of a five-carbon sugar (deoxyribose) attached to a single phosphate group and a nitrogen-containing base. The four different bases are called adenine (A), cytosine (C), guanine (G), and thymine (T). These bases are connected via a sugar phosphate backbone through the 3’-hydroxyl group of a sugar to the 5’-phosphate group of another sugar. Therefore, one end of the DNA carries an unlinked hydroxyl

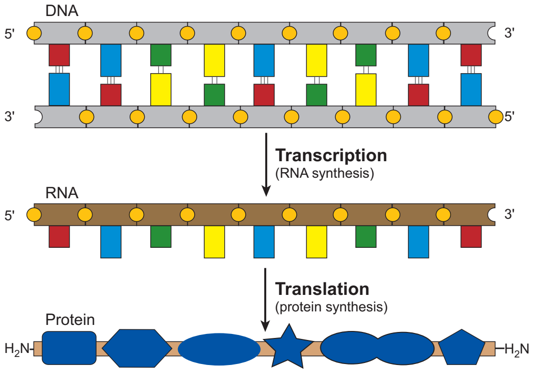


**Figure 1.1. From DNA to protein.** Genetic information is passed from DNA to RNA in a process called transcription and from RNA to protein in a process called translation.

group to the 3' position on the sugar ring (3' end) and the other end carries a free phosphate group at the 5' position on the sugar ring (5' end). It is the order of the base pairs from its 5' end to its 3' end that encodes the genetic information of a cell. Each DNA strand is paired with a complementary DNA strand which are held together via hydrogen bonds formed between the bases. This pairing is specific with an adenine forming two hydrogen bonds with a thymine and a guanine forming three hydrogen bonds with a cytosine. Chemical and structural features of the DNA chains force the DNA into the typical DNA double helix, the predominant form in the cell. [3]

Genomes of higher eukaryotes can be up to several billion base pairs long, e.g. the human genome has a length of around 3 billion base pairs [65]. Nowadays, a large number of genomes has been successfully sequenced and the number of available DNA sequences is rapidly increasing due to new high throughput sequencing techniques. The challenge now is to decipher the genetic code embedded in the genomic sequence [65].

A large number of genes, information containing-elements of DNA that determine characteristics of an organism by encoding functional cellular components such as proteins [3], have been already identified. The sequence of a gene is used as a template to synthesize ribonucleic acid (RNA) molecules in a process called transcription. RNA is chemically similar to DNA but contains a sugar ribose instead of a deoxyribose and the base uracil (U) instead of thymine. In addition, RNA is a single-stranded molecule. The RNA molecule copied from protein-coding genes is called messenger RNA (mRNA). It is used as a messenger molecule for the production of proteins by being translated into a chain of amino acids. Genes can be separated into distinct protein-coding regions (exons) and intervening non-coding regions (introns). The mRNA is formed by joining different exons whereas introns get excised from the primary transcript. Alternative splicing adds another layer of complexity of gene expression regulation by allowing the incorporation of different exons from the same primary transcript [194]. Once the mRNA has been formed, a poly-A tail is added to the 3' end of the mRNA and a cap structure is added to its 5' end. The mature mRNA is then transported out of the nucleus into the cytosol where it is translated into a protein, a long polymer chain of monomeric building blocks called amino acids. [3] The information flow from DNA to proteins is shown in Figure 1.1.

In higher eukaryotes, genes are embedded into large regions of non-coding DNA, e.g. only 1-2% of the DNA encodes for genes while most of the genome does not and has yet mostly unknown function [156]. It may be involved in regulatory processes [65, 164, 181]. Identifying regulatory regions and in particular transcription factor binding sites located many kilo bases away from their corresponding genes in the vast stretches of non-coding DNA remains a major challenge.

### 1.2.1 Measuring gene expression with microarrays

Expression microarrays allow the simultaneous measurement of transcription levels for every known gene in an organism [142] and have been used to identify disease specific gene signatures (for cancer e.g. [62, 279, 294]). Most microarray platforms are designed to address a specific set of questions in a specific organism [117]. The term microarray analysis is usually used for transcript analysis [117] and microarrays are commonly used to compare levels of expression of genes from samples from two different tissues or at two distinct experimental conditions (e.g. normal vs diseased tissue, treated vs untreated samples) [41]. However, microarrays can also be used for genotyping, epigenetic studies, structural variation, splice-variant analysis, and protein binding [117]. Although transcriptional profiling is the most widely used application, it focuses on a biological intermediate [117] with transcription being the first step in gene regulation. Still, the correlation between mRNA and protein abundance in the cell is not straightforward [8, 56, 109].

Microarrays are glass slides with artificially constructed grids of DNA [47]. One array can contain up to hundreds of thousands of spots which consist of what are known as probe sequences. The probes can be single-stranded cDNA (complementary DNA, the reverse transcribed product of mRNA) and long oligonucleotides (60-70 bp) or short oligonucleotides (25 bp as with Affymetrix® arrays) [11]. There are other differences between oligonucleotide microarrays and spotted cDNA microarrays. Affymetrix® arrays contain between 11 and 20 pairs of oligonucleotide probes per target RNA. One of each pair is the reverse complement to an ideally unique 25mer in the RNA and the other contains a single mutation [5]. In contrast, cDNA microarrays contain a single probe for each target RNA and the two different biological samples are represented by different colors. After hybridization, each color is scanned separately and relative expression levels are achieved by comparing the intensities [47]. For oligonucleotide microarrays, each microarray represents a single sample and provides an absolute measurement level for each RNA molecule, whereas for cDNA microarrays each microarray measures two samples and provides a relative measurement level for each RNA molecule [47].

A microarray experiment involves a number of distinct stages. First, RNA is extracted from biological samples, amplified and labeled with a fluorescent dye. It can then be hybridized to the arrays and the microarrays are processed to get intensities by scanning the arrays [257]. Following, the target quantity is measured indirectly by measuring the intensity of fluorescence of the spots on the array for each fluorescent dye [5]. The raw data produced by microarray experiments are monochrome images, that need to be transformed into gene expression matrices. The intensity is read by a camera and transformed into gray level values using image processing [61, 76, 244, 262]. In the end, one gets 4,000-50,000 measurements per biological sample [47].

### Microarray analysis

Biological replicates are essential to estimate and reduce measurement variability and biological differences between the cases [5]. However, technical replicates estimate and reduce only effects of measurement variability and are not required when making inferences about populations from samples [5]. To analyze differentially expressed genes under two conditions, at least five biological cases per group should be analyzed but larger numbers are preferable [215].

To make microarrays comparable and to reduce noise, the intensities must be quality-controlled and normalized to adjust for dye-bias and for any systematic variation in the technology. Intra- and inter-microarray variations can skew the interpretation of such expression data [47]. In addition, hybridization images can contain artifacts, such as bubbles and scratches [245]. A first quality assessment is done by a visual inspection of the images and plots of the raw data [11]. So called MA-plots are commonly used to plot the log ratios M with

$$M = \log_2 \frac{R}{G} \tag{1.1}$$

against the average intensity values A with

$$A = \frac{1}{2} \log_2(R \cdot G) \tag{1.2}$$

where $R$ and $G$ represent the intensity levels for a given spot. MA-plots give a good impression of the distribution of the raw data. As most genes are not expected to be differentially expressed, the majority of points should lie in a cloud around $M = 0$.

Results from individual experiments need to be normalized with respect to each other to account for experimental variation in RNA amounts, specific activity of probe labels, and standard handling errors [59] and is an important step of the preprocessing of microarray data. Individual intensities are adjusted to be able to make comparisons both within the array as well as between arrays in the experiment. These adjustments are necessary to remove purely technical differences that do not represent biological variation and to be able to identify true differentially expressed genes [11]. Normalization can include the adjustment of the overall brightness of each scanned microarray image [300], using expression levels of housekeeping genes [75] or assuming that most genes are not differentially expressed [309]. Various normalization approaches have been developed [75, 122, 222, 256, 309]. After normalization is completed, differentially expressed genes or functional groups classifying the samples into meaningful groups can be identified [47, 257].

The earliest approach to identify genes whose expression is significantly different between the two conditions is a simple fold-change criterion to detect genes of interest. However, it is perceived as an inadequate test statistic [190] as it does not incorporate variance. More sophisticated statistical tests achieve more reliable identification of differentially expressed genes [41]. Significance has been evaluated in many different ways, including parametric [274] and non-parametric tests [214], analysis of variance [139], and many others [47]. Common statistics for differential expression are the $t$-statistic and its non-parametric counterpart the Wilcoxon statistics [11]. Another method to identify differentially expressed genes is based on calculating **rank products** which is similar to the fold-change but overcomes its most significant limitations [41]. It is based on the assumption that the probability of gene being differentially expressed increases with the number of times a gene is differentially expressed in replicate experiments.

The rank product is defined as

$$RP_g = (\prod_{i=1}^{k} r_{i,g})^{\frac{1}{k}} \tag{1.3}$$

where $r_{i,g}$ is the position of gene $g$ in the list of genes in the $i$th replicate sorted by decreasing fold change for an experiment examining $n$ genes in $k$ replicates [41]. Significance is then assessed by permuting the expression values of the genes for each single array.

Testing thousand of genes of transcripts for differential expression simultaneously produces a large amount of false-positives. This testing of many hypotheses within a single study is called multiple testing and different multiple testing correction approaches have been developed [5]. One of the most popular multiple testing corrections includes the Bonferroni correction [35] where the p-value of each gene is multiplied with the total number of genes. A less stringent correction is the Benjamini-Hochberg correction [26] where the p-value of each gene is multiplied by the total number of all genes divided by the rank of the p-value compared to all p-values.

An interesting development is the testing of groups of genes instead of single genes where one is interested in a group-wise effect (e.g. [99, 247]) which leads to an increase of power [11]. These genes usually share common features, such as all genes from a pathway.

### 1.2.2 Regulation of gene expression

Spatial and temporal expression of genes is crucial for development, differentiation, and all biological processes [179]. There exist various different mechanisms to regulate eukaryotic gene expression at various steps, including transcription, mRNA splicing and processing, transport, translation, stability, post-translational modification of proteins, and degradation [179]. However, transcription initiation is believed to be the most regulated step [179]. Transcription is regulated by *cis*-regulatory elements, such as promoters and distal regulatory elements, e.g. enhancers, silencers, insulators, and locus control regions [179]. These elements contain recognition sites for trans-acting proteins binding to these elements [158, 243] that either repress or enhance transcription [179]. The genetic information encoded in the DNA of eukaryotic genes requires the regulated synthesis of specific RNAs by molecular machines called RNA polymerases. The protein-coding genes are transcribed by RNA polymerase II [90, 243] and transcription initiation requires, in addition to RNA polymerase II, the binding of regulatory elements and co-factors to *cis*-regulatory sequences [239]. It is the interplay between promoters, proximal

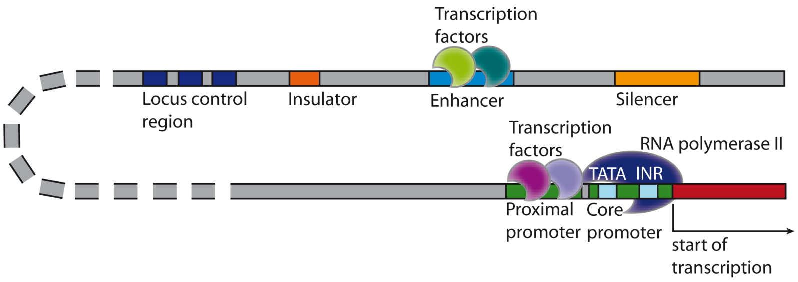


**Figure 1.2. Overview of transcriptional regulatory elements in the genome.** The promoter is composed of a core and proximal promoter. General transcription factors bind to core promoter regions through recognition of common elements such as TATA boxes and initiators (INR). Promoter activity can be increased by site-specific transcription factors binding to proximal promoter regions. Distal regulatory elements can include enhancers, silencers, insulators and locus control regions. Promoter activity can be further stimulated by site-specific factors binding to enhancers. In contrast, transcriptional activity can be repressed by transcription factors binding to silencers. Enhancer blocking insulators ensure enhancer interaction with the promoter of the right gene.

and distal regulatory elements as well as their binding factors and cofactors that contribute to the precise nature of the transcriptional output of any promoter [179]. Figure 1.2 gives an overview of all regulatory elements involved in transcription.

**Promoter** Core and nearby proximal promoters are the basic elements that regulate transcription [239]. They contain binding sites for transcription factors and common sequence elements that recruit the general transcriptional machinery to the transcription start site (TSS) [81] as well as additional chromatin-modifying factors [81, 158]. The region around the TSS is referred to as the core promoter and is approximately 100 bp in length. It is sufficient for directing transcription initiation by the basal transcriptional machinery [158] and defines the position of the transcription initiation site and direction of transcription [255]. It contains an AT-rich site called the TATA box [158] which is located 25 to 30 bp upstream of the TSS in higher eukaryotes [265] and serves as the binding site for the TATA-binding protein (TBP) [158]. The core promoter can also contain an initiator element (Inr) [254]. Factors binding to Inr may facilitate the recruitment of the transcriptional machinery [50, 137]. Promoters in higher eukaryotes are highly diverse. Core promoters can contain either element, Inr or TATA box, both elements or neither element [158]. Core promoters of many genes do not contain any of the known core promoter elements [102] and TATA-containing promoters are rather the exception than the rule [65]. So called null promoters often have multiple TSS [96, 170] that are in close proximity to each other [243] and most human genes have alternative promoters [52, 143] which are thought to be used in different contexts and tissues. In mammals, promoters containing a TATA box are usually associated with tissue- or context-specific genes [248] and require a finer regulation [53]. However, the majority of human genes (80-90%) have promoters close to CpG islands [141], stretches of DNA 500-2kb in length with high C+G content [179]. Methylation of these islands is associated with transcriptional silencing and blocks transcription factor binding to their recognition sequences [179]. Promoters containing CpG islands are thought to be involved in the regulation of ubiquitously expressed genes [114]. The proximal promoter lies immediately upstream from the core promoter (up to a few hundred bais pairs) and contains multiple binding sites for activators [179]. However, the distinction between proximal and core promoter elements may be blurred in mammals [206], as sequence-specific TFs might also contribute to the positioning of RNA Polymerase II at the TSS [187].

**Enhancer** Similar to promoters, enhancers contain binding sites for transcription factors but can be located far upstream from the TSS [81]. An enhancer is sometimes also referred to as any regulatory element with binding sites for sequence-specific transcription factors [13] but the term is mostly used for distal regulatory regions. Enhancers usually contain clusters of DNA-binding sites for more than one type of transcriptional activators [13, 158]. They influence transcription independent of their orientation and distance from the TSS, which can be upstream (as great as 85 kb), downstream of the promoter in an intron, or even beyond the 3' end of a gene [32]. Enhancers are highly modular and a single promoter can be acted upon various enhancers in a time- and tissue-specific manner [15]. Enhancers can increase transcriptional activity indirectly through chromatin remodeling or directly through interactions with the general transcriptional machinery [32].

There exist several enhancer models. In one model, binding of multiple transcriptional regulators to the enhancer can lead to the formation of enhanceosomes [158]. The stability and function of an enhanceosome is dependent on the arrangement of binding sites, the interaction of the activators and the addition of architectural proteins [51]. In contrast, the billboard enhancer model is more flexible with independent interactions of each binding factor with the basal machinery and exact binding site locations are less critical [14].

Interaction of enhancers with the promoter occur by looping out the intervening DNA between these elements [179].

**Silencer** Silencers are similar to enhancers but repress promoter activity in an orientation- and position-independent manner instead of enhancing transcription [204]. Repressors binding to silencers can inhibit transcription through different mechanisms, e.g. interfering with activator binding, preventing recruitment of the transcriptional machinery, and modifying the chromatin structure [81, 204]. Methylation of a CpG dinucleotide motif has been implicated in silencing in higher eukaryotes [149].

**Insulator** Enhancer blocking insulators are negative regulatory elements lying between an enhancer and the promoter [239]. They can also block genes from being affected by the transcriptional activity of neighboring genes and prevent spreading of repressive chromatin [179].

**Locus control regions** Similar to enhancers, locus control regions (LCRs) contain multiple binding sites for activators [158]. Whereas enhancer function can be diminished by the chromatin structure of the site of integration, LCRs can stimulate transcription independent of the chromatin structure but are limited by orientation and distance [84]. LCRs consists of groups of regulatory elements involved in regulating an entire gene cluster [179].

**Chromatin structure** To fit DNA into the nucleus, DNA is packaged into a nucleoprotein complex known as chromatin [158]. The repeating unit of chromatin is the nucleosome which contains 146 bp of DNA wrapped 1.65 turns around an octamer of histone molecules [158]. Higher-order chromatin structure is composed of nucleosomes coiled into chromatin fibers [312]. Histones can be modified in various ways, including acetylation, methylation, and phosphorylation [239]. These modifications can epigenetically control the expression of genes by controlling the accessibility of chromatin [239]. Transcriptionally active chromatin (euchromatin) contains many sites which are hypersensitive to DNases open for transcription [104]. Nucleosomes can prevent transcription initiation by restricting access of transcriptional regulators to the DNA [158]. Removing nucleosomes can enhance binding of activators and the transcriptional machinery [151, 130]. The actual TSS region has been shown to be free of nucleosomes [17].

Binding of general transcription factors to the core promoter results usually just in low transcriptional activity. The basal machinery can be defined as those factors that are essential for basal transcription *in vitro* from an isolated core promoter [255]. Tissue-specific and developmentally regulated expression of genes requires a finer tuning, and the recruitment of additional transcriptional elements located in upstream, intronic, or downstream regions [299]. Site-specific factors binding to the proximal promoter can support the recruitment or stabilization of interactions of general factors at the core promoter, thereby increasing transcriptional activity [81]. Binding of regulatory factors to distal enhancer regions and recruitment of histone-modifying enzymes such as Swi/Snf and SAGA (PCAF) to promoters [126, 218] can generate a more favorable environment for transcription which leads to a further increase in promoter activity [81]. The general transcriptional machinery includes subunits of RNA polymerases and complexes such as TFIID [81] and co-activators. An RNA polymerase II-holoenzyme consisting of general transcription factors and a multi protein complex called the Srb/Mediator is recruited by transcriptional activators [175]. The mediator complex provides activator targets and can integrate multiple regulatory signals [158]. RNA polymerase II has been shown to contain 10-12 subunits with diverse functions including start site selection, transcriptional elongation rates, and interactions with activators [10]. RNA polymerase II is associated with elongation factors [232] and protein complexes involved in RNA capping, polyadenylation and splicing [28].

Co-activators may support the action of activators by protein-protein interactions with DNA-bound activators [258]. These co-activators can play a crucial role in regulation and can switch activators to repressors [160]. Multiple copies of the same factor or different factors can act synergistically to increase transcription greater than the sum of individual activity [179].

Transcriptional regulation is a balanced process of positive and negative regulators [158]. Transcriptional activators and repressors modify the chromatin structure to make it accessible to the transcriptional machinery and recruit the initiation apparatus to promoters [158]. The transcriptional apparatus can be recruited in multiple steps or in a single step if already fully assembled [158]. Activators can also increase the elongation rate for polymerase by stimulating the rate of promoter escape, its processivity, or facilitate reinitiation of transcription [305]. Repressors can compete with activators for binding sites or can interact with components of the transcriptional machinery and chromatin [204]. Transcription consists of a series of steps, including promoter melting, clearance, and escape, before a fully functional RNA polymerase II elongation complex is formed. After the formation

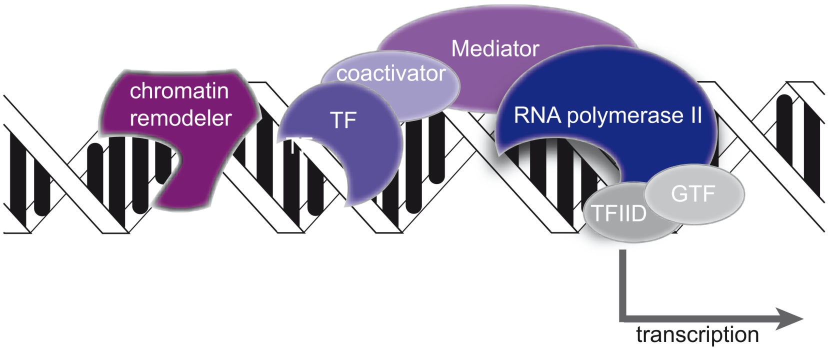


**Figure 1.3. Transcription initiation by RNA polymerase II.** RNA polymerase II and various general transcription factors (GTF), e.g. TFIID, form the pre-initiation complex around the transcription start site. Other regulatory proteins, such as Mediator, chromatin remodelers, coactivators, and sequence-specific transcription factors (TFs), are involved in transcriptional regulation.

of a stable transcription initiation complex (Figure 1.3), the promoter is cleared and elongation is induced to produce an RNA transcript [158]. RNA polymerase II is phosphorylated to switch from initiation to elongation and cofactors associated with the polymerase are exchanged [158].

After primary transcripts are produced by RNA polymerase II, they are modified at both ends and are subjected to splicing [158]. 5' end modification is crucial for further processing, localization, and translation [281]. 5' ends are capped with a methylated guanosin triphosphate [158]. In contrast, 3' ends are cleaved and polyadenylated [158]. This modification is essential for transcript termination, transport, translation, and stability of the transcript [307]. mRNA is proofread and aberrant RNA molecules are degraded [116]. Mature mRNA is then exported to the cytoplasm [198].

Regulatory factors located in promoter proximal regions do not always activate or repress transcription in the classical sense but serve as tethering elements recruiting distal enhancers to the core promoter [49]. It is possible that some regulatory factors recruit some enhancers, while other combinations recruit other enhancers [164]. Different complexes of regulatory elements are necessary for temporal- and tissue-specific regulation of distinct *cis*-DNA elements [164]. These different complexes binding at promoter and enhancer elements offer a variety of combinations of distinct gene expression patterns [164].

### 1.2.3 Transcription factors

Transcription factors can be described as proteins that regulate the production of RNA in different modes of action [182]. They regulate the expression of genes by binding to sequence-specific binding sites that can occur genome-wide [182]. Transcription factors rarely work alone but cooperate with each other either by direct interaction or through co-activators or co-repressors [81] and transcription factor bindings sites were identified to cluster together to regulate transcription cooperatively [29]. It is yet unknown what determines the set of binding sites bound in a particular context but the influence of chromatin accessibility is assumed to play a major role [48].

There exist approximately 200-300 transcription factors that bind to core promoter elements and 1400 sequence-specific transcription factors. Transcription factors consist of two functional domains, a DNA-binding and an activation domain that is crucial for the TF to stimulate transcription [179, 192]. There exist various classes of transcription factors with more than 100 known DNA-binding domains and specific DNA binding sequences [179] and approximately 12 to 15 structurally distinct DNA-binding domains are known from eukaryotic TFs [112]. Grouping of TFs according to their binding domain can provide insights into their function, e.g. homeodomain containing TFs are often involved in developmental processes [173]. The DNA-binding domain can be composed of contiguous amino acids (e.g. homeodomain, MADS box) or dispersed within the primary sequence (e.g. Zn-fingers) [299]. Three types of TF domains account for 80% of the repertoire in the human and mouse genome: the $C_2H_2$zinc-finger, homeodomain and helix-loop-helix [103, 280].

The binding sequence is a rather short and degenerate sequence of 6-12 bp which is described in a consensus sequence whereas the binding specificity is dictated by just 4-6 bp [179]. Differences in the binding sequence can affect the strength of the activator binding which can have implications for crucial situations such as embryonic development when transcription factors are distributed in a concentration gradient [179]. Specific factors determine the set of genes to be transcribed and are either ubiquitously expressed or are rather tissue-specific [280].

Transcription factors can bind close or far away from regulated genes, upstream, downstream or in the introns of the genes they regulate [45]. Transcription factors binding in close proximity to the TSS are thought to stabilize general transcription factors at core promoter elements, whereas transcription factors binding to distal regions of a gene induce interactions between distal elements and the general transcriptional machinery bound at the TSS [81]. The analysis of 1% of the human genome has shown that transcription factors binding primarily at proximal promoter regions are rather the exception than the rule [65, 54, 55] as less than 10% of the analyzed factors had the majority of their binding sites within 2.5kb of a TSS [65], e.g. YY1 [114] and E2F1 [31]. Distal sites have been shown to be involved in tissue-specific regulation [239]. Transcription factors binding rather at distal positions include p53 [55], ER [54], NF$\kappa$B [178], CREBP [80], and STAT1,2 [113]. The genomic distribution patterns of these factors resemble distal regulatory elements such as enhancers [114].

General transcription factors required for promoter binding by RNA polymerase II in vitro include TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH [211]. The mediator provides targets for transcriptional activators and by passing signals to RNA polymerase II and components of the initiation apparatus [158].

## 1.3 Identification of transcription factor binding sites

One of the challenges in genomic research is the identification of all functional elements, including those factors that regulate gene expression. The identification of these elements will help to uncover mechanisms in various diseases [179]. The availability of genome sequences of various organisms, microarrays and the development of computational methods have facilitated the identification of *cis*-regulatory sites on a genome-wide scale [52, 53, 55] and allow the analysis of transcription factors under various conditions [202].

### 1.3.1 Experimental techniques

The identification of transcription factor binding sites using experimental approaches is crucial for a better understanding of biological function of TFs, for the analysis of tissue- and temporal-specific effects on gene expression [164], and for an improvement of computational predictions [77].

If the regulatory factors are unknown, alterations of chromatin structure and experimental manipulation of defined DNA segments can lead to the identification of the location of a functional element [77]. Direct measurements of interactions of a regulatory factor with the DNA provide more precise information when the TF involved has been identified [77].

Traditionally, TF binding specificity has been determined using footprinting methods that identify the DNA region protected by a bound protein, nitrocellulose binding assays, gel-shift analysis, or reporter constructs [45]. Several sequencing-based high-throughput methods on a genome-wide scale require the isolation of cDNAs, sequencing of their 5' ends and mapping of these fragments to a genomic DNA sequence [243]. Chromosome conformation capture can be used to identify interactions between chromosomal regions such as regulatory elements to their corresponding transcript [301].

**Reporter assays** Reporter assays are the classical experimental approach to characterize regulatory elements. The candidate transcriptional regulatory element, a promoter, and a reporter gene are tested *in vitro* or *in vivo* in a synthetic construct [138, 146] and the change in production of the reporter protein in response to the candidate regulatory element is measured [77].

**EMSA** Electrophoretic mobility shift assay (EMSA) is a traditional approach for the identification of interactions between DNA and proteins [87, 93]. This 'gel-shift' assay can be used to verify the ability of a protein to recognize and bind a target DNA sequence [77].

**DNaseI hypersensitivity** DNase hypersensitive sites, nucleosome-depleted regions digested by DNAseI, are markers for functional regions in non-coding sequences [57, 104]. DNase hypersensitivity assays map changes in chromatin structure [77] and detect sites of open chromatin likely to contain functional transcription factor binding sites by sequencing the flanking sites of DNaseI cleavage sites [111, 239].

**ChIP** Chromatin immunoprecipitation is one of the most powerful experimental techniques for the *in vivo* mapping of DNA-associated proteins when the regulatory factor is known [77, 153]. Binding sites for a specific TF can be isolated by antibodies that recognize the specific TF bound to their target DNA [239]. A large number of binding regions can be identified at the same time in living cells [239]. This assay captures *in vivo* interactions between DNA and a protein by cross-linking proteins to their DNA recognition site [77]. The cells are lyzed and DNA is fragmented into small pieces followed by immunoprecipitation using a TF-specific antibody [77]. Reversal of the cross-linking releases the DNA for subsequent detection by PCR amplification [77]. ChIP is used to identify DNA-bound factors on a genomic scale by hybridizing DNA sequence segments to promoter or tiling microarray (ChIP-chip) [233, 250]. The identified ChIP products can also be identified by ultra-high-throughput sequencing (ChIP-seq) [21].

**ChIP-chip** High-throughput variations of the ChIP technique amplify all genomic sequences with binding sites for the given protein [77]. These sequence fragments are then hybridized to a genome-wide tiling or promoter microarray [77, 153]. Whereas tiling arrays cover the whole genome sequence, promoter microarrays cover the upstream regions of the TSS for every gene (1-10 kb) [142, 153]. One of the key issues is the identification of the best binding sites among all identified potential sites which requires computational methods [77]. In addition, the microarray design and probe density dictates the ability to identify binding regions [239].

**ChIP-Seq** Chromatin-immunoprecipitated DNA can also be sequenced using massively parallel sequencing, often referred to as next-generation sequencing [153]. The sequence reads are mapped to a reference genome [153] and a peak-finding algorithm is used to determine binding site locations [213]. Unfortunately, due to repetitive DNA sequences, not all sequence reads can be mapped unambiguously [153]. ChIP-Seq has a finer resolution than ChIP-chip in large genomes (25-200 bp compared to 200 bp) [153].

**SELEX** ChIP and subsequent motif discovery may miss binding sites due to partial occupancy or low resolution [153]. The identification of all binding sites under all possible biological conditions for each TF remains also elusive [153]. However, *in vitro* proteins bind to DNA probes regardless of the condition [153]. Systematic evolution of ligands by exponential enrichment (SELEX) [263] is a high-throughput approach to screen short, random oligonucleotide probes for recognition by a specific protein *in vitro* [77]. This way, short DNA sequences with a high affinity to the transcription factor of interest are selected from randomized short double-stranded DNAs from a genome-wide library [208].

### 1.3.2 Computational approaches

A number of bioinformatics approaches have been developed for the identification of both known and unknown transcriptional regulatory elements [179]. Upstream sequences can be scanned for motifs of known transcription factor binding sites extracted from databases such as TRANSFAC [183]. It is also possible to analyze sets of coexpressed genes and identify common sequence motifs in their upstream region [179] under the assumption that co-expressed genes are also co-regulated [169]. If the set of genes is coexpressed, the expression might be mediated by common regulatory elements [284]. Coexpressed genes can be identified by e.g. microarray expression experiments. These sequences can not only be used to identify binding sites for known transcription factors but also for the *de novo* identification of binding motifs which are overrepresented in the set of upstream sequences of the genes of interest.

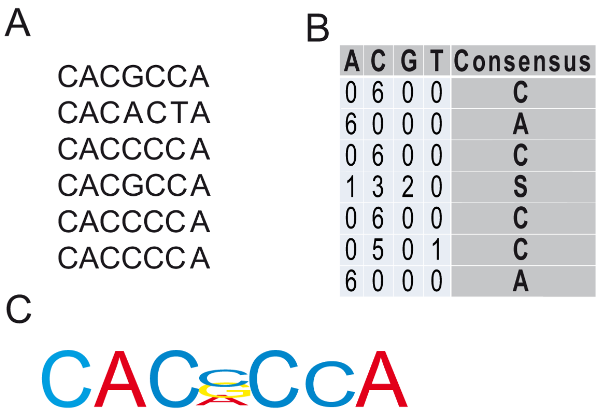
Description of binding motifs are built from experimentally determined transcription factor binding sites in DNA and can be used to form a consensus sequence or a position weight matrix (PWM), also called position specific score matrix (PSSM), [259] which can in turn be used to scan a putative regulatory region for motif occurrences [284]. Consensus sequences are simple strings over the 4-letter alphabet [A,C,G,T] that forms DNA sequences [107]. The degenerate IUPAC nucleic acid code [132] also describes variation in a specific position [107]. However, degenerate consensus sequences contain little information about the actual nucleotide frequencies at the different positions of the binding profile [45]. Thus, PWMs or PSSMs are used to describe the nucleotide preferences for a specific factor [45].

### PWMs

To construct a PWM, identified binding sites are aligned and the distribution of each base in each position of the binding motif is used as the weight in the PWM (see Figure 1.4). Hence, the elements of a PWM describe the likelihood of a nucleotide at a certain position [45]. The observed number of occurrences of each nucleotide at each position is represented in a *count matrix* [227]. This number is divided by the number of total sequences so all rows sum up to one [227]. A constant is often added to the matrix to avoid the occurrence of zero counts. This allows a minimal chance of a nucleotide at that position to occur rather than no chance at all [227]. For computational purposes the matrix is then transferred into a PWM using a logarithmic scale [290]. The PWM is slid over a sequence and in each sequence window a score is determined that captures the overlap between the sequence region and the PWM [111, 227]. There are a number of ways to use the determined score to decide if a match is an actual predicted binding site [111]. If the score exceeds a pre-determined threshold, the match is counted as a hit of the PWM at the identified sequence region [227]. A score threshold is often chosen in a way that the probability of finding a false hit by chance is at most 0.05 which limits false positive detections [227]. Rahmann and colleagues [227] developed a method that also accounts for the probability of detecting a true hit when there is a binding site present in the sequence (called the power).

A drawback of PWMs is the assumption that each position in the binding motif is independent towards the binding of the TF [27, 107], although this has been proven wrong in a number of situations [46, 176].

TRANSFAC [183] and JASPAR [287] are two major databases for eukaryotic PWMs [284]. Although these PWM libraries are incomplete, the search



CACGCCA
CACACTA
CACCCCA
CACGCCA
CACCCCA
CACCCCA

| A | C | G | T | Consensus |
|---|---|---|---|---|
| 0 | 6 | 0 | 0 | C |
| 6 | 0 | 0 | 0 | A |
| 0 | 6 | 0 | 0 | C |
| 1 | 3 | 2 | 0 | S |
| 0 | 6 | 0 | 0 | C |
| 0 | 5 | 0 | 1 | C |
| 6 | 0 | 0 | 0 | A |

**Figure 1.4. Representation of transcription factor binding sites.** **A** An example of six aligned sequences. **B** A position weight matrix for the given sequences and the consensus sequence derived from the matrix. S represents G or C. The score for each nucleotide at each position is obtained from the observed occurrence of that nucleotide at the corresponding position in the given sequences. **C** Sequence logo of the given sequences. The higher a letter the higher the probability of that nucleotide at the given position.

for occurrences of known motifs is less complex than the *de novo* identification of motifs [273]. A major drawback of scanning long sequences using PWMs is the inevitable large number of false positive hits due to the low information content of the binding sites [284].

Furthermore, predicted binding sites may not be functional binding sites and may not be bound *in vivo* due to various reasons such as the chromatin structure [45]. In addition to better motif descriptions, clustering of binding sites, evolutionary conservation of binding sites, and enrichment of regulatory sites in co-regulated genes can help to improve binding site predictions [283, 284].

**Phylogenetic footprinting** The term phylogenetic footprinting refers to the identification of conserved regulatory patterns in orthologous sequences using phylogenetic comparisons [45, 284]. The availability of a large number of genome sequences has enabled the comparison among different organisms. It is assumed that *cis*-regulatory elements are more conserved than non-coding sequences [284] and that orthologous genes are regulated by the same mechanisms in different species [290].

***cis*-regulatory modules** Transcriptional regulation occurs through the combined action of multiple transcription factors [174]. Thus, transcriptional regulatory sequences often comprise multiple binding sites for multiple regulatory factors in close proximity to each other [169]. Elements with a large number of binding sites are called *cis*-regulatory modules (CRM). Clustering of transcription factor binding sites is often used for the identification of CRMs [284]. These clusters can contain multiple binding site for the same TF (homotypic) or multiple binding sites for multiple TFs (heterotypic) [288]. CRMs can also be identified in conserved sequences when the sequences are aligned [284]. Searching for clusters of co-occurrences of binding sites has been shown to increase prediction specificity without loosing sensitivity [29].

Detection of CRMs can be performed following three different approaches [169]. To identify CRMs in a specific and well-studied process, genome-wide binding sites for a predetermined set of transcription factors are detected, often using a combination of PWMs, e.g. [110, 228]. A second approach identifies regulatory elements in a set of co-expressed genes, e.g. [272, 278]. The last approach yields at identifying genome-wide binding sites for any combination of transcription factors without any assumptions on a specific set of TFs or any specific process. This approach is more general and does not assume specific sets of transcription factors working cooperatively. However, only a few methods have been developed to identify sets of interacting transcription factors without prior knowledge on a genome-wide scale. One example is PReMod [33] where statistically significant clusters of up to five transcription factors were searched in whole-genome alignments. The method is based on the assumption that regulatory regions consist frequently of clusters of binding sites for a few different transcription factors and that these clusters are more conserved than their flanking intergenic sequences. Transcription factor binding sites were predicted in the human genome within a human-mouse-rat alignment block using vertebrate PWMs obtained from TRANSFAC [183]. To reduce false positive hits, the developed scoring method favors simultaneous matches in all three species. In a second step, clusters of putative transcription factor binding sites were identified. For this, regions of at most 2 kilobases (kb) were determined that were significantly enriched with binding sites for one up to five different transcription factors.

## 1.4 Regulation in development

Animal development is a fascinating process. Out of a single fertilized egg an embryo develops and embryonic cells differentiate into distinct cell types and organs building the adult body. All these processes are driven by an intrinsic blueprint of development written in the four-letter alphabet of the genomic DNA sequence [260] and depend on the precise control of gene expression at the level of transcription [206]. Multiple cells in an organism must develop in a coordinated fashion. For this, multiple genes must be activated at the same time in response to the same stimulus [291]. Interaction and communication of cells through signals and signal transduction pathways, which in turn induces a particular combination of TFs, is crucial for accurate pattern formation [306]. The precise patterns of gene expression are crucial for development which are controlled by transcription factors binding to *cis*-regulatory modules [310]. Thereby, TFs determine the rate of transcription and mediate the accurate activation or repression of a particular gene in a time- and tissue-specific manner, e.g. in the appropriate cell types or regions of the developing organism [12].

The importance of TFs in development and differentiation has been demonstrated in a number of cases [306] as they initiate specific developmental programs [306]. One example are the *hox* genes, which are involved in the correct formation of specific body segments and the anterior-posterior patterning of most metazoans [124, 148]. These genes are generally located within tightly regulated clusters and their complex expression patterns during development are regulated by local and long-range *cis*-regulatory DNA elements. The expression of Hoxb1 in the mouse hind brain is the best characterized example of a vertebrate Hox enhancer [82]. Other examples include Pax6, which controls eye development [18] and MyoD, which is crucial for muscle formation [210]. Recently, a specific combination of TFs has been shown to be sufficient to reprogram differentiated cells into pluripotent embryonic stem cells by a specific combination of TFs [135, 165, 269].

A specific spatio-temporal output of expression is achieved by integrating the input of multiple transcription factors in CRMs [14]. Numerous *cis*-regulatory elements direct the expression of a particular gene, each in a different pattern [199].

A detailed knowledge of the location of all developmental CRMs, a comprehensive map of their combinatorial and temporal binding profiles and the ability to predict their spatio-temporal activity is necessary to understand global *cis*-regulatory networks [310]. A central role in the regulation of developmental transcription has been attributed to DNA enhancer elements and transcription factors binding to these regions [206]. Although the role of distant acting regulatory sequences directing spatial and temporal expression patterns has been established in development, the identification of these sequences is still limited [261]. To predict spatio-temporal activity of enhancers, a number of sequence-based models has been applied [249, 311]. However, these methods are only accurate when tailored to individual CRMs [311] or small numbers of regulatory modules [249]. Besides this, genome-wide ChIP studies revealed extensive patterns of TF occupancy in a number of developmental contexts in different organisms, including *Drosophila* [310], mouse [154], and fish [195]. However, a static map of TF binding does not reflect the dynamic nature of gene expression and dynamic properties of *cis*-regulatory networks are crucial to understand temporal expression patterns [295]. A number of computational methods attempted to infer temporal regulation of gene expression mainly by testing enrichment of TF motifs in differentially expressed groups of genes [67, 92, 296]. Zinzen and coworkers [310] constructed a high-resolution atlas of *cis*-regulatory modules performing ChIP-chip experiments for five transcription factors at consecutive time-points describing their temporal and combinatorial occupancy during *Drosophila* mesoderm development [310]. As many biological processes are spatially and temporally controlled at the level of transcription, understanding the mechanisms of transcriptional regulation of gene expression will help to understand the molecular mechanisms of differentiation and development.

Alterations in *cis*-regulatory sequences responsible for proper transcription are essential for morphological diversification and evolution of developmental mechanisms [270]. As many developmental enhancers have a more flexible arrangement of binding sites than enhanceosomes, they are described more accurately by the billboard enhancer model. The exact composition of billboard enhancers are subject to rapid change in evolution, keeping the overall output constant [172]. The position of individual binding sites within CRMs involved in embryonic patterning is highly flexible [202]. Enhancers directing similar expression patterns can have different binding site arrangements [203]. Whereas regulation of terminal differentiation seems to be very simple, regulatory sequences for early patterning can be quite long and contain various different binding sites [202]. This can be explained by the fact that in early patterning events, multiple binding sites are necessary for sensing small differences in the concentrations and combinations of regulatory factors. However, only after cells express the relevant TFs, terminal differentiation can occur [202].

Although many genes involved in specific developmental processes have been identified, genes encoding transcription factors and cell signaling components need to be further characterized to elucidated extensive gene regulatory networks [163]. Regulation of development has been extensively analyzed in the sea urchin embryo [68, 209], frog [144], worm [131], and *Drosophila* [310]. However, little work has been done in mammals with few exceptions [253]. The comparison of sequences between different organisms has led to the identification of many thousands of conserved non-coding elements (CNEs) also between distantly related species (such as human and puffer fish) [166]. Experimental testing of randomly selected CNEs in mice [216, 286] and zebra fish [251, 298] has identified CNEs as potential enhancers and many CNEs show spatial- or temporal-specific enhancer activity [166]. One example is the identification of vertebrate brain region-specific enhancers through a high throughput analysis of expression-pattern associated CNEs in zebra fish [166]. Highly conserved non-coding sequences have also been shown to be associated with vertebrate development when comparing orthologous sequences between human and puffer fish [298]. Most of the identified sequences are located upstream of genes involved in developmental regulation [298] and a number of these sequences have been shown to have some function *in vivo* (e.g. [201]). In addition, ultraconserved elements (perfectly conserved regions of at least 200 bp) between human, mouse, and rat [25] as well as between human, mouse, and puffer fish [242] have also been shown to be located nearby genes encoding key regulators of development and transcription [25, 34, 298]. Furthermore, these ultraconserved elements have been shown to serve as long-range enhancers during mouse development [216]. Therefore, extreme evolutionary non-coding conservation can serve as a powerful predictor for mammalian tissue-specific enhancers [216].

Most developmental enhancers (irrespective of the size of the genome) are typically between 200 bp to 1 kb in length [162] and contain multiple binding sites for different classes of sequence-specific TFs [12]. Additionally, these enhancers often contain binding sites for repressors that inhibit expression in inappropriate tissues [188]. The combination of computational and experimental approaches has greatly improved the collection of developmental enhancers and has allowed investigating the exact arrangements of binding sites of developmental enhancers [162].

Although the textbook view of developmental transcription is that regulation is mediated by TFs and enhancer sequences, recent reports have also implicated TFII complexes and core promoter elements in the precise regulation of developmental transcription [206]. So far no universal eukaryotic promoter elements have been identified, e.g. the TATA box occurs only in ~10-20% of eukaryotic genes [23, 205]. Furthermore, core promoter elements and TFIID complexes can be highly diverse. Hence, it was suggested that core promoter elements may be adapted to the transcription initiation machinery of specific cells [206]. In addition, enhancer elements were shown to interact with different core promoters during distinct stages of development [206].

## 1.5 Regulation of cholesterol biosynthesis

Cholesterol is central in human metabolism as cholesterol and its derivates, steroids and bile acids, act as signal transducers and solubilizers of other lipids [128]. In addition, cholesterol is a substantial component of cellular membranes modulating the function of membrane proteins and it participates in several membrane trafficking and transmembrane signaling processes [128]. Dysfunction of cholesterol metabolism has been implicated in various diseases such as cardiac diseases, dementia, diabetes, and cancer [127, 185].

Mammalian cells acquire cholesterol by *de novo* synthesis (see Figure 1.5) in the endoplasmic reticulum (ER) and by endocytosis of lipoproteins [184]. Cholesterol esters in the core of lipoproteins (of which ~70% are low-density lipoproteins [22]) are hydrolyzed in late endosomes and lysosomes, and free cholesterol is released into the cell [184]. Feedback control mechanism of cholesterol is mediated by a cell surface receptor for a plasma cholesterol transport protein called low density lipoprotein receptor (LDLR) [100]. These receptors bind LDL and carry it into the cell by receptor-mediated endocytosis [43]. The internalized lipoprotein is delivered to lysosomes where its cholesterol esters are hydrolyzed [43].

The sterol regulatory element binding proteins (SREBPs) are key regulators of cholesterol and fatty acid metabolism [120] by regulating multiple genes involved in cholesterol biosynthesis and uptake [44]. SREBPs belong to a large class of TFs containing basic-helix-loop-helix-leucine zipper (bHLH-Zip) domains [240] and are synthesized as inactive precursors bound to the membranes of the ER [44, 101]. A two-step proteolytic process of cleavage is required in order to release their amino-terminal bHLH-Zip containing domain into the nucleus. SREBPs bind like all bHLH proteins to E-boxes (5'-CANNTG-3', with N representing any base) and a specific DNA sequence, the sterol regulatory element (SRE) (5'-TCACNCCAC-3') [140]. SCAP (SREBP cleavage activating protein) is a required activator of SREBP cleavage and the activity is abolished by sterols [44]. SCAP contains a sterol-sensing domain regulating the transport of SREBP from the ER to the Golgi [252]. In the Golgi the cleavage is initiated by a membrane-bound serine protease termed Site-1 protease (S1P) that clips SREBP at site 1. This cleavage breaks the covalent bond between the two transmembrane domains of SREBP but both halves remain attached to the membrane. A second membrane-bound zinc metalloproteinase termed Site-2 protease (S2P) clips the first transmembrane fragment at site 2, which releases the active domain into the cytosol from where it enters the nucleus [44].

Three isoforms have been identified in mammals, SREBP1a, SREBP1c, and SREBP2 [101] which are encoded by two different genes (SREBF-1

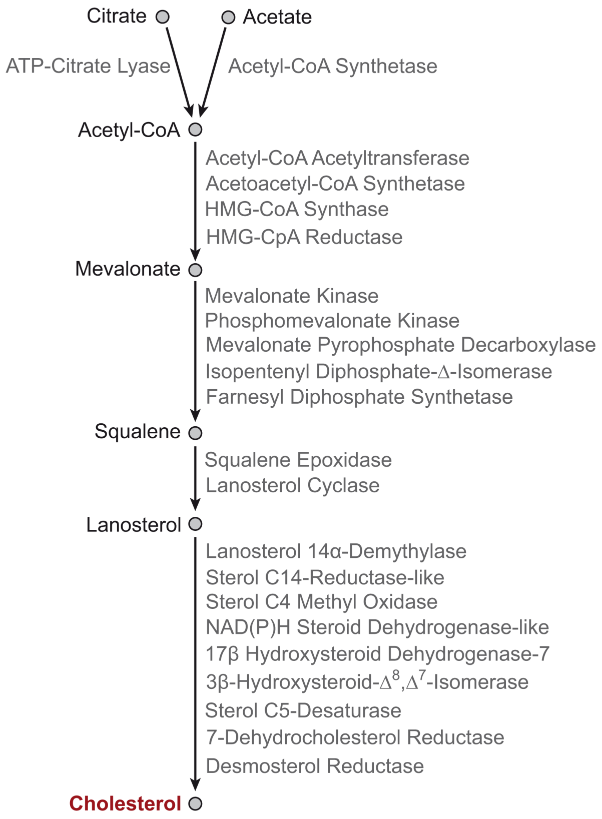

Citrate
Acetate
ATP-Citrate Lyase
Acetyl-CoA Synthetase
Acetyl-CoA
Acetyl-CoA Acetyltransferase
Acetoacetyl-CoA Synthetase
HMG-CoA Synthase
HMG-CpA Reductase
Mevalonate
Mevalonate Kinase
Phosphomevalonate Kinase
Mevalonate Pyrophosphate Decarboxylase
Isopentenyl Diphosphate-Δ-Isomerase
Farnesyl Diphosphate Synthetase
Squalene
Squalene Epoxidase
Lanosterol Cyclase
Lanosterol
Lanosterol 14α-Demythylase
Sterol C14-Reductase-like
Sterol C4 Methyl Oxidase
NAD(P)H Steroid Dehydrogenase-like
17β Hydroxysteroid Dehydrogenase-7
3β-Hydroxysteroid-Δ8,Δ7-Isomerase
Sterol C5-Desaturase
7-Dehydrocholesterol Reductase
Desmosterol Reductase
Cholesterol

**Figure 1.5. Human cholesterol biosynthesis.** Key intermediate metabolites of the pathway are represented as circles. Arrows indicate the cholesterol pathway. CoA: coenzyme A.

and SREBF-2) [44]. SREBP1a is involved in the regulation of cholesterol and fatty acid biosynthesis, whereas SREBP1c and SREBP2 regulate fatty acid metabolism and cholesterol biosynthesis, respectively [120]. SREBPs are weak transcriptional regulators themselves and act in combination with other transcription factors, i.e. Sp1 transcription factor (SP1) and nuclear transcription factor-Y (NF-Y) [72, 134, 234, 241]. Although many SREBP-responsive genes involved in the uptake and synthesis of cholesterol and fatty acids have been described [22, 121, 231], it is suggested that only a smaller fraction of SREBP target genes have been discovered so far [121]. Furthermore, little is known about interactions and simultaneous binding of SREBP with other transcription factors [231]. Diseases caused by dysfunctional cholesterol uptake and synthesis like Familial Hypercholesterolemia and Niemann-Pick Disease Type C are not fully understood and open questions remain concerning the regulation of cholesterol metabolism and molecular interactions of known cholesterol-regulating factors [22].

## 1.6 Machine learning

Machine learning emerged from the subfield of computer science known as *artificial intelligence (AI)*. Machine learning systems “learn” from previous experience. Accumulated experience (e.g. through experimental data) allows a machine to develop new knowledge which leads to a better performance on a specific task over time. Learning from experience is the central idea to the different types of problems encountered in machine learning, especially problems involved in classification. The general goal of all problems is the identification of a systematic way of classifying a new example. Classification is based on knowledge obtained from *learning* (or *training*) *samples* together with measurements obtained from a similar new example. The number of classes needs to be finite and known and the class of each example needs to be determined and known. [133]

Machine learning can be divided into *supervised learning* and *unsupervised learning*.

**Unsupervised learning** The goal of unsupervised learning is the exploration of characteristics of the input variable when there is no appropriate information about an output variable available. This approach was not followed in this thesis.

**Supervised learning** The goal of supervised learning is to find a function of the input variables to approximate the known output variables. For this, the learning algorithm receives a set of continuous or categorical input variables and a correct output variable. This way, the relationships between the input variables and the output variable can be analyzed. If the output variable is continuous a regression problem is faced, whereas a categorical output variable faces a classification problem.

The concept of *generalization* aims at making good predictions when applied to a data set that is independent of the data used to fit the model. Not using an independent data set will result in overestimating the model’s predictive accuracy. One way of creating an independent data set is to hold back a proportion of the data set from the model fitting and use it for prediction. Usually, the data set is separated into three non-overlapping and independent data sets. [133]

**Training set** The training set is used for the assessment of the data, preliminary testing, looking for patterns, trying different models, and eliminating outliers.

**Validation set** The validation set is used to assess the different models and to select the best model possible.

**Test set** The test set is used to assess the performance of the specified final model.

To assess the performance of a particular model, the *prediction error* can be used as a measure of prediction accuracy. In classification, a classifier is built from the training set and used to predict the classes of the test set. The proportion of all misclassified samples in the test set is then defined as the prediction error. Other methods to assess the test error are based on *cross-validation* [264] and *bootstraping* [74] and are used when limited data samples are available.

***V*-fold cross-validation** During cross-validation, the entire data set is divided randomly into equally sized $V$ non-overlapping groups with $V$ being any number between 2 and the overall sample size. One group is then successfully removed from the entire set and the other $V$-1 groups are used as the training set to fit the model. The omitted group serves as the test set and its output variable is predicted using the fitted model to determine the prediction error of the omitted group. This procedure is repeated $V$ times, each time removing a different group. The overall test error is estimated by averaging over the obtained $V$ prediction errors. This way, the entire data set is used in a more efficient manner than the mere division into a training and an independent test set.

**Bootstrapping** A *bootstrap sample* is randomly drawn from the entire set with replacement. Using this sample, a model is fitted and the prediction error assessed with the remaining data. This procedure is repeated at least ~1000 times, each time assessing the prediction error. The test error is then estimated by averaging all the prediction errors.

When learning a model one has to be careful not to overfit the model. Overfitting occurs when the model is too large or complicated, or when the data set contains too many parameters relative to the size of the training set. It usually estimates well the training set but results in a large prediction error on the test set. [133]

### 1.6.1 Decision trees

Decision tree learning is one of the most extensively used methods for inductive inference. Tree-based methods have been used in a variety of fields such as biomedical and genetic research, marketing, political science, speech recognition, and other applied sciences. Decision trees are intuitive and easy to interpret as the sequences of decisions made to assign a class label to an input is easy to follow [145]. Furthermore, they can be easily extended to categorical rather than numerical variables. Furthermore, decision tree learning methods are robust to errors and can cope with missing values [193]. However, a challenge in decision trees remains the induction of decision tress of small size and depth [145].

Examples are classified by sorting them down the decision tree from the root to some leaf node. At each node a test of some variable is specified and the example is moved down the branch from that node corresponding to the value of the variable in the given example. An example for a decision tree is given in Figure 1.6. Decision trees are best suited to problems where instances are represented by a fixed set of variables and their values with a discrete output variable [193].

A simple and powerful method to infer classification rules from a set of labeled examples is the top-down induction of decision trees [223]. One of the earlier approaches are Friedman and Breiman's work resulting in the CART system [40, 88] and the ID3 algorithm [223] with its successor C4.5 [225]. The central idea of the nonparametric statistical method of *classification and regression trees (CART)* [40] is an algorithm known as *recursive partitioning*. This involves a step-by-step procedure to construct a decision tree by either splitting or non splitting each node in the tree into two daughter nodes. The CART algorithm (or the related C4.5 methodology) asks a sequence of boolean questions and the results are therefore relatively easy to understand and to interpret. In the CART methodology, the input space is partitioned into a number of non-overlapping rectangular or cuboid regions. Each region is viewed homogeneous to predict the output. Classes are assigned to the regions with sides parallel to the input space. This kind of partition corresponds to a classification tree. Similar to classification trees, regression trees are constructed by recursive-partitioning, generally referred to as recursive-partitioning regression. [133]

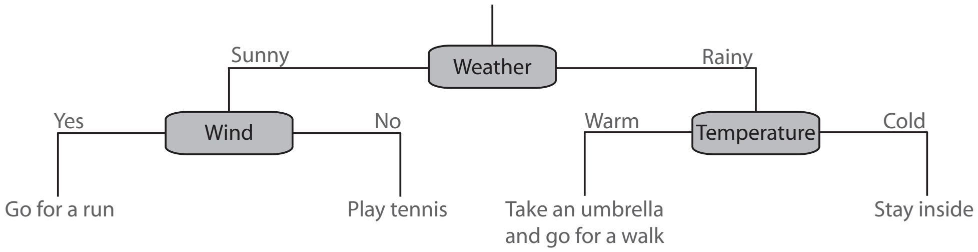


**Figure 1.6. A decision tree for the concept of *Saturday afternoon activity*.** An example is classified by passing it down the tree to the appropriate leaf note. This tree classifies Saturday afternoon according to what activity is suitable for the given weather, e.g. if it is rainy and cold outside the appropriate activity is staying inside.

### Classification Trees

A classification tree is the outcome of asking an ordered sequence of questions that terminates with the prediction of a class. Each question depends on the answer to the previous question in the process.

At each node in the tree a decision rule is implemented splitting the examples into two or more partitions. The starting point of a classification tree at the top of the tree is called the *root node*. It consists of the entire training set that is successively passed down the tree. A terminal or nonterminal *node* consists of a subset of the set of variables. A nonterminal node is also called a *parent node* and a binary split divides the node into two daughter nodes. The binary split is determined by a Boolean condition on the value of a single variable. The observed value of a variable can either satisfy ("yes") or not satisfy ("no") the condition at that split. Observations satisfying the condition for the variable at a particular node are passed down to one of the daughter nodes, all other observations not satisfying the condition drop down to the other daughter node. A *terminal node* or *leaf node* does not split and is assigned a class label depending on the class of the majority of the observations at that node (called the plurality rule). Each observation is passed down the tree until it falls into one of the terminal nodes where it is assigned a class label. The set of all terminal nodes of a tree is called a *partition* of the data. If a tree has only a single split with two terminal nodes, it is called a *stump*.

To grow a classification tree, one has to determine how to choose the Boolean conditions for splitting at each node, which criterion should be used to split a parent node into its daughter nodes, how to determine a node to be terminal, and how to assign a class to a terminal node. [133]

### Tree growing procedure

The tree growing algorithm needs to decide at each node which variable is split “best”. All variables present at the node need to be considered for the split and each one needs to be evaluated to determine the best variable for the split. However, for each given variable the best split needs to be determined first. To determine the best possible split, a measure for the **goodness of a split** is needed which is defined by the node impurity function. The two most commonly used impurity functions are the *entropy function* [223] and the *Gini diversity index* [40]. The entropy function is defined as

$$i(\lambda) = -\sum_{k=1}^{K} p(k|\lambda) log(p(k|\lambda)) \tag{1.4}$$

where $p(k|\lambda)$ is the probability that an observation is correctly classified as class $k$ at node $\lambda$. The Gini diverstiy index is defined as

$$i(\lambda) = 1 - \sum_{k} p(k|\lambda)^2 \tag{1.5}$$

where $p(k|\lambda)$ is also the probability that an observation is correctly classified as class $k$ at node $\lambda$. The Gini index is often the default function in tree growing software as it is frequently the best splitting rule. It focuses on separating one class at a time from the remaining data and thereby tries to produce pure nodes, whereas entropy aims at equalizing sample sizes in the generated subsets at each split [38]. The reduction in impurity gained by splitting the parent node into its daughter nodes gives the goodness-of-split at a particular node.

The sequential splitting process of growing a tree is called **recursive partitioning**. Each tree growing procedure starts with the root node, which consists of the entire training set. The tree algorithm determines the best split at the root node for each variable using the chosen “goodness-of-split” criterion. The split with the largest value over the best splits of all single variables is chosen as the best split at the root node. According to the determined split, the root node is split into two daughter nodes. Each daughter node is then split in the same way as the root node and the subsequent nodes are split accordingly. This results in a greedy search for an acceptable decision tree where earlier choices are never reconsidered [193]. In a binary tree every node has exactly two daughter nodes. A tree is saturated if the procedure is conducted until none of the nodes can be split any further. A terminal node can also be determined based on a defined stopping criteria. To restrict the growth of a tree, a node can be declared terminal if it is smaller than a predefined threshold. The growing procedure can also be restricted by setting a minimal value for the improvement of the goodness-of-split value at a node. However, it is usually better to grow a saturated tree and then "prune" it back [40]. The idea of pruning is to let the tree grow "large" and then successively remove branches with little statistical validity [91, 191, 224] until the tree has obtained the "right size" using a bottom up approach [40]. Therefore, a pruned tree is a subtree of the original saturated tree.

A good estimate of the misclassification rate can be computed by using an independent test set or cross-validation to determine the best subtree. [133]

## 1.6.2 Ensemble learning

How to lower the generalization error of a learning algorithm by reducing the bias or the variance is one of the most important research topics in machine learning. This is related to the idea of "instability" of a prediction or classification method. A classification is unstable if small perturbations of the training set lead to major changes in the resulting classifier. Due to overfitting, unstable classifiers have high variance and low bias. In contrast, underfitting leads to a high bias. By this definition, decision trees are unstable. However, instability of a classifier can be used to improve the accuracy of the learning algorithm. By perturbing the training set, an ensemble of different base classifiers is generated. Using these combined classifiers is called *ensemble learning* or *committee-based learning* and their success often depends on the degree of instability of the base classifier. Ensemble learning algorithms generate many classifiers and aggregate their results [167]. Growing an ensemble of trees and determining the most popular class has significantly improved classification accuracy [39]. Two methods for ensemble learning are *bagging* [37] and *boosting* [86] which differ in the way perturbations are generated. Whereas bagging was designed to reduce variance, boosting appears to rather reduce bias. Another example for ensemble learning is *random forest* [39].

### Bagging

Bagging is an acronym for "bootstrap aggregating" [37]. Perturbations of the training set are generated by random and independent drawings from the training set. It was the first procedure successfully combining an ensemble of learning algorithms that improved the performance over a single learning algorithm. Bagging starts by drawing $B$ bootstrap samples from the training set where each bootstrap sample is gained by repeating sampling with replacement from the training set. For each bootstrap sample a classification tree is grown independently of earlier trees and each sample is dropped down each of the bootstrap trees. The class of each sample is determined by class which was predicted by the majority of trees. This classification procedure is called the majority-vote rule. Those observations not included in the bootstrap sample are called out-of-bag (OOB) observations. These OOB observations serve as an independent test set. The OOB misclassification rate is then determined by the proportion of classes where the predicted class differs from the actual class for all observations in the training set. [133]

### Boosting

In contrast to bagging, boosting [86] is an iterative process where each classifier is dependent on the performance of those built before [297]. Thereby, the performance on previously misclassified samples is improved. Whereas in bagging all single classifiers are equally weighted, boosting weights classifiers according to their contribution in performance. In terms of decision trees, successive trees give extra weights to incorrect predictions by earlier trees [167]. Boosting aims at enhancing the accuracy of samples that are difficult to predict using a "weak" binary classification learning algorithm. A "weak" classifier is only marginally better than random guessing and classifies correctly barely more than 50% of the time. The term "boosting" derives from the idea of creating a "strong" classifier by improving ("boosting") the performance of a single classifier. This improvement is achieved by combining classification votes from an ensemble of similar "weak" classifiers. [133]

### Random forest

Random forest is one of the most effective ensemble methods available and is an extension of the idea of bagging [39]. Whereas in bagging randomization is only used in choosing the data set to grow the tree on, in random forest randomization is also a crucial part of constructing each tree, thereby adding another layer of randomization to bagging. Random forest follows the bagging procedure by drawing $n$ bootstrap samples from the original data set. However, these approaches differ in the way the trees are grown from the bootstrap samples. An unpruned classification tree is grown for each bootstrap sample. Whereas in single decision trees each node is split according to the best split among all variables, in random forest the best split at each node is determined among a randomly chosen subset of all variables. Bagging constitutes a special case of random forest when all variables are used at each node instead of a subset. New data is then predicted by a majority vote of all trees in the forest. The error is estimated using the OOB observations of each bootstrap sample. Using the additional layer of randomization, the correlation between the different tree-structured classifiers is reduced.

Random forest is easy to use as there are only two tuning parameters, the number of bootstrap samples and the number of variables randomly chosen as a subset at each node. In addition, random forest cannot overfit as the generalization error converges to a limit when the number of trees in the forest is increased. Furthermore, random forest can be used to evaluate the variables in a data set and assess the *importance* of each variable. This is achieved by classifying the OOB observations and compute the OOB error rate. The OOB values for the specific variable are then permuted while all other variables remain unchanged and the altered OOB observations are re-classified. If a variable is important, the altered data leads to a poorer classification. The difference between the two computed error rates averaged over all trees in the forest serves as a measure for the importance of that variable. To identify structure in the data or for unsupervised learning, a *proximity measure* can be estimated. Proximity between two instances is given as the number of trees where both instances fall in the same terminal node under the assumption that "similar" observations fall in the same terminal node more often than dissimilar ones. [133]

## 1.7 Network analysis

To understand cellular processes, it is necessary to analyze cellular molecules in the framework of pathways and networks. Experimental studies and large-scale screens have provided interaction data that can be assembled into a network format. The network can then be analyzed for significant biological properties through its topological structure [308]. The relationships between the different biological entities, e.g. interactions between proteins, can be described with the language of graph theory which offers a mathematical abstraction [122]. A graph consists of a set of nodes and a set of edges. The edges connect the nodes, representing the relationship between them. For example, nodes could represent different proteins and the edges physical interactions between the proteins. Examples of biological graphs include regulatory networks, signal transduction networks, protein-protein interaction networks, and metabolic networks. Depending on the characteristics of the biological data, networks can be directed or undirected [308]. In case of transcriptional regulatory networks, edges within a graph can be directed to represent the direction of regulation [122].

**Metabolic networks** One well-established type of biological networks are metabolic networks. It can be easily constructed when key biochemical relationships between key metabolic genes are known. It represents biochemical reactions between the different substrates facilitated by metabolic enzymes and can represent certain pathways such as cholesterol biosynthesis (see section 1.5).

**Signaling networks** Extracellular signals are connected to the control of transcription factors via signal transduction pathways. These networks describe the interactions between signaling molecules within the cell from the extracellular input to the specific transcription factors involved in the distinct process.

**Regulatory networks** Large-scale identification of transcription factor binding sites by ChIP-chip or ChIP-Seq has allowed the construction of transcriptional regulatory networks, e.g. [292]. These networks represent the regulation between transcription factors and their respective genes.

**Protein-protein interaction networks** Interactions between proteins can be assembled into a protein-protein interaction (PPI) network. These networks represent the largest and most diverse biological data sets available to date [308]. Tightly connected proteins are often involved in similar processes. Functional annotations of interacting proteins may indicate potential roles for unannotated genes and improve our understanding of true pathological mechanism of a disease [308]. PPI network models are often used as a simplification of more elaborated signaling networks [246].

### 1.7.1 Graph theory

A graph $G = (V, E)$ is specified by its set of nodes $V$ and its set of edges $E$. Each element of $E$ consists of a pair $u, v$ of elements of $V$ and edges can be assigned weights, directions, and types. Two nodes are *adjacent* to each other if they are connected by an edge. Likewise, two edges are *adjacent* if they are joined by a node. If all nodes in a graph are connected with an edge, the graph is called a *complete graph.* [122]

A graph can be represented by its *adjacency matrix* which is a square matrix $A$ whose rows and columns correspond to nodes. Its elements $A_{ij}$ denote the presence of an edge from node $i$ to node $j$ and possibly the weight of the edge. The adjacency matrix is symmetric for undirected graphs. [122]

The first step in understanding network architecture and performance is the analysis of its network topology. The most important and commonly used topological features in cell biology include degree, clustering coefficient, shortest path length, and betweenness centrality [308].

The simplest measure to characterize the role of a node in a network is the node degree or connectivity [6]. The **degree** or **connectivity** of a node $v$ is equal to the number of edges incident at node $v$. A node with many connections has a higher node degree, which reflects its importance in the network [308]. The degree for directed networks can be divided into in-degree and out-degree that represent the number of incoming and outgoing edges, respectively. One example are transcriptional regulatory networks, where the in-degree can represent the number of transcription factors regulating a target genes whereas the out-degree reflects the number of target genes regulated by transcription factors [142]. Most target genes have only a small number of transcriptional regulators, whereas just a small set of transcription factors has a high number of connections [106].

The **clustering coefficient** of a node measures the degree to which the neighborhood of a node resembles a completely connected subgraph (clique). It is defined as the ratio of the number of actual edges between the node's neighbors to all possible edges between them. The clustering coefficient is given by

$$C(v) = \frac{2E}{k_v(k_v - 1)} \tag{1.6}$$

where $E$ is the number of edges connecting the immediate neighbors of node $v$ and $k_v$ is the degree of node $v$. The clustering coefficient measures the cliquishness, or transitivity, of the local neighborhood [6] and quantifies the probability of two nodes that are both neighbors of the same third node also being connected to each other [98].

The **shortest path length** between two nodes represent the shortest distance between the two nodes. The maximal length of the shortest path in a network is called the *graph diameter* [308].

The **betweenness centrality** is defined as the fraction of shortest paths between all pairs of nodes passing through a node or edge. The betweenness centrality is given by

$$C(v) = \sum_{s \neq v \neq t \in V} \frac{s_{st}(v)}{s_{st}} \tag{1.7}$$

where $s_{st}(v)$ denotes the number of shortest paths between node $s$ to node $t$ and node $v$ lying on a shortest path between $s$ and $t$ and $s_{st}$ denotes the number of all shortest paths from $s$ to $t$. Betweenness centrality is an estimator of the traffic load through a node and the rate at which signals pass along the edges [85, 200, 308].

There exist three distinct models of networks according to their topology, random networks, scale-free networks, and hierarchical networks [20]. These network models are distinguished by their distribution of node degree [142]. The **connectivity distribution** given by

$$P(k) \sim \frac{N_k}{N} \tag{1.8}$$

where $N_k$ is the number of nodes with $k$ neighbors. It offers a more detailed insight into the structure of a graph. Erdös and Rény [78] showed that the connectivity of simple random graphs follows a Poisson distribution. However, in many real networks the degree distribution $P(k)$ follows a power-law distribution

$$P(k) \sim k^{-\gamma} \tag{1.9}$$

where $\gamma$ is a constant depending on the network usually in the range of $2 < \gamma < 3$ [1, 19]. In these network, the majority of nodes only have a few connections whereas only a small number of nodes are highly connected [6]. These networks are called scale-free. Nodes with many connections are also called *hubs* which often play a crucial role in cellular networks [20, 136, 220]. Another common property of many networks is the community structure, the division of nodes into highly connected groups in the network with only sparse connections between them [200].

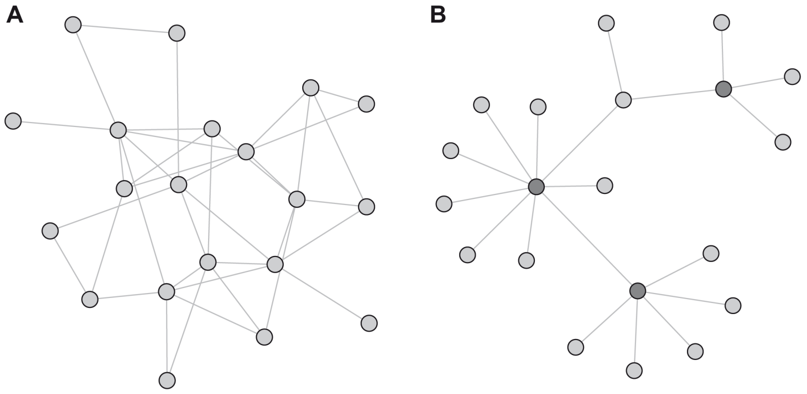


**Figure 1.7. Random and scale-free networks. A** A random network with a Poisson connectivity distribution. **B** A scale-free network with a power-law connectivity distribution where most nodes are scarcely connected and only few nodes, called hubs, have many connections (shown in dark gray).

The majority of cellular (e.g. metabolic networks, protein-protein-interaction networks) as well as non-cellular networks (e.g. social world networks, the World Wide Web) seem to approximate a scale free topology [20]. In scale-free networks, the nodes usually are highly connected such that each node can be reached from every other node in a minimal number of steps (called the *small world property*) [64]. These networks have therefore a small graph diameter [189] and a high clustering coefficient [308]. The scale-free topology provides robustness to the network with increased flexibility to random perturbations where the loss of individual nodes usually has no effect on the overall network topology. Nevertheless, it is susceptible to targeted attacks at heavily connected critical hubs [2, 6] and mutations affecting hubs are more likely to cause a defect [308].

Although the analysis of cellular networks has given details into the biological system of interest, they usually represent a static representation of the biological system. However, the cell is far from a static environment and new approaches are needed to incorporate the dynamic nature of biological systems [6].

# Chapter 2

# Methods

## 2.1 Identification of spatio-temporal specific regulatory modules

The complete workflow of the analysis is shown in Figure 2.1.

### 2.1.1 Identification of transcription factor binding sites

Sequences from 10,000 base pairs (bp) upstream to 100 bp downstream of the transcription start site (TSS) for 32,290 human genes (Build 36.3) as well as for 33,063 genes for mouse (Build 37.1) and 27,110 genes for rat (Build 4.1) were retrieved from the National Center for Biotechnology Information (NCBI, `ftp://ftp.ncbi.nlm.nih.gov/`). Transcription factor (TF) annotations and associated position weight matrices (PWMs) were obtained from TRANSFAC (Release 12.1) [183] yielding 549 human transcription factors (455 PWMs), 407 transcription factors (TFs) for mouse (410 PWMs), and 366 TFs for rat (471 PWMs). Each PWM binding site was mapped to the corresponding binding site of its associated TFs. TFs sharing the same PWM were grouped together. The grouping resulted in 152 TF-groups for human (see Table A.1), 139 for mouse, and 141 TF-groups for rat which were used for further analysis. The detection of TF binding sites based on the respective PWMs was performed with the software package R (`www.r-project.org`) as described previously [227, 294]. Predicted binding sites with a p-value above P = 0.05 were discarded.

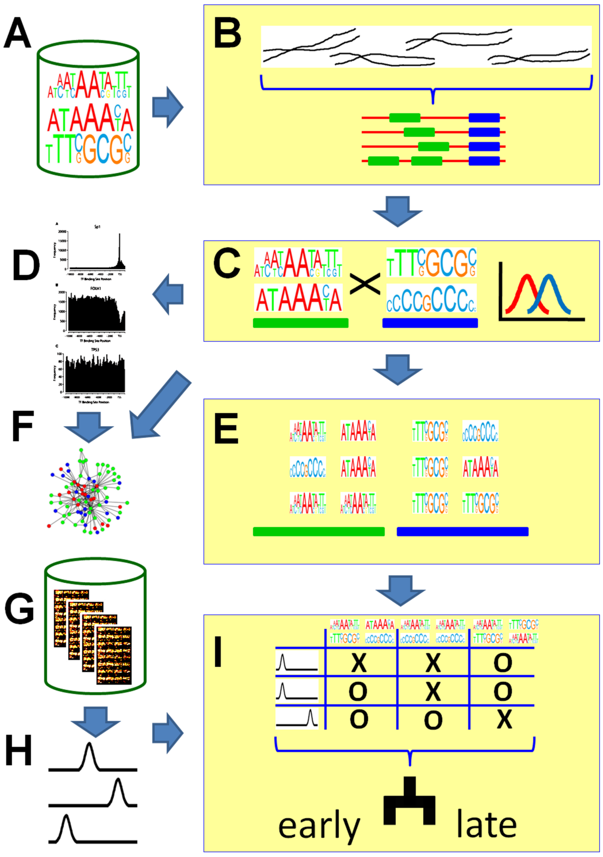


**Figure 2.1. Workflow for the identification of tissue- and temporal-specific regulatory modules in development and differentiation.** **A** Motifs (position weight matrices) for transcription factor binding sites were collected from a database. **B** Upstream sequences were gathered for each transcript. Promoters (+/- 100 base pairs of the transcription start site, TSS) and enhancers (defined by accumulation of binding motifs and phylogenetic conservation, 2,000-10,000 base pairs upstream of the TSS) were selected. **C** Statistical and combinatorial analysis of transcription factor binding sites of promoters and enhancers. **D** Characterization of single motifs with respect to their distributions in the observed sequences (0-10,000 bases upstream of the TSS). **E** Assembly of regulatory modules. A regulatory module consisted of a pair of transcription factors binding at the promoter region and a pair of transcription factors binding at the enhancer region. **F** Network analysis. **G** Gene expression data was taken from microarray studies of the development of several mouse tissues and of the differentiation of human stem cells. **H** A time series analysis was performed to identify genes being differentially expressed at distinct (developmental) time intervals and tissues/cell types. **I** The regulatory modules were used to predict differential expression of developmental time intervals.

### 2.1.2 Identification of combinations of transcription factors

Predicted TF binding sites were combined into pairs of TFs. Regions 100 base pairs (bp) upstream and downstream of the transcription start site (TSS) were used as promoters, and regions starting 2,000 bp and ending 10,000 bp upstream of the TSS were used as potential enhancer regions. Combinations of TFs for promoters were obtained by pairing non-overlapping TF binding sites co-occurring in the promoter region of a gene using a sliding window of 20 bp. Only pairs occurring in at least 10 genes were taken into further consideration. To decrease false positives of predicted transcription factor binding sites, only conserved binding sites were analyzed in enhancer regions. To determine the conservation of human, mouse and rat binding sites, we analyzed pair-wise alignments between human and chimp, mouse and rat, and rat and mouse, respectively. Chained and netted pair-wise alignments of human (UCSC version hg18) and chimp (UCSC version panTro2), of mouse (UCSC version mm9) and rat (UCSC version rn4), and of rat (UCSC version rn4) and mouse (UCSC version mm9) were downloaded from UCSC [235] in the axtNet format (`ftp://hgdownload.cse.ucsc.edu/`). Conserved regions between human and chimp, mouse and rat, and rat and mouse were determined by the given aligned regions in the alignment files. Predicted binding sites were compared to the identified conserved regions and taken if binding sites occurred in these conserved regions. Pairs of non-overlapping co-occurring transcription factors in enhancer regions were determined using a sliding window of 20 bp (same size as for promoter regions). To analyze enhancer regions with a comparable size to promoter regions, we regarded sequences of a 200 bp sliding window. As enhancer were shown to consist of clusters of TFs [13, 158], only regions in which at least 10 binding sites occurred were considered as enhancer regions and TF pairs occurring in at least 10 genes were considered further.

### 2.1.3 Identification of regulatory modules

Regulatory modules were constructed by combining two TF pairs occurring at the respective promoter and enhancer regions of a gene. Regulatory modules occurring in at least 10 genes were taken for further analysis. Hence, regulatory modules consisted of a combination of a pair of co-occurring transcription factors at the promoter and the enhancer region. To show that combinations of TF pairs of enhancers and promoters show better specificity than combining only TF pairs binding at the promoter, we also constructed the latter combinations. For this, two pairs of co-occurring transcription factors in the promoter region were combined for each gene and taken for further analyses if the regulatory module occurred in at least 10 genes.

### 2.1.4 Gene expression analyses of mouse embryonic development and human stem cell differentiation

Gene expression data of mouse embryonic development and differentiation of human stem cells were retrieved from the Gene Expression Omnibus (`http://www.ncbi.nlm.nih.gov/geo/`). For mouse embryonic development, gene expression data was analyzed comprising early cardiac development (GSE1479), the developing prefrontal cortex (GSE4675), facial prominences (GSE7759), early development of the brain (GSE8091), development of the liver (GSE13149), ovary development (GSE5334), and development of testis (GSE4818). Quality was assessed by manual inspection of probe intensity distributions of each array and discarded if the MA-plots showed abnormal distributions. Figure 2.2A shows an MA-plot of a good quality array with an expected distribution where most genes show no differential expression. In contrast, Figure 2.2B demonstrated an MA-plot of a bad quality array with a skewed distribution that was discarded from the analysis. This quality control was also applied in Schramm and coworkers [247]. We discarded three samples from the dataset of early brain development, four samples from ovary development and one sample from testis development. For differentiation of human stem cells, we analyzed gene expression data of cardiomyocytes (GSE13834), chondrogenic differentiation (GSE10315), myoblast differentiation (GSE3780), myelopoiesis (GSE12837), and neural differentiation (GSE9940). Similar to the data sets for mouse, we discarded data with low quality. We discarded one sample from the cadiomyocytes, four samples from differentiation of chondrogenesis, 24 samples from differentiation of myoblasts, 11 samples from myelopoiesis and four samples from neural differentiation. The data was analyzed using the affymetrix package [94] of R (`www.r-project.org`) and normalized with VSN normalization [123]. For better comparability, for each gene expression study, time points were grouped into three time intervals: early, mid, and late expression, e.g. in the human myelopoiesis data set (GSE12837), the haematopoietic stem/progenitor cells (HSC) were grouped at the early time interval, myeloid precursors at the mid time interval and terminally differentiated cells at the late time interval. Each dataset was tested for differentially expressed genes between the different time intervals using the Rank Product Test [41]. Significant genes were determined using a cutoff for false positives smaller than 5% (false discovery rate $< 0.05$).

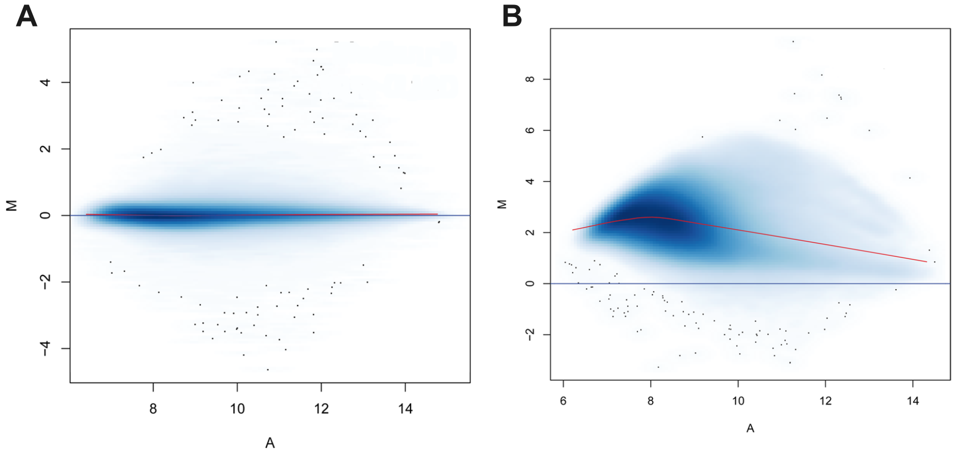


**Figure 2.2. MA-plots for quality control. A** MA-plot of a good quality array with an expected distribution. **B** MA-plot of a bad quality array with a skewed distribution.

### 2.1.5 Estimating tissue and time specificity for TFs, combinations of TFs and regulatory modules

For each TF we determined genes with binding sites for the TF identified by our PWM-scans and regarded them as potentially regulated by the specific TF. Using Fisher's Exact tests, we tested if these regulated genes were significantly enriched in the list of differentially expressed genes of each time interval for each gene expression study (tissue). We defined this TF to be tissue-specific if such an enrichment occurred only for one tissue (number of tissues = one), otherwise we specified this TF to regulate two or more tissues (number of tissues $> 1$). Similarly, we defined the TF to be time interval-specific if we determined an enrichment of its regulated genes in the list of differentially expressed genes of a tissue at one time interval (number of time intervals = one), and more than one time interval otherwise (number of time intervals $> 1$). This enrichment analysis was conducted for all TFs. The results were summarized for all TFs and the percentage of TFs per time interval and tissue identified, yielding the results shown in Figure 3.1A and Figure 3.3A. The same procedure was carried out for pairs of TFs at promoters (Figure 3.1B and Figure 3.3B), pairs of TFs at enhancers (Figure 3.1C and Figure 3.3C), and regulatory modules (Figure 3.1D and Figure 3.3D). To assess the signifcance of temporal specificity of regulatory modules compared to pairs of TFs at promoters, a Fisher's Exact test was conducted to test if the number of regulatory modules specific for a single time interval was enriched compared to the number of TF pairs at promoters specific at a single time interval.

### 2.1.6 Prediction of time intervals using regulatory modules

To identify relevant regulatory modules for temporal regulation of gene expression during development and differentiation and to estimate their potential power to regulate distinct gene groups for the progression of development, we employed the method of random forest as a machine learning method (classifier). We set up a classification task for two classes. For this, we identified all genes that were differentially expressed at only one time interval. As little data was available for the mid time interval (n = 7) and to simplify classification, we used only two time intervals (early and late). The early time interval constituted the first class and the late time interval the second class. The classifiers were trained to predict the correct time interval for each gene, using the information which specific regulatory modules were regulating the respective gene (regulatory modules served as features for the classifier). We trained 10,000 decision trees yielding an ensemble classifier (random forest) using the package randomForest [167] (`http://cran.r-project.org/web/packages/randomForest`) in R (`www.r-project.org`) with 80 variables randomly selected at each node (parameter $m_{try}$), a maximum node size of 2 (parameter *maxnodes*), and enabling the assessment of variable importance (parameter *importance*). To identify regulatory modules with the best discriminative behavior, we applied the Gini criterion which minimizes the impurity of the children nodes at each split in the tree. To focus on the best descriptors, we used the top 5% of the features for classification. A 10 times 10-fold cross-validation was applied to determine the performance of the classifier (yielding accuracy, sensitivity and specificity for the classifier). For comparison, we also trained a random forest using pairs of co-occurring transcription factors at promoters as features with the same parameters as for regulatory modules. Similar to regulatory modules, the most important pairs of transcription factors at promoters were identified according to the Gini criterion. The top 5% of the features were used for predictions and the performance of the classifier was determined employing a 10 times 10-fold cross-validation.

### 2.1.7 Definition of TFs with TSS-enriched, TSS-depleted and uniformly distributed binding sites

For each transcription factor, the distribution of binding sites was determined with respect to the annotated transcription start site (TSS) for all genes. Transcription factors were grouped into three categories: transcription factors with binding sites predominantly around the TSS (TSS-enriched-BS), transcription factors with a depletion of binding sites at the TSS (TSS-depleted-BS), and transcription factors showing a uniform distribution of binding sites (uniformly-distributed-BS). For this grouping, a Wilcoxon signed-rank test was conducted for each transcription factor to test if the distribution of binding sites at the TSS (+/- 100 bp around TSS) follows the distribution of the remaining binding sites. To correct for multiple testing, a Benjamini-Hochberg correction [26] was applied. Transcription factors with $P < 0.05$ and a difference of the medians of the distributions of at least four bp were classified as transcription factors preferentially binding at the TSS (TSS-enriched-BS) or as transcription factors with binding sites depleted around the TSS (TSS-depleted-BS) depending on the sign of the difference of the medians of the distributions. All other transcription factors were termed transcription factors with a uniform distribution of binding sites (uniformly-distributed-BS).

To further distinguish TFs with TSS-enriched-BS from TFs with TSS-depleted-BS, we determined the ratio (log-ratio) of the number of binding sites at promoters and enhancers per transcription factor. Binding sites at the promoter region and the enhancer region were counted per transcription factor for all genes. The number of binding sites was then normalized according to the width of the binding region (200 bp for the promoter region and 8,000 bp for the enhancer region). The log ratio of binding sites at promoters and enhancers was determined and compared between transcription factors preferentially binding at the TSS (TSS-enriched-BS) and transcription factors with a depletion of binding sites close to the TSS (TSS-depleted-BS). The difference of the log ratios between the two classes of TFs was determined using a Wilcoxon test.

### 2.1.8 Constructing the networks

Using the identified co-occurring transcription factor pairs as links (see section 2.1.2), two networks were constructed, one for promoters and one for enhancers. To assess if pairs of TFs of the same group (TSS-enriched-BS, TSS-depleted-BS, uniformly-distributed-BS) occurred more often than expected by chance, we performed a permutation test with 10,000 permutations of the class labels. Connectivity and betweenness centrality were determined for each node in the network and their maxima were identified for both networks for transcription factors of the categories TSS-enriched-BS and TSS-depleted-BS. In addition, a protein-protein-interaction network of TFs was constructed using physical binding information from a public repository (BIND [16]) and each transcription factor was associated to its corresponding protein in the network. This network was analyzed for the same properties as the described promoter and enhancer networks.

## 2.2 Identification of novel putative SREBP target genes

The complete workflow is shown in Figure 2.3. Whereas I conducted all bioinformatics analyses, all experimental procedures were carried out by Jessica Schilde and Heiko Runz at the Institute of Human Genetics of the University Hospital Heidelberg in the group of Heiko Runz "Regulation of cellular cholesterol metabolism".

### 2.2.1 Gene expression analysis of HeLa cells and patient fibroblast cell lines

Gene expression data of HeLa cells were taken from Bartz and Kern and co-workers [22] and were downloaded from the public microarray database ArrayExpress (Acc.No. E-TABM-599). The raw data was normalized with VSN normalization [123]. Of the original cDNA probe set, 40,847 cDNA-clones (Unigene Build 215) were mapped to 17,848 human genes (NCBI Build 36.3) for further analysis. Median values were taken if a gene was represented by multiple probes on the array. The fibroblasts were cultivated under similar cell culture conditions as the HeLa cells as described in [22] and mRNA of three biological replicates each was hybridized against Illumina Human Sentrix-8 chips. Labeling, hybridization and scanning of the Illumina chips was performed in the Genomics and Proteomics Core Facility of the German Cancer Research Center according to Illumina's recommended protocols. The raw data of the fibroblasts was analyzed using the lumi package [73] of R (`www.r-project.org`) and normalized with RSN normalization [168]. Both datasets were tested for differentially expressed genes of normal and sterol-depleted conditions using the Rank Product Test [41]. Significant genes were determined using a cutoff of false positives smaller than 5% (false discovery rate $< 5\%$). Differentially regulated genes were compared

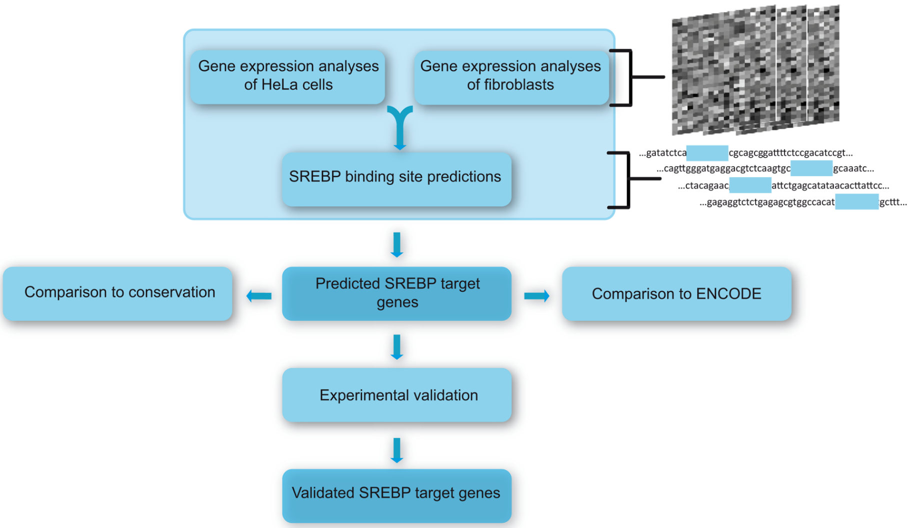

Gene expression analyses of HeLa cells
Gene expression analyses of fibroblasts
SREBP binding site predictions
...gatatctca cgcagcggattttctccgacatccgt...
...cagttgggatgaggacgtctcaagtgc gcaaatc...
...ctacagaac attctgagcatataacacttattcc...
...gagaggtctctgagagcgtggccacat gcttt...
Predicted SREBP target genes
Comparison to conservation
Comparison to ENCODE
Experimental validation
Validated SREBP target genes

**Figure 2.3. Workflow of data integration and identification of SREBP target genes.** Differentially expressed genes in sterol-depleted medium were identified in fibroblasts and HeLa cells. Promoter regions of differentially expressed genes were screened *in silico* for SREBP binding sites and genes selected with binding motifs for SREBP. This yielded putative SREBP target genes which were compared to genes coding for enzymes in cholesterol biosynthesis and any known relationships to SREBP. Identified binding sites were compared to ChIP-Seq data from the ENCODE project [65] for SREBP and NF-Y (ChIP-Seq data for SREBP1a and SREBP2 were taken from a study by the lab of Michael Snyder at Yale University and NF-YA and NF-YB from a study by the lab of Kevin Struhl at Harvard and the data was downloaded from UCSC [236]) and their conservation to chimp and mouse was assessed. Selected genes were validated experimentally using qRT-PCR and SREBP knockdown experiments.

to already known genes associated to cholesterol and fatty acid biosynthesis using Gene Ontology terms (`www.geneontology.org`). The mapping of associated Gene Ontology terms for each gene was downloaded from NCBI (Build 36.3) and parsed for the terms sterol, steroid, lipid and fatty acid. In addition, enrichment of genes encoding for the 22 enzymes necessary for cholesterol biosynthesis (*acly*, *acas2*, *acat2*, *hmgcs1*, *hmgcr*, *mvk*, *pmvk*, *mvd*, *idi1*, *fdft1*, *sqle*, *lss*, *cyp51a1*, *tm7sf2*, *sc4mol*, *h105e3*, *hsd17b7*, *ebp*, *sc5dl*, *dhcr7*, *dhcr24*; see [22]) was tested using a Fisher's exact test.

### 2.2.2 Genome-wide *in silico* promoter screen and identification of genes with SREBP binding sites

Sequences from 10,000 bp upstream to 1,000 bp downstream of the transcription start site for 32,121 human genes were retrieved from NCBI (Build 36.3) (`ftp://ftp.ncbi.nlm.nih.gov/`). We used a total of 21 position weight matrices (PWMs) for SREBP1 and its isoforms a and c as well as 7 for SREBP2, 6 for SP1, 5 for NF-Y and its isoforms alpha and beta, and 4 PWMs for the LXR isoforms LXR-alpha and LXR-beta which were taken from TRANSFAC (Release 12.1) [183]. The promoter screen using these position-weight matrices was conducted as described in [227, 294] with R (`http://www.r-project.org`). Predicted binding sites with a p-value below 0.05 were used for our analysis where each PWM binding site corresponds to a binding site of its associated transcription factors. For genes with predicted SREBP binding sites, binding sites for SP1, NF-Y, and LXR were determined. Predicted SREBP target genes were compared to already known genes associated to cholesterol and fatty acid biosynthesis in the same way as described for differentially expressed genes (see section 2.2.1). Gene Ontology enrichment analysis for the identified SREBP target genes was conducted using topGO [4] with R (`http://www.r-project.org`) using the classic algorithm for scoring significance of GO terms. To correct for multiple testing, a Benjamini-Hochberg correction [26] was applied. GOterms with a p-value $< 0.05$ were taken for further considerations. Additional biological roles of identified SREBP target genes were also identified using topGO [4] with R (`http://www.r-project.org`) using the weight algorithm for scoring significance of GO terms. GOterms with a p-value $< 0.05$ were taken for further considerations.

### 2.2.3 Identification of putative SREBP target genes and comparison to existing sequence data of chromatin immunopreciptation screens (ChIP-Seq)

We compared predicted SREBP target genes to ChIP-Seq data from the ENCODE project [65] for SREBP and NF-Y. ChIP-Seq data for SREBP1a and SREBP2 were taken from a study in HepG2 cells by the lab of Michael Snyder at Yale University and NF-YA and NF-YB from a study in K-562 cells by the lab of Kevin Struhl at Harvard. For sterol deprivation, HepG2 cells were cultured with pravastatin (2 $\mu$M; Sigma) in DMEM with 0.5% BSA for 16 h. The data was downloaded from UCSC [236]. Genome coordinates of peak hits were compared to gene annotations (NCBI Build 36.3) and target genes were determined using the same settings as for the *in silico* promoter screen. Binding sites occurring within a range of -10kb and +1kb of the annotated transcription start site of a gene were included in the analysis. The determined genes were then compared to the list of identified putative SREBP target genes. Predicted target sites were also analyzed for evolutionary conservation. To determine the conservation of binding sites of the predicted SREBP target genes, we analyzed pair-wise alignments between human and mouse, and human and chimp, respectively. Chained and netted pair-wise alignments of human (UCSC version hg18) with chimp (UCSC version panTro2) and mouse (UCSC version mm9) were downloaded from UCSC [236] in the axtNet format (`ftp://hgdownload.cse.ucsc.edu/`). Conserved regions between human-chimp and human-mouse were determined by the given aligned regions in the alignment files. Predicted binding sites for SREBP were compared to the identified conserved regions and taken as conserved if binding sites occurred in conserved regions. To compare the conservation of SREBP binding sites to the conservation of all TF binding sites in the promoter sequences of the predicted SREBP target genes, the number of conserved bindings sites for all transcription factors in chimp and mouse was determined in the same way as described for SREBP. A one-sided Wilcoxon test was conducted to assess the signifcance level of the conservation of SREBP binding sites compared to the number of conserved binding sites for all transcription factors.

### 2.2.4 Cell culture and sterol depletion

HeLa kyoto cells and human fibroblasts (KOA-1) were plated onto 100 mm cell culture dishes (SPL life sciences) and cultivated at 37 °C, 5% $CO_2$ in either DMEM/ 1g/l Glucose/with L-Glutamine (PAA), 1% (v/v) Penicillin/Streptomycin (100x) (PAA) and 5% FBS (Biochrom) (HeLa cells) or DMEM/ 1g/l Glucose/ with L-Glutamine, 1% (v/v) Penicillin/Streptomycin (100x), 1% (v/v) Amphotericin B (250 $\mu$g/ml) (PAA) and 10% FBS (KOA1). After reaching a cell density of ~60%, cells were either cultivated in control media or sterol depleted media without FBS but 0.5 % LDS (Pan Biotech) (HeLa cells) or 5% LDS (KOA-1). After 96 hours sterol depleted cells were additionally treated with 1% (w/v) (2-Hydroxypropyl)-$\beta$-cyclodextrin (HPCD, Sigma) for 3 hours and afterwards cultivated for additional 3 hours in sterol depleted media.

### 2.2.5 siRNA treatment

For SREBP knockdown experiments HeLa cells were plated onto 100 mm cell culture dishes and cultivated at 37 °C, 5% $CO_2$ in DMEM/ 1g/l Glucose/ with L-Glutamine and 5% FBS one day before siRNA treatment. Transfection was performed with Oligofectamine$^{TM}$Reagent (Invitrogen) using a negative control siRNA (Silencer® Select 4390843, Ambion), siRNA against *srebf1* (Silencer® Select 4392420, Ambion), *srebf2* (Silencer® Select 4390824, Ambion) or both. Medium was changed 24 hours after transfection and cells were either cultivated in control or sterol depleted media, respectively. 48 hours after transfection sterol depleted cells were treated with 1% HPCD as described in section 2.2.4.

### 2.2.6 RNA isolation and quantitative real-time PCR

For gene expression experiments via qRT-PCR RNA was isolated with the InviTrap® Spin Cell RNA Mini Kit (Invitek). RNA was then reverse transcribed by RevertAid$^{TM}$H Minus m-MuL V Reverse Transcriptase (200 u/$\mu$l) (Fermentas) using random primers (Invitrogen). QRT-PCR was performed with Power SYBR®Green (Applied Biosystems) using a 7500 Fast Real-Time PCR System (Applied Biosystems). Three or four independent RNA samples were analyzed for each gene and differential expression was calculated in relation to housekeeping gene *rpl19*.

# Chapter 3

# Results

## 3.1 Identification of spatio-temporal specific regulatory modules

### 3.1.1 Identifying regulatory modules

To identify regulatory modules, we performed a genome-wide screen for transcription factor binding sites using position weight matrices (PWM-scans) for all annotated human genes and transcription factors [227, 294]. Figure 2.1 depicts the workflow of the method. The sequence upstream and downstream (+/- 100 base pairs) of the annotated transcription start site (TSS) was termed promoter region whereas the studied enhancer region was further upstream of the TSS (2,000-10,000 base pairs upstream). To identify interacting transcription factors at promoters and enhancers, we selected pairs of co-occurring transcription factor binding sites in a defined window at the promoter and enhancer region for each gene, respectively. We then combined identified pairs of co-occurring transcription factors at the promoter and enhancer region for each gene to analyze combinations of promoter and enhancer interactions. These combinations were termed regulatory modules. After filtering (see section 2.1.3), we identified 129 regulatory modules binding at 340 genes. To generalize our investigations, we repeated the analysis and identified regulatory modules also for mouse and rat. Regulatory modules for mouse and rat showed similar results when applying the same settings as for human (see Table 3.1).

**Table 3.1. Overview of the number of identified transcriptional regulators for different organisms.**

| | Human | | Mouse | | Rat | |
|---|---|---|---|---|---|---|
| | Regulatory elements | Genes | Regulatory elements | Genes | Regulatory elements | Genes |
| Transcription factors (TFs) | 132 | 32121 | 123 | 33033 | 132 | 27110 |
| TF pairs at promoters | 111 | 3007 | 77 | 1931 | 74 | 1891 |
| TF pairs at enhancers | 579 | 11172 | 418 | 10985 | 585 | 8326 |
| Regulatory modules | 129 | 340 | 113 | 311 | 28 | 134 |

### 3.1.2 Identified regulatory modules regulate spatio-temporal gene expression in development

To investigate time- and tissue-specific regulatory roles of the identified regulatory modules in development, we analyzed time series of gene expression profiles of embryonic development in mouse and embryonic stem cell differentiation in human cells.

#### Human stem cell differentiation

We selected gene expression studies from a broad range of different human stem cells of different origin. Each study was regarded as tissue-specific. For better comparison among the different studies, we grouped time points for each gene expression study into three distinct time intervals we termed early, mid, and late expression. For each gene expression study, we identified differentially expressed genes at these time intervals and determined their respective regulation by transcription factors, pairs of co-occurring transcription factors and regulatory modules employing enrichment analyses (see chapter 2). We compared the number of enriched tissues and time intervals for single transcription factors, pairs of co-occurring transcription factors and regulatory modules. Strikingly, regulatory modules showed the highest tissue and temporal specificity. Figure 3.1 shows the results for human stem cells. Only 16% of transcription factors were specific for a single tissue whereas 76% of pairs of co-occurring transcription factors in promoter regions, 77% of pairs of co-occurring transcription factors in enhancer regions and 79% of regulatory modules showed specificity for a single tissue. Temporal specificity was even more distinctive. Whereas only 40% of the studied transcription factors were

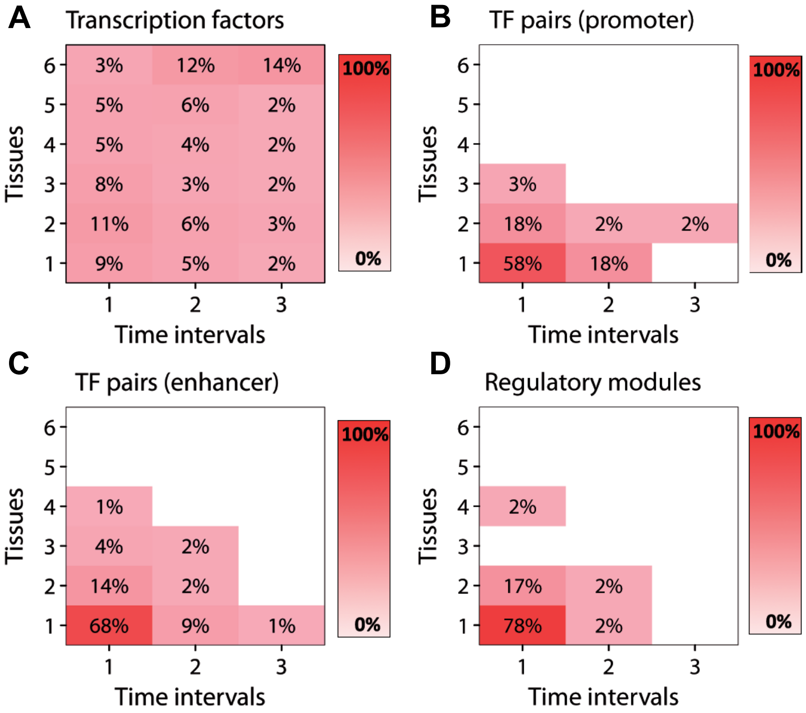


**Figure 3.1. Tissue and temporal specificity for each regulatory element during differentiation of human stem cells.** The number of tissues versus the number of time intervals is plotted for **A** transcription factors, **B** pairs of co-occurring transcription factors in promoter regions, **C** pairs of co-occurring transcription factors in enhancer regions, and **D** regulatory modules. The percentage of the different regulatory elements is indicated at each entry in the grid (i.e. 9% of transcription factors are specific for a single tissue and a single time interval). The color also represents the percentage of regulatory elements at each point in the grid; bright red indicates a high amount of regulatory elements whereas light pink indicates a low amount of regulatory elements at the respective entry.

specific for a single time interval, 79% of pairs of co-occurring transcription factors in promoter regions and 85% in enhancer regions and even 97% of the regulatory modules showed specificity for a single time interval in the data sets of human stem cells (Figure 3.2). Ravasi and coworkers [229] showed that pairs of transcription factors rather than single transcription factors determine tissue specificity. Surprisingly, the additional temporal specificity of regulatory modules is obtained by pairs of co-occurring transcription factors at enhancers (97% for regulatory modules versus 85% for pairs of co-occurring transcription factors, significance of the difference: P = 0.01).

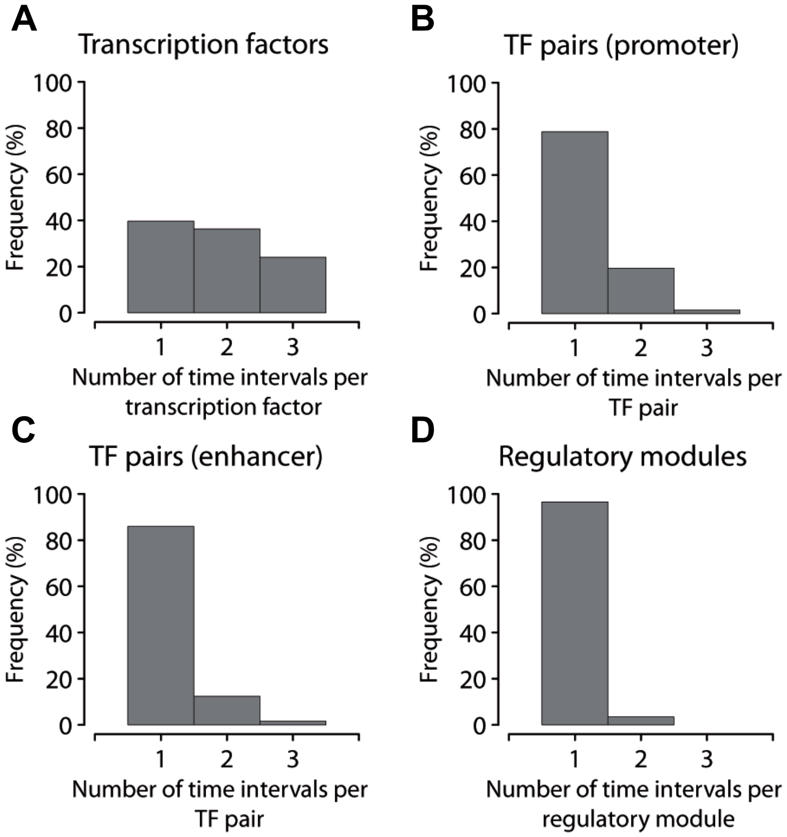


**Figure 3.2. Frequency distribution of the number of time intervals for human stem cell differentiation.** The histograms show the frequency distribution of the number of time intervals for **A** transcription factors, **B** pairs of co-occurring transcription factors in promoter regions, **C** pairs of co-occurring transcription factors in enhancer regions, and **D** regulatory modules.

### Mouse embryonic development

Similar results to the human stem cell differentiation data sets were obtained for mouse embryonic development. The analysis was conducted in the same manner as for human stem cell differentiation. As described in section 2.1.4, we selected gene expression studies from a broad range of different embryonic mouse tissues of different origin. Each study was regarded as tissue-specific. For better comparison among the different studies, we also grouped time points for each gene expression study into three distinct time intervals we termed early, mid, and late expression. For each gene expression study, we identified differentially expressed genes at these time intervals

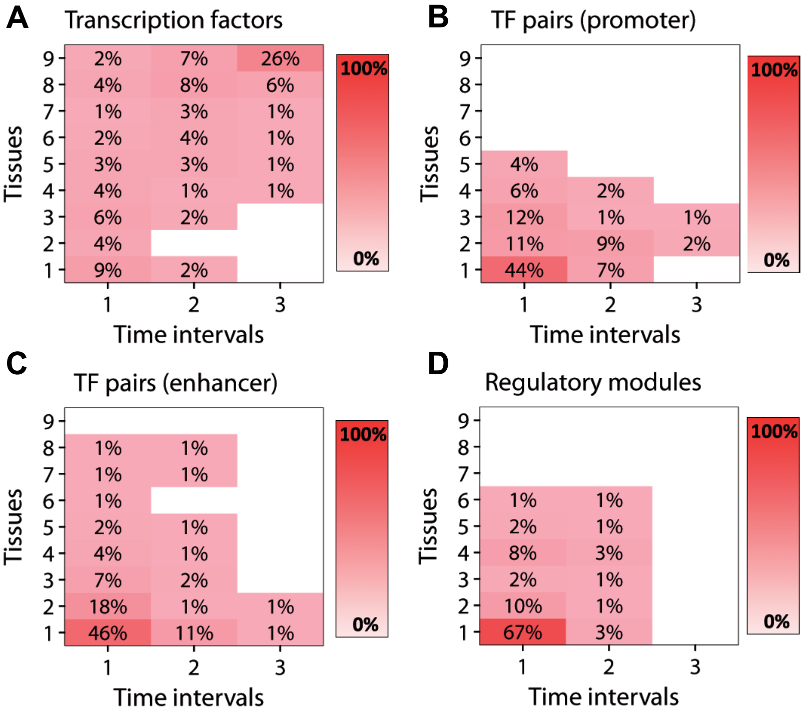


**Figure 3.3. Tissue and temporal specificity for each regulatory element in the embryonic development of mouse.** The number of tissues versus the number of time intervals is plotted for **A** transcription factors, **B** pairs of co-occurring transcription factors in promoter regions, **C** pairs of co-occurring transcription factors in enhancer regions, and **D** regulatory modules. The percentage of the different regulatory elements is indicated at each entry in the grid (i.e. 9% of transcription factors are specific for a single tissue and a single time interval). The color also represents the percentage of regulatory elements at each point in the grid; bright red indicates a high amount of regulatory elements whereas light pink indicates a low amount of regulatory elements at the respective entry.

and determined their respective regulation by transcription factors, pairs of co-occurring transcription factors and regulatory modules employing enrichment analyses (see chapter 2). We compared the number of enriched tissues and time intervals for single transcription factors, pairs of co-occurring transcription factors and regulatory modules. Similarly to the results of human stem cell differentiation, regulatory modules showed the highest tissue and temporal specificity. Figure 3.3 and Figure 3.4 show the results for mouse embryonic development. Whereas 34% of the transcription factors were specific for a single time interval and 11% for a single tissue, 77% and 52% of pairs

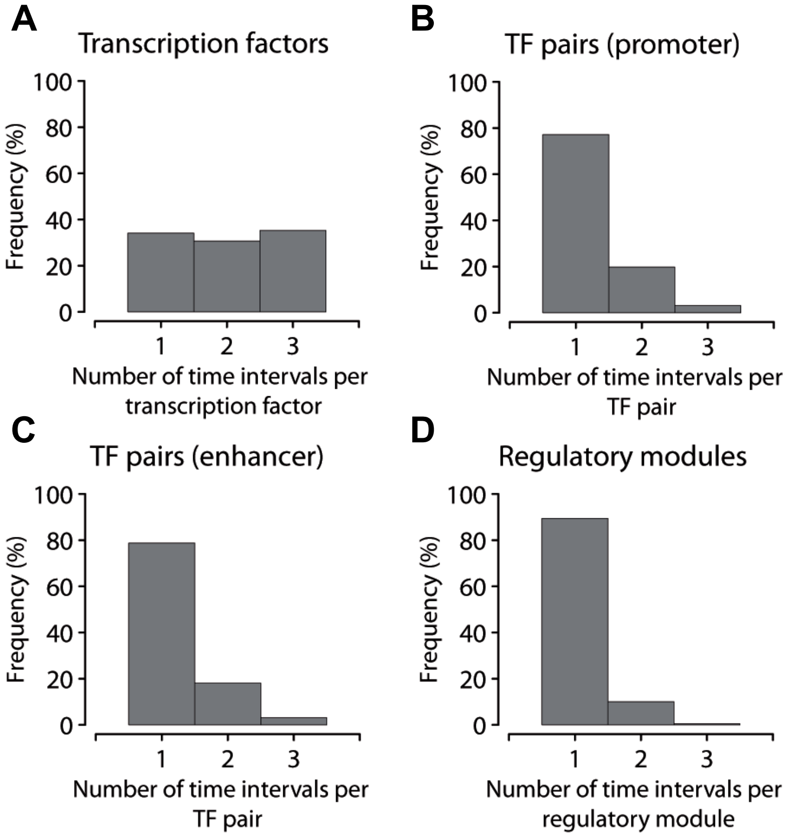


**Figure 3.4. Frequency distribution of the number of time intervals for mouse embryonic development.** The histograms show the frequency distribution of the number of time intervals for **A** transcription factors, **B** pairs of co-occurring transcription factors in promoter regions, **C** pairs of co-occurring transcription factors in enhancer regions, and **D** regulatory modules.

of co-occurring transcription factors at promoters, 79% and 58% of pairs at enhancers, and 89% and 69% of regulatory modules showed specificity for a single time interval and tissue, respectively.

### Enhancers determine temporal specificity

As seen for both mouse embryonic development and human stem cell differentiation, the combinations of regulatory factors at promoters and enhancers resulted in higher specificity of tissue and temporal regulation during development and differentiation. Concluding, transcription factors binding at promoters contributed significantly to tissue-specific regulation whereas regulatory factors at enhancers rather accounted for temporal specificity.

To cross-check the specificity of these regulatory modules, we constructed regulatory modules consisting of combinations of pairs of co-occurring transcription factors at promoters only and repeated the analysis. Using the same parameter settings, we identified a limited number of regulatory modules (n = 12) that did not allow any conclusion about tissue and temporal specificity. Even increasing the promoter region by a factor of ten (which resulted in a sufficient number of regulatory modules) revealed 83% of regulatory modules as tissue-specific but only 61% as time-specific. These results further support the fact that the combinations of promoter-enhancer interactions establish temporal specificity of gene expression.

### 3.1.3 Regulatory modules predict temporal gene expression in development

To identify regulatory modules (combinations of transcription factors) for temporal specificity, we further analyzed the active regulatory modules during human stem cell differentiation. We learned a classifier (of a random forest) to predict the time interval (now simplified for two categories, early and late) of differential expression for each gene based on its regulatory modules. This way, we were able to predict temporal differential expression based on the profile of regulatory modules. Specifically, we predicted the combination of pairs of co-occurring transcription factors in promoter and enhancer regions, which determine the temporal regulation observed during development. The top 10 regulatory modules explaining best temporal specificity are shown in Table 3.2. TF-groups SP1 (Sp1 transcription factor), EGR1 (early growth response 1) and E2F1 (E2F transcription factor 1) were the most observed transcription factors occurring in promoter regions, whereas members of the forkhead box family of transcription factors (FOXI1, FOXJ1, FOXD3, FOXF1, FOXL1, and FOXA1) and CDX1 (caudal type homeobox 1) were mostly found at enhancer regions. To validate our results, we performed a stratified 10 times 10-fold cross-validation and trained with the top 5% of regulatory modules yielding a considerably good prediction performance (70% accuracy, 73% sensitivity, 69% specificity). In comparison, pairs of co-occurring transcription factors at promoters were not sufficient to predict temporal gene expression and failed to detect differences between the time intervals (43% accuracy, 21% sensitivity, 76% specificity). These results support the specificity of the identified regulated modules for temporal gene expression.

**Table 3.2. Top ten of the list of identified regulatory modules explaining temporal specificity for the differentiation of human stem cells.**

| Regulatory modules[*] | Additional members of the TF-group | Binding preference[+] |
|---|---|---|
| SP1 SP1 - FOXI1 FOXJ1 | SP2, SP3, SP4 (SP1)<br>FOXD3, FOXF1, FOXF2 (FOXJ1) | TSS-enriched-BS (SP1)<br>TSS-depleted-BS (FOXI1,FOXJ1) |
| SP1 SP1 - FOXJ1 FOXJ1 | SP2, SP3, SP4 (SP1)<br>FOXD3, FOXF1, FOXF2 (FOXJ1) | TSS-enriched-BS (SP1)<br>TSS-depleted-BS (FOXJ1) |
| SP1 SP1 - CDX1 FOXA1 | SP2, SP3, SP4 (SP1)<br>CDX2 (CDX1)<br>FOXA2, FOXA3 (FOXA1) | TSS-enriched-BS (SP1)<br>TSS-depleted-BS (CDX1,FOXA1) |
| SP1 SP1 - FOXI1 FOXA1 | SP2, SP3, SP4 (SP1)<br>FOXA2, FOXA3 (FOXA1) | TSS-enriched-BS (SP1)<br>TSS-depleted-BS (FOXI1,FOXA1) |
| EGR1 SP1 - FOXI1 FOXA1 | EGR2, EGR3, EGR4 (EGR1)<br>SP2, SP3, SP4 (SP1)<br>FOXA2, FOXA3 (FOXA1) | TSS-enriched-BS (EGR1,SP1)<br>TSS-depleted-BS (FOXI1,FOXA1) |
| SP1 SP1 - FOXJ1 FOXJ2 | SP2, SP3, SP4 (SP1)<br>FOXD3, FOXF1, FOXF2 (FOXJ1) | TSS-enriched-BS (SP1)<br>TSS-depleted-BS (FOXJ1,FOXJ2) |
| SP1 SP1 - FOXL1 FOXL1 | SP2, SP3, SP4 (SP1) | TSS-enriched-BS (SP1)<br>TSS-depleted-BS (FOXL1) |
| E2F1 E2F1 - FOXL1 FOXA1 | E2F2, E2F3, E2F4, E2F5, E2F7, TFDP1 (E2F1)<br>FOXA2, FOXA3 (FOXA1) | TSS-enriched-BS (E2F1)<br>TSS-depleted-BS (FOXL1,FOXA1) |
| E2F1 EGR1 - FOXJ2 FOXL1 | E2F2, E2F3, E2F4, E2F5, E2F7, TFDP1 (E2F1)<br>EGR2, EGR3, EGR4 (EGR1) | TSS-enriched-BS (E2F1,EGR1)<br>TSS-depleted-BS (FOXJ1,FOXL1) |
| EGR1 SP1 - CDX1 FOXI1 | EGR2, EGR3, EGR4 (EGR1)<br>SP2, SP3, SP4 (SP1)<br>CDX2 (CDX1) | TSS-enriched-BS (EGR1,SP1)<br>TSS-depleted-BS (CDX1,FOXI1) |

[*] The first two transcription factors were identified at promoters, the last two at enhancers.

[+] TSS-enriched-BS: Transcription factors with binding sites predominantly around the TSS; TSS-depleted-BS: transcription factors with a depletion of binding sites at the TSS.

### 3.1.4 Transcription factors show distinct binding site distributions for promoter and enhancer regions

To identify differences among transcription factors binding preferentially either at promoter or enhancer regions of the identified regulatory modules, we analyzed the distributions of binding sites for all transcription factors with respect to the annotated transcription start site (TSS). Interestingly, we identified three different binding site distributions for the analyzed transcription factors. Figure 3.5 shows exemplarily the distributions for the TF-groups SP1, FOXA1 and TP53 (tumor protein p53). The distribution of SP1 showed an enrichment of binding sites close to the TSS (Figure 3.5A). Binding sites with these distributions were termed TSS-enriched-BS, whereas FOXA1 exhibited a depletion of binding sites at the TSS (TSS-depleted-BS, Figure 3.5B). We also observed rather uniform distributions for e.g. TP53 (uniformly-distributed-BS, Figure 3.5C). We investigated the motifs of these three groups and found that transcription factors with TSS-enriched-BS had binding sites with a higher GC content compared to the other transcription factors (P = 8.22E-14). This is consistent with reports that sequences at TSS are often GC rich [83, 114, 147]. All transcription factors occurring at promoters of the identified regulatory modules had TSS-enriched-BS, and 91% of the transcription factors at enhancers had TSS-depleted-BS. Notably, this tendency was even stronger for the regulatory modules selected by the classification algorithm (100% TSS-enriched-BS for the promoter pairs and 100% TSS-depleted-BS for the enhancer pairs of regulatory modules). When applying the analysis to all transcription factors analyzed, the majority of transcription factors (53%) showed a uniform distribution of binding sites with no preferential binding position (uniformly-distributed-BS). 21% of transcription factors were determined to preferentially bind close to the TSS (TSS-enriched-BS), whereas 26% transcription factors showed a depletion of binding sites around the TSS (TSS-depleted-BS). Binding preferences for all transcription factors are shown in Table A.1. It is to note that although previous studies identified transcription factors with preferential binding close to the TSS [83, 147, 271, 282, 302, 304], transcription factors showing a depletion of binding sites around the TSS or a uniform binding site distribution have been noted [83] but have not been quantified so far.

To further distinguish TFs with TSS-enriched-BS from TFs with TSS-depleted-BS, we determined the ratio (log-ratio) of the number of binding sites at promoters and enhancers per transcription factor. Binding sites at the promoter region (+/- 100 bp around the TSS) and the enhancer region (2,000 bp to 10,000 bp upstream of the TSS) were counted per transcription factor for all genes. The number of binding sites was then normalized

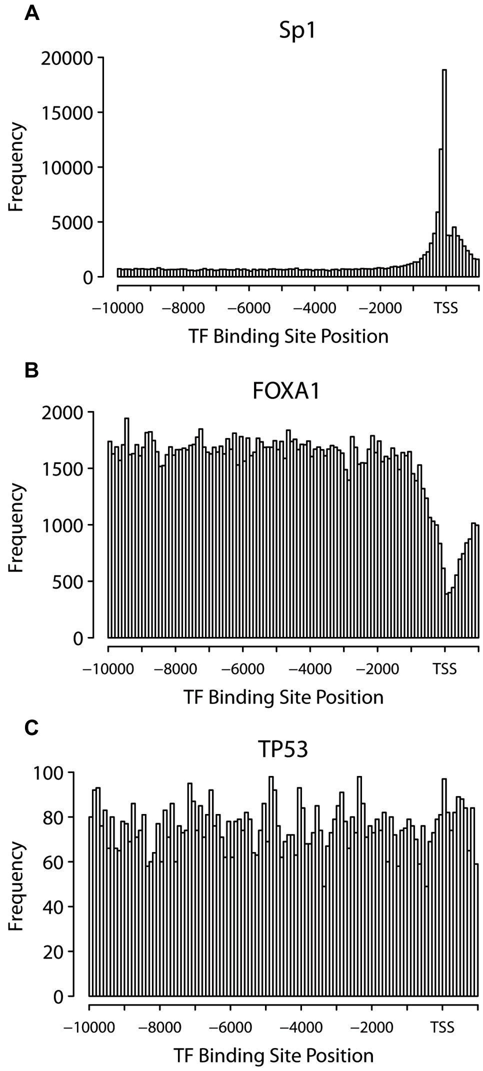


**Figure 3.5. Distribution of binding sites for different groups of transcription factors.** For different groups of transcription factors the distribution of binding sites with respect to the transcription start site is shown exemplarily for the transcription factors SP1, FOXA1, and TP53: **A** The distribution for SP1 which represents the distribution of binding sites for transcription factors preferentially binding at the transcription start site (TSS-enriched-BS), **B** the distribution of FOXA1 which represents the distribution of binding sites for transcription factors with a depletion of binding sites at the TSS (TSS-depleted-BS), and **C** the distribution for TP53 representing uniformly distributed binding sites (uniformly-distributed-BS).

according to the width of the binding region. The log ratio of binding sites at promoters and enhancers was determined and compared between transcription factors preferentially binding at the TSS (TSS-enriched-BS) and transcription factors with a depletion of binding sites close to the TSS (TSS-depleted-BS). We identified a major difference between the ratios for the two classes of transcription factors using a Wilcoxon test. Transcription factors with TSS-enriched-BS had significantly more binding sites at promoters whereas transcription factors with TSS-depleted-BS had more binding sites at enhancer regions (P < 2.2e-16, Figure 3.6).

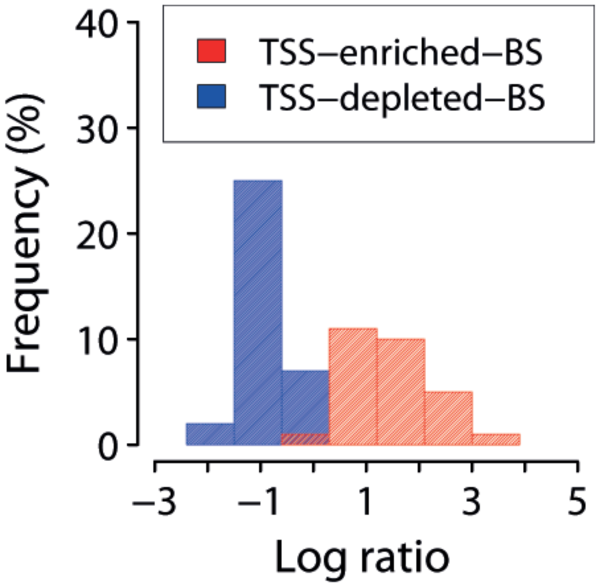


**Figure 3.6. Log ratio of binding sites at promoters and enhancers.** The distribution of the log-ratios of binding sites at promoters and enhancers is shown in red for transcription factors with preferential binding around the transcription start site (TSS-enriched-BS) and in blue for transcription factors with a depletion of binding sites at the TSS (TSS-depleted-BS).

### 3.1.5 Network analysis of transcription factors

To further analyze characteristics of transcription factors with different binding site distributions in promoter and enhancer regions, we constructed two networks. A link in the networks was set for each pair of co-occurring transcription factors identified at promoters for the promoter network and enhancer regions for the enhancer network. Interestingly, in both networks, the majority of transcription factors with TSS-enriched-BS was

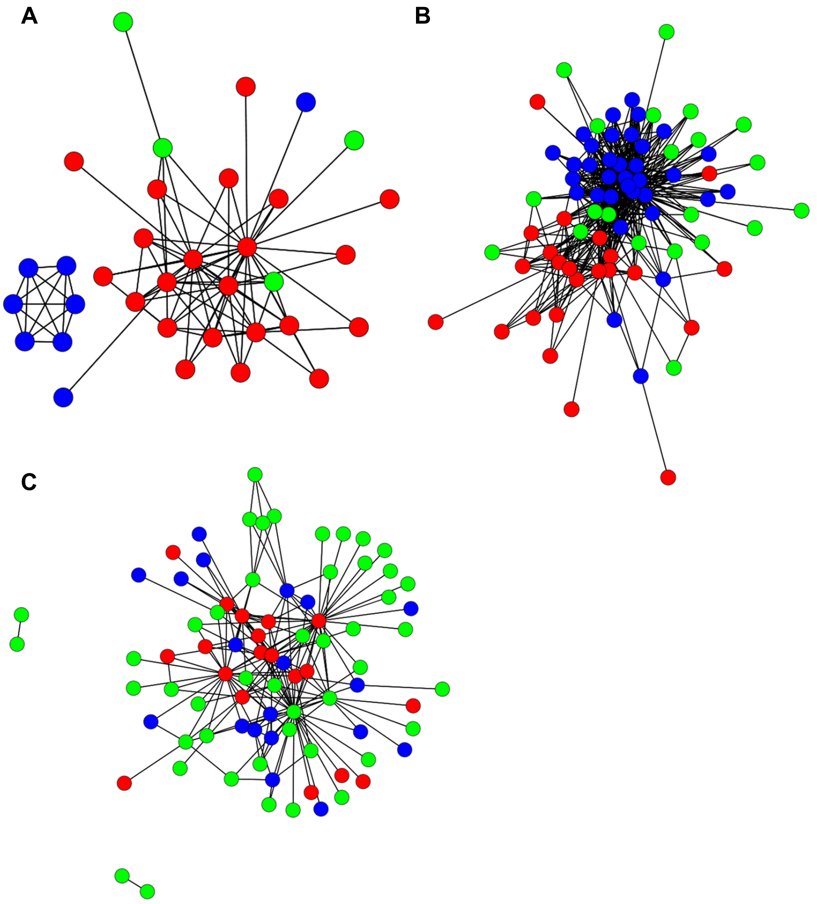


**Figure 3.7. Networks of transcription factor pairs (human).** The network of pairs of co-occurring transcription factors are shown for **A** the promoter regions and **B** the enhancer regions. **C** A network of transcription factors mapped onto a PPI network [16]. Transcription factors showing preferential binding around the transcription start site (TSS-enriched-BS) are marked in red, transcription factors with a depletion of binding sites around the TSS (TSS-depleted-BS) in blue and transcription factors showing no preferential binding (uniformly-distributed-BS) in green. In the promoter network, transcription factors with TSS-depleted-BS formed a small cluster, including FOXA1, FOXI1, FOXJ1A, FOXJ2, FOXL1, and CDX2.

adjacent to transcription factors of the same entity (TSS-enriched-BS) (significant (P = 0.002) for the promoter network and tendency (P = 0.1) for the enhancer network). Similarly, transcription factors with TSS-depleted-BS were preferentially adjacent to transcription factors with TSS-depleted-BS (P = 0.002 for the promoter network and P = 0.001 for

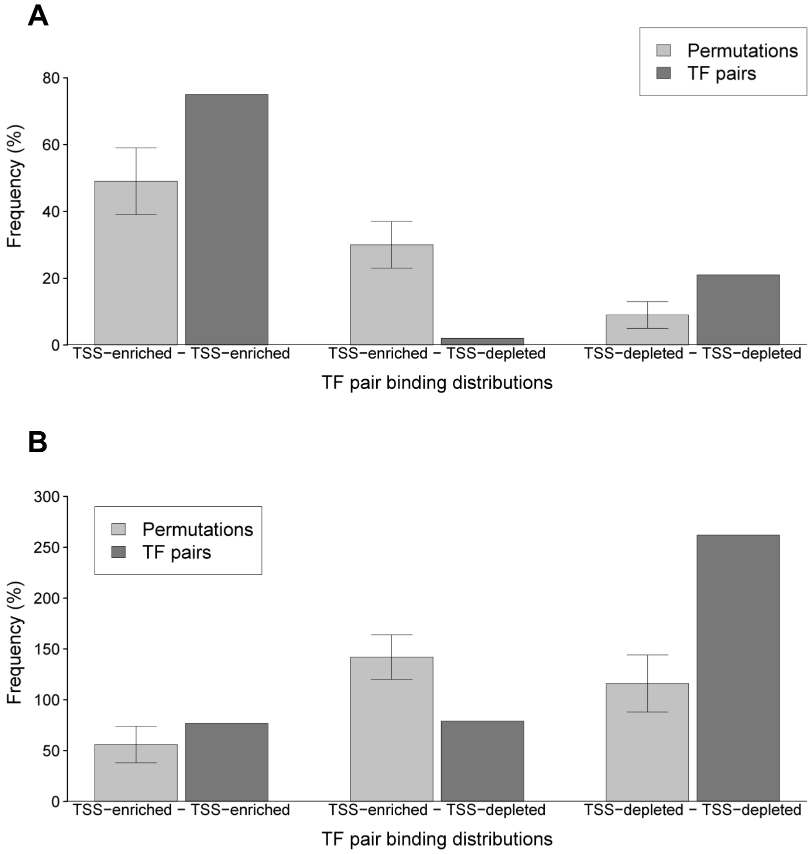


**Figure 3.8. Distribution of binding site preferences of pairs of co-occurring transcription factors in the promoter and the enhancer networks.** The number of pairs with specific binding site preferences is shown in dark gray. To get a null distribution, node labels were permuted 10,000 times and the mean number of TFs pairs with specific binding preferences determined (light gray). TSS-enriched-BS represent transcription factors showing preferential binding around the transcription start site, TSS-depleted-BS transcription factors with a depletion of binding sites at the TSS, and uniformly-distributed-BS transcription factors showing no preferential binding. The results for the promoter network are shown in **A**, the results for the enhancer network in **B**.

the enhancer network). In the promoter network, transcription factors with TSS-depleted-BS formed a small cluster and were not connected to transcription factors with TSS-enriched-BS. These TSS-enriched-BS transcription factors include FOXA1, FOXI1, FOXJ1A, FOXJ2, FOXL1, and CDX2. Figure 3.7 shows the networks and Figure 3.8 the distributions of

**Table 3.3. Overview of network properties.**

| | Promoter network | | Enhancer network | | PPI network | |
|---|---|---|---|---|---|---|
| | TSS-enriched | TSS-depleted | TSS-enriched | TSS-depleted | TSS-enriched | TSS-depleted |
| Quantity | 17 | 8 | 23 | 34 | 19 | 19 |
| Connectivity* | 50 | 12 | 124 | 64 | 62 | 20 |
| Betweenness centrality* | 351.5 | 0 | 349.4 | 1062.4 | 2055.2 | 401.3 |

* For each transcription factor group, the maximum is shown.

transcription factors in the promoter and enhancer networks.

As shown in Table 3.3, most transcription factors in the promoter network had TSS-enriched-BS (65%), whereas only 24% of transcription factors had TSS-depleted-BS. To further characterize distinct roles for transcription factors with different binding site distributions, we determined connectivity and betweenness centrality for each node in the networks. The transcription factor SP1 with TSS-enriched-BS had the highest connectivity and highest centrality in the promoter network with a connectivity of 50 and betweenness centrality of 351.5. In contrast, the highest connectivity of a transcription factor with TSS-depleted-BS was 12 and the betweenness centrality was zero for all transcription factors with TSS-depleted-BS. These results supported the fact that transcription factors with TSS-enriched-BS played a central role in the promoter network. These transcription factors constituted the main component of the network (Figure 5A) while transcription factors with TSS-depleted-BS formed a rather small and separated component. In contrast, transcription factors with TSS-depleted-BS played a central role in the enhancer network. These transcription factors constituted the core of the enhancer network (Figure 5B) with other transcription factors at its periphery. 44% of the transcription factors in the enhancer network had TSS-depleted-BS whereas only 29% of the transcription factors had TSS-enriched-BS. The fork head transcription factor FOXA1 (forkhead box A1) with TSS-depleted-BS had the highest connectivity (124) and centrality (1062.4) in the enhancer network. In contrast, the highest connectivity of a transcription factor with TSS-enriched-BS was 64 and the highest betweenness centrality was 349.4. In addition, we constructed a network of known transcription factor interactions (physical binding of pairs of transcription factors, obtained from a public repository [16]). Interestingly, the number of transcription factors with TSS-enriched-BS and TSS-depleted-BS was balanced (both 23%) and these transcription factors were located rather at the core of the network (Figure 5C). TBP (TATA box binding protein) with TSS-enriched-BS showed the highest connectivity (62) and betweenness centrality (2055.2) in the network compared to transcription factors with TSS-depleted-BS which had a maximum connectivity of 20 and a maximum centrality of 401.3.

## 3.2 Identification of novel putative SREBP target genes

The regulation of cholesterol biosynthesis is one example where transcriptional activation is achieved by the combined action of several transcription factors. The key regulator, the sterol regulatory element binding protein (SREBP), is a weak transcription factor itself and has been shown to work in co-operation with SP1 and NF-Y [72, 134, 234, 241]. Therefore, we aimed at identifying novel putative SREBP target genes by integrating *in silico* predictions of SREBP target genes with gene expression profiling of cholesterol-depleted cells when SREBP is induced.

### 3.2.1 Prediction of new putative SREBP target genes by integrating binding site predictions and gene expression profiling

To determine SREBP target genes we integrated gene expression data of cholesterol depleted cells and *in silico* predictions of SREBP target genes. The workflow is shown in Figure 2.3. We conducted gene expression analysis of two different cell lines cultured under control and sterol-depleted conditions. In addition to genome-wide gene expression analysis in HeLa cells [22], we analyzed a primary human skin fibroblast cell line from a healthy individual. In total, we yielded 189 significantly differentially expressed genes, of which 73 were up- and 116 down-regulated (see Table B.1). Among the differentially expressed genes, 42 genes have been described previously to be functionally relevant for cholesterol biosynthesis and lipid homeostasis. Of these, 16 genes encode cholesterol biosynthetic enzymes (all genes necessary for cholesterol biosynthesis in human: 22 (see [22])). For example, the rate-limiting enzyme for cholesterol synthesis is 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMGCR) and it was up-regulated in both studies ($P < 2.2e\text{-}16$ for both studies), with a mean fold change of 4.3 in the HeLa-cell experiments and 3.5 in the fibroblast experiments. To refine this list and to select genes with SREBP binding sites we performed a genome-wide *in silico* promoter screen for all human genes with annotated transcription start sites (TSS). As SREBP1 and 2 typically cooperate with the transcriptional co-factors SP1 and NF-Y [72, 134, 234, 241], we also screened for binding motifs of SP1 and NF-Y in the promoter region of the predicted SREBP target genes. In addition, we considered LXR which is known to regulate cellular cholesterol and is involved in its efflux [234]. Of the 22 human genes necessary for cholesterol biosynthesis we identified SREBP binding sites for 10 genes, SP1 binding sites for 19 genes, NF-Y binding sites for 15 genes, and LXR binding sites for 14 genes.

The integration of differentially expressed genes upon cellular cholesterol depletion and *in silico* predicted SREBP target genes yielded a total of 99 genes which are shown in Table 3.4.

**Table 3.4. Putative SREBP target genes determined by gene expression analysis and binding site prediction.**

***A Genes up-regulated under sterol-depleted conditions***

| | | Fibroblasts | | HeLa cells | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Symbol | Name | pfp[1] | FC[2] | pfp[1] | FC[2] | SREBP[3] | SP1[3] | NF-Y[3] | LXR[3] |
| HES6 | hairy and enhancer of split 6 (Drosophila) | 0 | 3.7806 | | | x | x | x | |
| CPSF1 | cleavage and polyadenylation specific factor 1, 160kDa | | | 0 | 2.5922 | x | x | x | x |
| CCNG2 | cyclin G2 | | | 1.00E-04 | 2.3832 | x | x | x | |
| SLCO2A1 | solute carrier organic anion transporter family, member 2A1 | | | 3.00E-04 | 2.1154 | x | x | | x |
| BHLHE40 | basic helix-loop-helix family, member e40 | 3.00E-04 | 2.6160 | | | x | x | | |
| FLVCR1 | feline leukemia virus subgroup C cellular receptor 1 | | | 7.00E-04 | 2.0854 | x | x | x | |
| TMEM97 | transmembrane protein 97 | 0.0014 | 2.6274 | 0.0017 | 2.0197 | x | x | | x |
| HSD17B7 | hydroxysteroid (17-beta) dehydrogenase 7 | | | 0.0018 | 1.9727 | x | x | | x |
| INSIG1 | insulin induced gene 1 | 0.0019 | 2.8316 | 0 | 3.2167 | x | x | x | |
| MVD | mevalonate (diphospho) decarboxylase | 0.0023 | 2.3226 | | | x | x | x | x |
| MAFB | v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian) | 0.0024 | 2.3337 | | | x | x | x | |

[1] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

[2] Fold change in fibroblasts or HeLa-cells

[3] x in columns 3-6 indicates predicted binding sites for SREBP, NF-Y, Sp1, and LXR.

| | | Fibroblasts | | HeLa cells | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Symbol | Name | pfp[1] | FC[2] | pfp[1] | FC[2] | SREBP[3] | SP1[3] | NF-Y[3] | LXR[3] |
| STC1 | stanniocalcin 1 | 0.0028 | 2.3742 | | | x | x | x | |
| IDI1 | isopentenyl-diphosphate delta isomerase 1 | 0.0028 | 2.2995 | 0.0013 | 2.1098 | x | x | x | x |
| FABP3 | fatty acid binding protein 3, muscle and heart (mammary-derived growth inhibitor) | 0.003 | 2.3738 | | | x | x | x | x |
| KLF6 | Kruppel-like factor 6 | 0.0031 | 3.1659 | | | x | x | x | x |
| PDGFRB | platelet-derived growth factor receptor, beta polypeptide | 0.0035 | 2.3345 | | | x | x | x | x |
| SCD | stearoyl-CoA desaturase (delta-9-desaturase) | 0.0038 | 2.1175 | 0.0011 | 1.9835 | x | x | | x |
| LPIN1 | lipin 1 | 0.004 | 2.0214 | 0.049 | 1.7186 | x | x | x | x |
| PFKFB4 | 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4 | 0.004 | 2.0521 | | | x | x | | |
| DHCR7 | 7-dehydrocholesterol reductase | 0.0043 | 1.9957 | 0.0034 | 1.883 | x | x | x | x |
| RGS4 | regulator of G-protein signaling 4 | 0.0043 | 2.2216 | | | x | | x | |
| C17orf59 | chromosome 17 open reading frame 59 | 0.0045 | 2.0269 | | | x | x | | x |
| FDFT1 | farnesyl-diphosphate farnesyltransferase 1 | | | 0.0072 | 1.8615 | x | x | x | x |
| MXRA5 | matrix-remodelling associated 5 | 0.0086 | 2.1595 | | | x | x | | x |
| LDLR | low density lipoprotein receptor | 0.0089 | 1.8743 | | | x | x | | x |
| C20orf20 | chromosome 20 open reading frame 20 | | | 0.0095 | 1.0461 | x | x | | |
| GAS1 | growth arrest-specific 1 | 0.0096 | 2.0526 | | | x | x | | x |
| TP53INP2 | tumor protein p53 inducible nuclear protein 2 | 0.0101 | 1.8091 | | | x | x | x | x |
| RASD1 | RAS, dexamethasone-induced 1 | | | 0.0117 | 1.0021 | x | x | | |
| DBC1 | deleted in bladder cancer 1 | 0.0128 | 1.906 | | | x | x | | |

[1] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

[2] Fold change in fibroblasts or HeLa-cells

[3] x in columns 3-6 indicates predicted binding sites for SREBP, NF-Y, Sp1, and LXR.

| | | Fibroblasts | | HeLa cells | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Symbol | Name | pfp[1] | FC[2] | pfp[1] | FC[2] | SREBP[3] | SP1[3] | NF-Y[3] | LXR[3] |
| SLC2A6 | solute carrier family 2 (facilitated glucose transporter), member 6 | 0.0132 | 1.8013 | | | x | x | | |
| TMEM55B | transmembrane protein 55B | 0.0133 | 1.7783 | | | x | x | x | |
| TP53INP1 | tumor protein p53 inducible nuclear protein 1 | 0.015 | 1.7629 | | | x | x | x | x |
| KCNJ2 | potassium inwardly-rectifying channel, subfamily J, member 2 | 0.0151 | 1.8601 | | | x | x | x | |
| ANGPTL2 | angiopoietin-like 2 | 0.0151 | 1.7884 | | | x | x | x | x |
| HSD17B12 | hydroxysteroid (17-beta) dehydrogenase 12 | 0.0155 | 1.7475 | | | x | | x | x |
| PDGFRA | platelet-derived growth factor receptor, alpha polypeptide | 0.0155 | 1.8534 | | | x | x | | x |
| SQLE | squalene epoxidase | 0.0157 | 1.7449 | 0 | 2.4489 | x | x | x | x |
| FBLN1 | fibulin 1 | 0.0167 | 2.4985 | | | x | x | | x |
| PCYT2 | phosphate cytidylyltrans-ferase 2, ethanolamine | | | 0.017 | 1.1331 | x | x | | |
| C3orf54 | chromosome 3 open reading frame 54 | 0.0175 | 1.7694 | | | x | x | | x |
| KLF13 | Kruppel-like factor 13 | 0.0183 | 1.669 | | | x | x | | x |
| FASN | fatty acid synthase | 0.0185 | 1.7046 | 1.00E-04 | 1.8882 | x | x | | x |
| MNT | MAX binding protein | 0.0185 | 1.678 | | | x | x | x | x |
| TOB1 | transducer of ERBB2, 1 | 0.0218 | 1.7352 | | | x | x | x | x |
| MYO1D | myosin ID | 0.0224 | 1.7072 | | | x | x | | |
| EPR1 | effector cell peptidase receptor 1 (non-protein coding) | | | 0.0275 | 1.0634 | x | x | | x |
| BIRC5 | baculoviral IAP repeat-containing 5 | | | 0.0275 | 1.0634 | x | x | x | |
| ELOVL6 | ELOVL family member 6, elongation of long chain fatty acids (FEN1/Elo2, SUR4/Elo3-like, yeast) | 0.0306 | 1.6842 | | | x | x | | x |

[1] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells
[2] Fold change in fibroblasts or HeLa-cells
[3] x in columns 3-6 indicates predicted binding sites for SREBP, NF-Y, Sp1, and LXR.

| | | Fibroblasts | | HeLa cells | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Symbol | Name | pfp[1] | FC[2] | pfp[1] | FC[2] | SREBP[3] | SP1[3] | NF-Y[3] | LXR[3] |
| DNAJB9 | DnaJ (Hsp40) homolog, subfamily B, member 9 | 0.0309 | 1.6304 | | | x | | | |
| RORB | RAR-related orphan receptor B | | | 0.0325 | 1.6022 | x | x | x | |
| ZCCHC14 | zinc finger, CCHC domain containing 14 | 0.0349 | 1.6118 | | | x | x | x | |
| FAM189B | family with sequence similarity 189, member B | | | 0.0364 | 1.6427 | x | x | | |
| MYO10 | myosin X | 0.0377 | 1.5884 | | x | x | | x | |
| SLIT3 | slit homolog 3 (Drosophila) | 0.038 | 1.6017 | | | x | | x | x |
| SAT1 | spermidine/spermine N1-acetyltransferase 1 | 0.0381 | 1.4897 | | | x | x | x | x |
| FRMD8 | FERM domain containing 8 | 0.0392 | 1.6735 | | | x | x | x | |
| MVK | mevalonate kinase | | | 0.0399 | 1.6572 | x | | x | |
| ZC3H12A | zinc finger CCCH-type containing 12A | 0.0415 | 1.5342 | | | x | x | | x |
| SLC26A6 | solute carrier family 26, member 6 | 0.0415 | 1.5668 | | | x | x | x | x |
| IER5L | immediate early response 5-like | 0.0419 | 1.5482 | | | x | x | x | |
| SLC2A3P1 | solute carrier family 2 (facilitated glucose transporter), member 3 pseudogene 1 | | | 0.0426 | 1.6067 | x | x | | |
| DDIT4 | DNA-damage-inducible transcript 4 | 0.0444 | 1.1515 | | | x | x | x | x |
| HCFC1R1 | host cell factor C1 regulator 1 (XPO1 dependent) | 0.0458 | 1.5468 | | | x | x | | |
| GP1BB | glycoprotein Ib (platelet), beta polypeptide | 0.0462 | 1.5687 | | | x | x | x | |
| SNAI1 | snail homolog 1 (Drosophila) | 0.0474 | 1.5163 | | | x | x | | x |
| MSX1 | msh homeobox 1 | 0.0485 | 1.4492 | | | x | x | x | |
| CXXC5 | CXXC finger protein 5 | 0.0495 | 1.5027 | | | x | x | x | |

[1] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

[2] Fold change in fibroblasts or HeLa-cells

[3] x in columns 3-6 indicates predicted binding sites for SREBP, NF-Y, Sp1, and LXR.

### *B Genes down-regulated under sterol-depleted conditions*

| | | Fibroblasts | | HeLa cells | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Symbol | Name | pfp[1] | FC[2] | pfp[1] | FC[2] | SREBP[3] | SP1[3] | NF-Y[3] | LXR[3] |
| CDC20 | cell division cycle 20 homolog (S. cerevisiae) | 0 | 0.3398 | | | x | x | | |
| LOC285556 | hypothetical protein LOC285556 | | | 0 | 0.6027 | x | | | |
| PTGES3 | prostaglandin E synthase 3 (cytosolic) | | | 5.00E-04 | 0.9244 | x | x | | x |
| CXCR2 | chemokine (C-X-C motif) receptor 2 | | | 0.0117 | 0.6894 | x | x | x | |
| ID2 | inhibitor of DNA binding 2, dominant negative helix-loop-helix protein | 0.0142 | 0.4843 | 0 | 0.463 | x | x | | |
| MAML2 | mastermind-like 2 (Drosophila) | | | 0.0157 | 0.5794 | x | x | | |
| ACTC1 | actin, alpha, cardiac muscle 1 | 0.0167 | 0.4843 | | | x | x | | |
| F3 | coagulation factor III (thromboplastin, tissue factor) | 0.0192 | 0.4966 | | | x | x | x | |
| SPOCD1 | SPOC domain containing 1 | 0.0194 | 0.4696 | | | x | | | x |
| EDN2 | endothelin 2 | | | 0.0195 | 0.5874 | x | x | | |
| S100P | S100 calcium binding protein P | 0.0207 | 0.4455 | | | x | x | | |
| CRIP1 | cysteine-rich protein 1 (intestinal) | 0.0208 | 0.5225 | | | x | x | | |
| TSC22D2 | TSC22 domain family, member 2 | 0.0209 | 0.5632 | | | x | x | x | |
| NTF3 | neurotrophin 3 | 0.0218 | 0.5795 | | | x | x | | x |
| TRIP13 | thyroid hormone receptor interactor 13 | 0.0226 | 0.5103 | | | x | x | x | x |
| KLF2 | Kruppel-like factor 2 (lung) | 0.0228 | 0.5312 | | | x | x | | x |
| IQGAP3 | IQ motif containing GTPase activating protein 3 | 0.0236 | 0.531 | | | x | | | x |
| AXL | AXL receptor tyrosine kinase | 0.0239 | 0.5541 | | | x | x | x | |
| PRC1 | protein regulator of cytokinesis 1 | 0.0242 | 0.5529 | | | x | x | | |
| PMP22 | peripheral myelin protein 22 | 0.0274 | 0.4008 | | | x | | x | |
| GBP3 | guanylate binding protein 3 | | | 0.0276 | 0.5946 | x | | x | |
| CDC42EP3 | CDC42 effector protein (Rho GTPase binding) 3 | 0.0303 | 0.614 | 0.0457 | 0.5798 | x | x | | |

[1] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

[2] Fold change in fibroblasts or HeLa-cells

[3] x in columns 3-6 indicates predicted binding sites for SREBP, NF-Y, Sp1, and LXR.

| | | Fibroblasts | | HeLa cells | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Symbol | Name | pfp[1] | FC[2] | pfp[1] | FC[2] | SREBP[3] | SP1[3] | NF-Y[3] | LXR[3] |
| ASPM | asp (abnormal spindle) homolog, microcephaly associated (Drosophila) | 0.0329 | 0.6177 | | | x | | x | |
| CEP55 | centrosomal protein 55kDa | 0.0334 | 0.6256 | | | x | x | x | x |
| DUSP1 | dual specificity phosphatase 1 | 0.0336 | 0.5624 | | | x | x | | x |
| OXTR | oxytocin receptor | 0.0336 | 0.6168 | | | x | | | x |
| TIPARP | TCDD-inducible poly(ADP-ribose) polymerase | 0.034 | 0.5414 | | | x | x | | |
| CA12 | carbonic anhydrase XII | 0.0345 | 0.6156 | | | x | x | x | x |
| RPL29 | ribosomal protein L29 | 0.0428 | 0.6387 | | | x | x | | x |
| UBXN1 | UBX domain protein 1 | | | 0.0442 | 0.7756 | x | x | | x |
| GADD45B | growth arrest and DNA-damage-inducible, beta | 0.0465 | 0.6555 | | | x | x | | x |

[1] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells
[2] Fold change in fibroblasts or HeLa-cells
[3] x in columns 3-6 indicates predicted binding sites for SREBP, NF-Y, Sp1, and LXR.

An enrichment analysis for Gene Ontology terms identified mainly cholesterol and lipid related biological processes (Table 3.6). 21 of the identified 99 genes have been previously associated to cholesterol or fatty acid metabolism and genes encoding cholesterol biosynthetic enzymes were enriched among the identified SREBP target genes when compared to all genes for cholesterol biosynthesis (P = 7.542e-11), yielding 78 novel putative SREBP target genes. Many of the novel identified SREBP target genes are involved in cellular processes such as cell proliferation (*pdgfra*, *pdgfrb*, *gas1*,*mnt*, *tob1*, *slit3*, *msx1*, *cdc20*, *cxrcr2*, *id2*, *edn2*, *pmp22*) and cell cycle (*ccng2*, *gas1*, *dbc1*, *tp53inp1*, *mnt*, *birc5*, *prc1*, *cdc20*, *id2*, *ppp1r14d*, *cep55*, *gadd45b*), are involved in the regulation of transcription or are transcription factors themselves (*mafb*, *klf6*, *klf13*, *msx1*, *maml2*, *ntf3*).

### 3.2.2 Comparison of identified target genes to sequencing results from chromatin immunoprecipitations (ChIP-Seq)

A comparison with ChIP-Seq data from the ENCODE project [65, 235] for SREBP1a and SREBP2 by the lab of Michael Snyder at Yale University and NF-YA and NF-YB by the lab of Kevin Struhl at Harvard showed a high

**Table 3.6. Identified Gene Ontology terms for the 99 identified putative SREBP target genes.**

| GO ID | GO Term | Annotated | Significant | Expected | P-value |
|---|---|---|---|---|---|
| GO:0008610 | lipid biosynthetic process | 188 | 15 | 2.43 | 8.75E-05 |
| GO:0016126 | sterol biosynthetic process | 17 | 6 | 0.22 | 0.0002 |
| GO:0006694 | steroid biosynthetic process | 47 | 8 | 0.61 | 0.0002 |
| GO:0008203 | cholesterol metabolic process | 47 | 8 | 0.61 | 0.0002 |
| GO:0016125 | sterol metabolic process | 49 | 8 | 0.63 | 0.0002 |
| GO:0006695 | cholesterol biosynthetic process | 14 | 5 | 0.18 | 0.0007 |
| GO:0008202 | steroid metabolic process | 107 | 10 | 1.38 | 0.0010 |
| GO:0008299 | isoprenoid biosynthetic process | 8 | 4 | 0.1 | 0.0015 |
| GO:0006629 | lipid metabolic process | 425 | 18 | 5.49 | 0.0047 |
| GO:0048008 | platelet-derived growth factor receptor signaling pathway | 12 | 4 | 0.16 | 0.0087 |
| GO:0008283 | cell proliferation | 575 | 20 | 7.43 | 0.0205 |
| GO:0042127 | regulation of cell proliferation | 414 | 16 | 5.35 | 0.0389 |

overlap with our predicted SREBP target genes. Using the same restrictions as for our *in silico* analysis, the ENCODE data set comprised 34 genes of our identified putative SREBP target genes and 55 NF-Y target genes. Table B.3 indicates all predicted binding sites for the 99 identified SREBP target genes as well as identified binding sites for SREBP and NF-Y of the ENCODE data set. Several of the predicted SREBP and NF-Y binding sites showed a high overlap with the identified ENCODE binding sites.

To analyze the conservation of binding sites of the predicted SREBP target genes, we employed pair-wise alignments between human, mouse and chimp, respectively. Interestingly, most of the predicted binding sites were conserved between human, mouse and chimp. 96% of the predicted SREBP target sites were conserved between human and chimp (compared to 94% of binding sites of all TFs), whereas 68% of the SREBP target sites were conserved between human and mouse (see Table B.4). In comparison, significantly less binding sites of all transcription factors (58%) were conserved between human and mouse in the upstream regions of the predicted SREBP target genes (P = 0.0151).

### 3.2.3 Experimental Validation of predicted SREBP target genes

For validation of our results we selected ten out of the 78 putative new SREBP target genes which were marginally described in association with cholesterol in the literature and mainly up-regulated at cholesterol depletion and performed qRT-PCR experiments. First, we analyzed the expression levels of *c17orf59*, *cpsf1*, *gbp3*, *hes6*, *klf6*, *klf13*, *mafb*, *slc2a6*, *tmem55b*, and *tob1* in human fibroblasts and HeLa cells at control and sterol-depleted growth conditions. The LDL receptor was used as a positive control and was significantly up-regulated under sterol-depleted conditions. Besides this, also *c17orf59*, *hes6*, *slc2a6*, *tmem55b*, and *tob1* were up-regulated at sterol-depleted growing conditions in both cell lines. *klf6*, *mafb*, and *klf13* were only up-regulated in human fibroblasts whereas *gbp3* showed an up-regulation in HeLa cells. Only for one (*cpsf1*) out of the ten genes chosen for validation, we were not able to see any up-regulating effect under sterol-depleted growth conditions (see Table 3.7). *c17orf59*, *slc2a6*, *tmem55b*, *hes6*, and

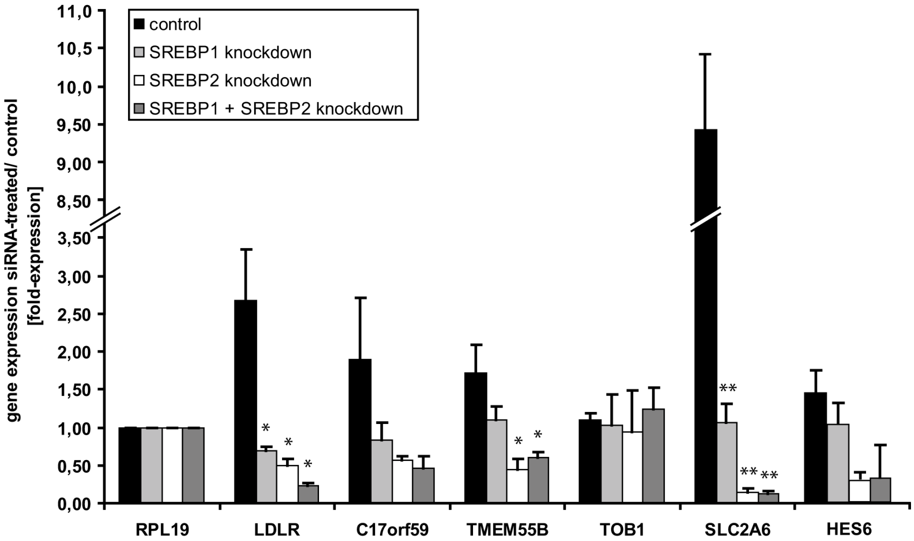


**Figure 3.9. Gene expression in sterol-depleted HeLa cells upon siRNA knockdown of SREBP**. In addition to the five candidate SREBP target genes, RPL19 and LDLR were used as negative and positive control, respectively. Significance of the results after SREBP knockdown are indicated by * ($P < 0.05$) and *** ($P < 0.001$). Knockdown experiments were conducted by Jessica Schilde.

**Table 3.7. Gene expression levels of ten putative new SREBP target genes under sterol-depleted growth conditions in human fibroblasts and HeLa cells.**

| | | Fibroblasts | | | HeLa cells | | |
|---|---|---|---|---|---|---|---|
| Symbol | Name | ratio* | SE+ | p-value# | ratio* | SE+ | p-value# |
| LDLR | low density lipoprotein receptor | 5.9 | 1.01 | *** | 3.34 | 1.63 | * |
| SLC2A6 | solute carrier family 2 (facilitated glucose transporter), member 6 | 4.50 | 2.36 | | 3.73 | 1.63 | |
| TMEM55B | transmembrane protein 55 B | 3.43 | 1.00 | | 2.41 | 1.39 | |
| TOB1 | transducer of ERBB2, 1 | 3.36 | 2.07 | | 2.31 | 0.98 | |
| C17orf59 | chromosome 17 open reading frame 59 | 1.86 | 0.21 | * | 1.68 | 0.49 | |
| HES6 | hairy and enhancer of split 6 (Drosophila) | 1.73 | 0.49 | | 1.82 | 1.38 | |
| KLF6 | Kruppel-like factor 6 | 4.77 | 1.20 | * | 1.65 | 1.93 | |
| MAFB | v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian) | 4.27 | 0.94 | * | 0.78 | 0.69 | |
| KFL13 | Kruppel-like factor 13 | 2.66 | 0.83 | | 0.98 | 0.32 | |
| GBP3 | guanylate binding protein 3 | 1.18 | 0.16 | | 2.99 | 2.75 | |
| CPSF1 | cleavage and polyadenylation specific factor 1, 160kDa | 0.88 | 0.26 | | 1.03 | 0.17 | |

* ratio between Ct (cycle threshold) values with regard to RPL19
+ standard error
# * p-value $< 0.05$, *** p-value $< 0.001$.

*tob1* (up-regulated in both cell systems) were selected for further analysis. To find out if the expression of these genes was dependent on SREBP, we analyzed HeLa cells in sterol-depleted conditions and treated them with interference constructs knocking down SREBP1 and SREBP2, each respectively, and mutually. This was compared to a mock treatment (control) and the expression of the putative target was quantified (using qRT-PCR). The results are shown in Figure 3.9. Four out of these five putative new SREBP target genes showed a much lower mRNA expression level in sterol depleted cells after SREBP knockdown compared to the control evidencing their regulatory dependency on SREBP. *c17orf57* and *slc2a6* showed a reduced expression level after SREBP1 as well as SREBP2 knockdown. In turn, *tmem55b* and *hes6* expression were only slightly reduced by SREBP1 knockdown but significantly reduced when SREBP2 was knocked down indicating *tmem55b* and *hes6* having been controlled solely by SREBP2 and *c17orf59* and *slc2a6* having been regulated by both transcription factors. This evidenced *tmem55b*, *c17orf59*, *hes6*, and *slc2a6* to be novel potential SREBP target genes, identified by the integration of transcription factor binding site predictions and gene expression profiles from cells in conditional cholesterol media.

# Chapter 4

# Discussion

## 4.1 Identification of spatio-temporal specific regulatory modules

It was suggested previously that combinations of different complexes of transcription factors offer a plethora of specific gene expression profiles [164]. Here, we identified *in silico* regulatory modules consisting of combinations of transcription factors binding at promoters and enhancers that determine specific tissue dependent and temporal regulation of gene expression during mouse embryonic development and human stem cell differentiation. In addition to tissue-specific regulation established by pairs of transcription factors as shown previously [229], we now also demonstrated that regulatory modules consisting of pairs of transcription factors at promoters and enhancers regulate the progression of gene expression patterns during development and differentiation. Furthermore, we found that these enhancer sites were rather depleted of Guanin-Cytosin (CpG). It was shown recently, that methylation-modifications of CpG-regions are a major regulation mechanism during development [36, 159, 207]. Enhancer regions therefore may contribute to a more constitutive regulation program during development which is rather independent from these methylation-modifications.

The identification of *in silico* transcription factor binding sites using PWMs can lead to a large number of false positives with no information about actual functional binding [284]. We addressed this issue by using clustering of binding sites as well as including conservation of binding sites and dismissed binding sites occurring only in a small number of genes. Additionally, we incorporated gene expression data into our analysis to identify transcription factors, combination of transcription factors, and regulatory modules that are “active” in a specific tissue at a specific time interval.

Analyzing the identified regulatory modules revealed a number of transcription factors binding preferentially either at promoters or enhancers. For human stem cell differentiation, we identified SP1 to preferentially bind at promoters. Although SP1 is ubiquitously expressed and regulates gene expression of many constitutively expressed genes [97, 129], its expression was shown to change at different developmental stages and in different cell types, suggesting specific roles in distinct developmental processes [238]. As SP1-null mice died prenatal, SP1 was shown to be essential for mouse embryonic development [177].

In contrast, members of the FOX (forkhead box) family of transcription factors had binding sites preferentially located at enhancers of the identified regulatory modules providing temporal specificity during human stem cell differentiation. FOX transcription factors have been identified to bind at enhancers in a number of studies [58, 63, 71, 105, 186, 303]. FOX proteins are substantial in a variety of cellular processes including development, differentiation, proliferation, apoptosis, and migration [197]. As FOX proteins are regulators for a multitude of biological processes, their deregulation can contribute significantly to tumorigenesis and cancer progression [197]. Various members of this family have been identified previously to be implicated in development [9, 30, 42, 60, 89, 125, 152, 155, 157, 217, 275, 289, 293].

The CDX family was another group of transcription factors we identified at enhancers of our regulatory modules. *cdx* genes are closely related to the Hox cluster and are expressed during embryonic development and gut morphogenesis [24]. CDX2 is specifically required during early development and *cdx2*-null mice are nonviable as the blastocyst fails to implant into the uterus [266].

Chromatin loops can overcome large distances between long-range enhancers and proximal promoters and may lead to false positive hits at promoter regions when screening promoters with ChIP-chip assays [118, 196]. So far, most ChIP sequencing studies neglect to assess indirect binding of transcription factors. However, it was shown for HNF4A (hepatocyte nuclear factor 4, alpha) that identified binding sites at the promoter occurred mainly at distal regulatory elements [226]. Our results support the fact that key transcription factors bind preferentially at either promoters or enhancers and that the interactions between those elements are crucial for specific gene regulation. Therefore, indirect binding is a central issue that cannot be neglected in prospective transcription factor binding site analyses and specifically when analyzing ChIP-chip and ChIP-Seq experiments.

## 4.2 Identification of novel putative SREBP target genes

Cholesterol biosynthesis in mammals is one example of combinatorial transcriptional regulation. Prediction of SREBP binding sites and its most common co-factors SP1 and NF-Y allowed the identification of novel potential target genes of SREBP and gained new insights into the regulation of cholesterol biosynthesis. Candidate genes were identified by genome-wide gene expression analyses of sterol-depleted cultured cells in two very distinct human cell lines. To further minimize the number of false positives, we selected genes with *in silico* predicted binding sites for SREBP and its most commonly co-occurring transcription factors SP1, NF-Y, and LXR. Genes encoding cholesterol biosynthetic enzymes were enriched among the identified genes indicating that our integrated approach suited well to identify also novel SREBP target genes. In addition, comparisons to binding sites of SREBP and NF-Y identified by ChIP-Seq experiments revealed a high overlap to our *in silico* predictions. We identified a number of genes that may constitute SREBP target genes and play substantial roles in cholesterol metabolism and specifically, cholesterol's control of central cellular decisions. We identified 78 putative SREBP target genes which have not previously been linked to processes in which cholesterol is involved. Of these, we selected ten genes for experimental validation. For most of these genes, we validated their upregulation in response to cholesterol-depletion. Four of these genes (*slc2a6*, *tmem55b*, *hes6*, and *c17orf59*) showed distinct down-regulation in response to SREBP knockdown indicating their regulatory control by SREBP.

Of the validated genes, *slc2a6* showed the highest effect of both SREBP1 and SREBP2 knockdowns. It functions as a sugar-transport facilitator with glucose-transporter activity, though its specific function and stimulus has not yet been identified [276]. Glucose supports absorption and transport of cellular cholesterol [230] and SREBP1c has been shown to be sensitive to high levels of carbohydrates such as glucose [119]. Ravid and coworkers showed that high extracellular glucose concentration had a positive effect on the regulation of cholesterol transport [230]. This treatment caused also an increase in protein expression of NPC1L1 and CD36 which are involved in cholesterol uptake [230]. It is to note that NPC1L1 is similar to NPC1 which loss-of-function causes the NPC disease 1 and is similarly involved in cholesterol uptake and trafficking [7, 69]. Interestingly, in this study, protein levels of SREBP2 were decreased whereas LXR levels increased due to high glucose concentrations [230]. Taken Ravid et al's, Horten et al's, and our findings together comparing SREBP1 and SREBP2, we speculate that two distinct mechanisms of cholesterol uptake are regulated by SREBP2 and SREBP1, respectively, enhancing cholesterol/glucose co-transport (rather SREBP1 regulated) and direct uptake via the LDL receptor (rather SREBP2 regulated). In any way, SLC2A6 may use the uptake of glucose to support uptake of cholesterol into the cell.

Chromosome 17 open reading frame 59 (C17orf59) has been previously identified as a potential regulator of cholesterol [22] and demonstrated a decrease in expression after SREBP knockdown. It is an uncharacterized protein coding gene that requires further functional analysis.

*tmem55b* also showed a decrease in expression after SREBP knockdown (predominantly after knockdown of SREBP2). It catalyzes the hydrolysis of phosphatidylinositol 4,5-bisphosphate (PtdIns-4,5-P2) to phosphatidylinositol 5-phosphate (PtdIns-5-P) and appears to affect the lysosomal degradation of internalized plasma membrane receptors [277]. PtdIns-4,5-P2 is involved in the regulation of signal transduction, exocytosis/endocytosis, actin dynamics, and ion channel and transport function [212, 268]. After cholesterol depletion, PtdIns-4,5-P2 levels were decreased in the plasma membrane [150] and the organization of PtdIns-4,5-P2 in the plasma membrane were disrupted [219]. It is reasonable that in physiological conditions, high cholesterol levels are accomplished with high internalization of LDL which needs to be lysosomal degraded and a regulation of this feedback mechanism is mediated by SREBP. Interestingly, increasing levels of PtdIns-5-P have been shown to mediate p53-dependent apoptosis [313]. In this context, cholesterol depletion causes SREBP induction, which causes PtdIns-5-P and p53 mediated apoptosis. We found further links of our detected SREBP target genes to the cellular decision of apoptosis/cell cycle arrest and proliferation. Specifically, HES6, GBP3 and TOB1 showed up-regulation in cells in cholesterol depleted medium and all have been reported for distinctively anti-proliferative effects [79, 108, 171, 180, 237]. Reduced cholesterol may result in a stress response that is mediated by SREBP diminishing proliferation and inducing apoptosis.

Our integrated approach has revealed new potential SREBP target genes being functionally relevant to cellular regulation mediated by cholesterol to control metabolism and signaling in health and disease.

## 4.3 Outlook

Time- and tissue-specific regulation of gene expression is central not only during development but in all processes of a cell in an organism. Here, we identified regulatory modules that determined specific gene regulation during development and revealed distinct binding site distributions for transcription factors binding preferentially at promoter or enhancer regions. The *in silico* identification of combinations of transcription factors binding at promoters and enhancers yielded generic insights into the temporal regulation of gene expression and improved our understanding of enhancer function.

The identified interactions of transcription factors identified in the regulatory modules require now experimental validation in a time- and tissue-specific manner. Knockdown experiments in embryonic mice can demonstrate the importance of a distinct transcription factor in a specific tissue at a specific time point. Another option is the use of fluorescent dyes (e.g. by fluorescence *in situ* hybridization) to demonstrate the activity of a distinct transcription factor at a specific time point in a specific tissue.

Although we applied stringent parameters we were able to identify a number of interesting regulatory modules. These regulatory modules were defined to consist of four transcription factors in total, with two transcription factors each binding either at a promoter region or an enhancer region, respectively. It remains to analyze the interactions between the transcription factors binding at the promoter and the transcription factor binding at the enhancer region in more detail and thereby uncover additional co-factors involved in the regulation of development and differentiation.

Another interesting aspect is the hypothesis that tumorigenesis resembles embryonic development and many key regulators in development have been shown to be implicated in tumor progression, see e.g. [70, 95, 161, 221]. It is therefore intriguing to apply the identified regulatory modules to tumor samples to uncover key regulators in tumorigenesis that will improve our understanding of tumor progression.

In addition, it has become clearer that histone modifications play an important role in transcriptional regulation [115]. The combination of the predicted regulatory modules together with histone marker maps at promoters and enhancers might elucidate further insights into the regulatory impact of both mechanisms.

We employed combinatorial promoter analyses of SREBP and known co-factors, used gene expression data, and identified new putative SREBP target genes. Further analyses may reveal the exact mechanism by which the identified SREBP target genes are involved in cholesterol homeostasis. Their detailed mechanism may improve our understanding of cholesterol related diseases, such as Familal Hypercholesterolemia and Niemann-Pick Disease Type C, when cholesterol uptake is impaired. Further experiments are currently conducted to elucidate the function of C17orf59 and TMEM55B.

Although we identified putative SREBP target genes that have differential regulation upon sterol-depletion and were able to validate a small number of genes in functional experiments, a direct influence of SREBP remains yet to be shown. Reporter assays using the upstream sequence of the gene under consideration upon induction of SREBP might help excluding any indirect effects of SREBP.

Interestingly, SREBP binding sites were not predicted for all genes known to be involved in cholesterol biosynthesis. However, these predictions can only be as good as the existing known binding motifs and SREBP might bind to additional, yet unknown binding motifs. Scanning upstream sequences of the identified putative SREBP target genes for over presented short sequences might identify further binding motifs and transcription factors involved in the regulation of cholesterol biosynthesis.

# References

[1] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[2] R. Albert, H. Jeong, and A. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, Jul 2000.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 4 edition, 2002.

[4] A. Alexa, J. Rahnenführer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–1607, Jul 2006.

[5] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55–65, Jan 2006.

[6] E. Almaas. Biological impacts and context of network theory. *J Exp Biol*, 210(Pt 9):1548–1558, May 2007.

[7] S. W. Altmann, H. R. Davis, L.-J. Zhu, X. Yao, L. M. Hoos, et al. Niemann-pick c1 like 1 protein is critical for intestinal cholesterol absorption. *Science*, 303(5661):1201–1204, Feb 2004.

[8] L. Anderson and J. Seilhamer. A comparison of selected mrna and protein abundances in human liver. *Electrophoresis*, 18(3-4):533–537, 1997.

[9] S. L. Ang and J. Rossant. Hnf-3 beta is essential for node and notochord formation in mouse development. *Cell*, 78(4):561–74, 1994.

[10] J. Archambault and J. D. Friesen. Genetics of eukaryotic rna polymerases i, ii, and iii. *Microbiol Rev*, 57(3):703–724, Sep 1993.

[11] N. J. Armstrong and M. A. van de Wiel. Microarray data analysis: from hypotheses to conclusions using gene expression data. *Cell Oncol*, 26(5-6):279–290, 2004.

[12] M. I. Arnone and E. H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–1864, May 1997.

[13] D. N. Arnosti. Analysis and function of transcriptional regulatory elements: insights from drosophila. *Annu Rev Entomol*, 48:579–602, 2003.

[14] D. N. Arnosti and M. M. Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem*, 94(5):890–898, Apr 2005.

[15] M. L. Atchison. Enhancers: mechanisms of action and cell specificity. *Annu Rev Cell Biol*, 4:127–153, 1988.

[16] G. D. Bader, D. Betel, and C. W. Hogue. Bind: the biomolecular interaction network database. *Nucleic Acids Res*, 31(1):248–50, 2003.

[17] L. Bai and A. V. Morozov. Gene regulation by nucleosome positioning. *Trends Genet*, 26(11):476–483, Nov 2010.

[18] N. E. Baker. Master regulatory genes; telling them what to do. *Bioessays*, 23(9):763–766, Sep 2001.

[19] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct 1999.

[20] A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, Feb 2004.

[21] A. Barski and K. Zhao. Genomic location analysis by chip-seq. *J Cell Biochem*, 107(1):11–18, May 2009.

[22] F. Bartz, L. Kern, D. Erz, M. Zhu, D. Gilbert, et al. Identification of cholesterol-regulating genes by targeted rnai screening. *Cell Metab*, 10(1):63–75, Jul 2009.

[23] A. D. Basehoar, S. J. Zanton, and B. F. Pugh. Identification and distinct regulation of yeast tata box-containing genes. *Cell*, 116(5):699–709, Mar 2004.

[24] F. Beck and E. J. Stringer. The role of cdx genes in the gut and in axial development. *Biochem Soc Trans*, 38(2):353–7, 2010.

[25] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, et al. Ultraconserved elements in the human genome. *Science*, 304(5675):1321–5, 2004.

[26] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[27] P. V. Benos, A. S. Lapedes, and G. D. Stormo. Is there a code for protein-DNA recognition? Probab(ilistical)ly... *Bioessays*, 24(5):466–475, May 2002.

[28] D. Bentley. Coupling rna polymerase ii transcription with pre-mrna processing. *Curr Opin Cell Biol*, 11(3):347–351, Jun 1999.

[29] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, et al. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proc Natl Acad Sci U S A*, 99(2):757–762, Jan 2002.

[30] V. Besnard, S. E. Wert, K. H. Kaestner, and J. A. Whitsett. Stage-specific regulation of respiratory epithelial cell differentiation by foxa1. *Am J Physiol Lung Cell Mol Physiol*, 289(5):L750–9, 2005.

[31] M. Bieda, X. Xu, M. A. Singer, R. Green, and P. J. Farnham. Unbiased location analysis of e2f1-binding sites suggests a widespread role for e2f1 in the human genome. *Genome Res*, 16(5):595–605, May 2006.

[32] E. M. Blackwood and J. T. Kadonaga. Going the distance: a current view of enhancer action. *Science*, 281(5373):60–63, Jul 1998.

[33] M. Blanchette, A. R. Bataille, X. Chen, C. Poitras, J. Laganiere, et al. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res*, 16(5):656–68, 2006.

[34] D. Boffelli, M. A. Nobrega, and E. M. Rubin. Comparative genomics at the vertebrate extremes. *Nat Rev Genet*, 5(6):456–465, Jun 2004.

[35] C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60. Rome, 1935.

[36] J. Borgel, S. Guibert, Y. Li, H. Chiba, D. Schubeler, et al. Targets and dynamics of promoter dna methylation during early mouse development. *Nat Genet*, 42(12):1093–100, 2010.

[37] L. Breiman. Bagging predictors. *Mach. Learn.*, 24:123–140, August 1996.

[38] L. Breiman. Technical note: Some properties of splitting criteria. *Machine Learning*, 24(1):41–47–47, 1996-07-01.

[39] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[40] L. Breiman, J. Friedman, R. Olshen, and P. Stone. *Classification and Regression Trees.* Chapman and Hall/CRC, 1984.

[41] R. Breitling and P. Herzyk. Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. *J Bioinform Comput Biol*, 3(5):1171–89, 2005.

[42] S. L. Brody, X. H. Yan, M. K. Wuerffel, S. K. Song, and S. D. Shapiro. Ciliogenesis and left-right axis defects in forkhead factor hfh-4-null mice. *Am J Respir Cell Mol Biol*, 23(1):45–51, 2000.

[43] M. S. Brown and J. L. Goldstein. A receptor-mediated pathway for cholesterol homeostasis. *Science*, 232(4746):34–47, Apr 1986.

[44] M. S. Brown and J. L. Goldstein. The srebp pathway: regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor. *Cell*, 89(3):331–340, May 1997.

[45] M. L. Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biol*, 5(1):201, 2003.

[46] M. L. Bulyk, P. L. F. Johnson, and G. M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–1261, Mar 2002.

[47] A. Butte. The use and analysis of microarray data. *Nat Rev Drug Discov*, 1(12):951–960, Dec 2002.

[48] B. R. Cairns. The logic of chromatin architecture and remodelling at promoters. *Nature*, 461(7261):193–198, Sep 2009.

[49] V. C. Calhoun, A. Stathopoulos, and M. Levine. Promoter-proximal tethering elements regulate enhancer-promoter specificity in the drosophila antennapedia complex. *Proc Natl Acad Sci U S A*, 99(14):9243–7, 2002.

[50] J. Carcamo, L. Buckbinder, and D. Reinberg. The initiator directs the assembly of a transcription factor iid-dependent transcription complex. *Proc Natl Acad Sci U S A*, 88(18):8052–8056, Sep 1991.

[51] M. Carey. The enhanceosome and transcriptional synergy. *Cell*, 92(1):5–8, Jan 1998.

[52] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, et al. The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563, Sep 2005.

[53] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–635, Jun 2006.

[54] J. S. Carroll, X. S. Liu, A. S. Brodsky, W. Li, C. A. Meyer, et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein foxa1. *Cell*, 122(1):33–43, Jul 2005.

[55] S. Cawley, S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell*, 116(4):499–509, Feb 2004.

[56] J. E. Celis, M. Kruhøffer, I. Gromova, C. Frederiksen, M. Ostergaard, et al. Gene expression profiling: monitoring transcription and translation products using dna microarrays and proteomics. *FEBS Lett*, 480(1):2–16, Aug 2000.

[57] S. Cereghini, S. Saragosti, M. Yaniv, and D. H. Hamer. Sv40-alpha-globulin hybrid minichromosomes. differences in dnase i hypersensitivity of promoter and enhancer sequences. *Eur J Biochem*, 144(3):545–553, Nov 1984.

[58] D. Chaya, T. Hayamizu, M. Bustin, and K. S. Zaret. Transcription factor foxa (hnf3) on a nucleosome at an enhancer complex in liver chromatin. *J Biol Chem*, 276(48):44385–9, 2001.

[59] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker. Analysis of microarray data using z score transformation. *J Mol Diagn*, 5(2):73–81, May 2003.

[60] J. Chen, H. J. Knowles, J. L. Hebert, and B. P. Hackett. Mutation of the mouse hepatocyte nuclear factor/forkhead homologue 4 gene results in an absence of cilia and random left-right asymmetry. *J Clin Invest*, 102(6):1077–82, 1998.

[61] Y. Chen, E. R. Dougherty, and M. L. Bittner. Ratio-based decisions and the quantitative analysis of cdna microarray images. *Journal Of Biomedical Optics*, 2:364–374, 1997.

[62] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3:140, 2007.

[63] L. A. Cirillo, F. R. Lin, I. Cuesta, D. Friedman, M. Jarnik, et al. Opening of compacted chromatin by early developmental transcription factors hnf3 (foxa) and gata-4. *Mol Cell*, 9(2):279–89, 2002.

[64] R. Cohen and S. Havlin. Scale-free networks are ultrasmall. *Phys. Rev. Lett.*, 90(5):058701, Feb 2003.

[65] E. N. C. O. D. E. P. Consortium, E. Birney, J. A. Stamatoyannopoulos, et al. Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816, Jun 2007.

[66] I. H. G. S. Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004.

[67] D. Das, Z. Nahlé, and M. Q. Zhang. Adaptively inferring human transcriptional subnetworks. *Mol Syst Biol*, 2:2006.0029, 2006.

[68] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Calestani, et al. A genomic regulatory network for development. *Science*, 295(5560):1669–78, 2002.

[69] J. P. Davies, B. Levy, and Y. A. Ioannou. Evidence for a niemann-pick c (npc) gene family: identification and characterization of npc1l1. *Genomics*, 65(2):137–145, Apr 2000.

[70] N. P. de Castro, M. C. Rangel, T. Nagaoka, D. S. Salomon, and C. Bianco. Cripto-1: an embryonic gene that promotes tumorigenesis. *Future Oncol*, 6(7):1127–1142, Jul 2010.

[71] S. De Val, N. C. Chi, S. M. Meadows, S. Minovitsky, J. P. Anderson, et al. Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. *Cell*, 135(6):1053–64, 2008.

[72] K. A. Dooley, S. Millinder, and T. F. Osborne. Sterol regulation of 3-hydroxy-3-methylglutaryl-coenzyme a synthase gene through a direct interaction between sterol regulatory element binding protein and the trimeric ccaat-binding factor/nuclear factor y. *J Biol Chem*, 273(3):1349–1356, Jan 1998.

[73] P. Du, W. A. Kibbe, and S. M. Lin. lumi: a pipeline for processing illumina microarray. *Bioinformatics*, May 2008.

[74] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[75] B. Eickhoff, B. Korn, M. Schick, A. Poustka, and J. van der Bosch. Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Res*, 27(22):e33, Nov 1999.

[76] C. T. Ekstrøm, S. Bak, C. Kristensen, and M. Rudemo. Spot shape modelling and data transformations for microarrays. *Bioinformatics*, 20(14):2270–2278, Sep 2004.

[77] L. Elnitski, V. X. Jin, P. J. Farnham, and S. J. M. Jones. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res*, 16(12):1455–1464, Dec 2006.

[78] P. Erdös and A. Rényi. On the evolution of random graphs. *Publ Math Inst Hung Acad Sci*, 5:17–60, 1960.

[79] B. Eun, Y. Lee, S. Hong, J. Kim, H.-W. Lee, et al. Hes6 controls cell proliferation via interaction with camp-response element-binding protein-binding protein in the promyelocytic leukemia nuclear body. *J Biol Chem*, 283(9):5939–5949, Feb 2008.

[80] G. Euskirchen, T. E. Royce, P. Bertone, R. Martone, J. L. Rinn, et al. Creb binds to multiple loci on human chromosome 22. *Mol Cell Biol*, 24(9):3804–3814, May 2004.

[81] P. J. Farnham. Insights from genomic profiling of transcription factors. *Nat Rev Genet*, 10(9):605–616, Sep 2009.

[82] E. Ferretti, F. Cambronero, S. Tümpel, E. Longobardi, L. M. Wiedemann, et al. Hoxb1 enhancer and control of rhombomere 4 expression: complex interplay between prep1-pbx1-hoxb1 binding sites. *Mol Cell Biol*, 25(19):8541–8552, Oct 2005.

[83] P. C. FitzGerald, A. Shlyakhtenko, A. A. Mir, and C. Vinson. Clustering of dna sequences in human promoters. *Genome Res*, 14(8):1562–74, 2004.

[84] P. Fraser and F. Grosveld. Locus control regions, chromatin activation and transcription. *Curr Opin Cell Biol*, 10(3):361–365, Jun 1998.

[85] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, March 1977.

[86] Y. Freund and R. Schapire. Discussion of breiman. *Annals of Statistics*, 26:824–832, 1998.

[87] M. Fried and D. M. Crothers. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res*, 9(23):6505–6525, Dec 1981.

[88] J. H. Friedman. A recursive partitioning decision rule for nonparametric classification. *IEEE Trans. Comput.*, 26:404–408, April 1977.

[89] J. R. Friedman and K. H. Kaestner. The foxa family of transcription factors in development and metabolism. *Cell Mol Life Sci*, 63(19-20):2317–28, 2006.

[90] N. J. Fuda, M. B. Ardehali, and J. T. Lis. Defining mechanisms that regulate rna polymerase ii transcription in vivo. *Nature*, 461(7261):186–192, Sep 2009.

[91] J. Fürnkranz. Pruning algorithms for rule learning. *Mach. Learn.*, 27:139–172, May 1997.

[92] F. Gao, B. C. Foat, and H. J. Bussemaker. Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data. *BMC Bioinformatics*, 5:31, Mar 2004.

[93] M. M. Garner and A. Revzin. A gel electrophoresis method for quantifying the binding of proteins to specific dna regions: application to components of the escherichia coli lactose operon regulatory system. *Nucleic Acids Res*, 9(13):3047–3060, Jul 1981.

[94] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry. affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–15, 2004.

[95] X. Ge and X. Wang. Role of wnt canonical pathway in hematological malignancies. *J Hematol Oncol*, 3:33, 2010.

[96] Y. Geng and L. F. Johnson. Lack of an initiator element is responsible for multiple transcriptional initiation sites of the tata-less mouse thymidylate synthase promoter. *Mol Cell Biol*, 13(8):4894–4903, Aug 1993.

[97] D. Gidoni, J. T. Kadonaga, H. Barrera-Saldana, K. Takahashi, P. Chambon, et al. Bidirectional sv40 transcription mediated by tandem sp1 binding interactions. *Science*, 230(4725):511–7, 1985.

[98] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*, 99(12):7821–7826, Jun 2002.

[99] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, Jan 2004.

[100] J. L. Goldstein and M. S. Brown. Familial hypercholesterolemia: identification of a defect in the regulation of 3-hydroxy-3-methylglutaryl coenzyme a reductase activity associated with overproduction of cholesterol. *Proc Natl Acad Sci U S A*, 70(10):2804–2808, Oct 1973.

[101] J. L. Goldstein, R. B. Rawson, and M. S. Brown. Mutant mammalian cells as tools to delineate the sterol regulatory element-binding protein pathway for feedback regulation of lipid synthesis. *Arch Biochem Biophys*, 397(2):139–148, Jan 2002.

[102] J. A. Goodrich and R. Tjian. Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. *Nat Rev Genet*, 11(8):549–558, Aug 2010.

[103] P. A. Gray, H. Fu, P. Luo, Q. Zhao, J. Yu, et al. Mouse brain organization revealed through direct genome-scale tf expression analysis. *Science*, 306(5705):2255–2257, Dec 2004.

[104] D. S. Gross and W. T. Garrard. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem*, 57:159–197, 1988.

[105] R. Gualdi, P. Bossard, M. Zheng, Y. Hamada, J. R. Coleman, et al. Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. *Genes Dev*, 10(13):1670–82, 1996.

[106] N. Guelzim, S. Bottani, P. Bourgine, and F. Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet*, 31(1):60–63, May 2002.

[107] D. GuhaThakurta. Computational identification of transcriptional regulatory elements in dna sequence. *Nucleic Acids Res*, 34(12):3585–3598, 2006.

[108] F. Guéhenneux, L. Duret, M. B. Callanan, R. Bouhas, S. Hayette, et al. Cloning of the mouse btg3 gene and definition of a new gene family (the btg family) involved in the negative control of the cell cycle. *Leukemia*, 11(3):370–375, Mar 1997.

[109] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mrna abundance in yeast. *Mol Cell Biol*, 19(3):1720–1730, Mar 1999.

[110] O. Hallikas, K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen, et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124(1):47–59, Jan 2006.

[111] S. Hannenhalli. Eukaryotic transcription factor binding sites - modeling and integrative search methods. *Bioinformatics*, 24(11):1325–1331, Jun 2008.

[112] S. C. Harrison. A structural taxonomy of dna-binding domains. *Nature*, 353(6346):715–719, Oct 1991.

[113] S. E. Hartman, P. Bertone, A. K. Nath, T. E. Royce, M. Gerstein, et al. Global changes in STAT target selection and transcription regulation upon interferon treatments. *Genes Dev*, 19(24):2953–2968, Dec 2005.

[114] N. D. Heintzman and B. Ren. The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome. *Cell Mol Life Sci*, 64(4):386–400, 2007.

[115] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 39(3):311–318, Mar 2007.

[116] P. Hilleren and R. Parker. Mechanisms of mrna surveillance in eukaryotes. *Annu Rev Genet*, 33:229–260, 1999.

[117] J. D. Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet*, 7(3):200–210, Mar 2006.

[118] S. Horike, S. Cai, M. Miyano, J. F. Cheng, and T. Kohwi-Shigematsu. Loss of silent-chromatin looping and impaired imprinting of dlx5 in rett syndrome. *Nat Genet*, 37(1):31–40, 2005.

[119] J. D. Horton, Y. Bashmakov, I. Shimomura, and H. Shimano. Regulation of sterol regulatory element binding proteins in livers of fasted and refed mice. *Proc Natl Acad Sci U S A*, 95(11):5987–5992, May 1998.

[120] J. D. Horton, J. L. Goldstein, and M. S. Brown. Srebps: activators of the complete program of cholesterol and fatty acid synthesis in the liver. *J Clin Invest*, 109(9):1125–1131, May 2002.

[121] J. D. Horton, N. A. Shah, J. A. Warrington, N. N. Anderson, S. W. Park, et al. Combined analysis of oligonucleotide microarray data from transgenic and knockout mice identifies direct srebp target genes. *Proc Natl Acad Sci U S A*, 100(21):12027–12032, Oct 2003.

[122] W. Huber, V. J. Carey, L. Long, S. Falcon, and R. Gentleman. Graphs in molecular biology. *BMC Bioinformatics*, 8 Suppl 6:S8, 2007.

[123] W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002.

[124] S. D. Hueber and I. Lohmann. Shaping segments: Hox gene function in the genomic age. *Bioessays*, 30(10):965–979, Oct 2008.

[125] M. Hulander, W. Wurst, P. Carlsson, and S. Enerback. The winged helix transcription factor fkh10 is required for normal development of the inner ear. *Nat Genet*, 20(4):374–6, 1998.

[126] K. Ikeda, D. J. Steger, A. Eberharter, and J. L. Workman. Activation domain-specific and general transcription stimulation by native histone acetyltransferase complexes. *Mol Cell Biol*, 19(1):855–863, Jan 1999.

[127] E. Ikonen. Mechanisms for cellular cholesterol transport: defects and human disease. *Physiol Rev*, 86(4):1237–1261, Oct 2006.

[128] E. Ikonen. Cellular cholesterol trafficking and compartmentalization. *Nat Rev Mol Cell Biol*, 9(2):125–138, Feb 2008.

[129] H. Imataka, K. Sogawa, K. Yasumoto, Y. Kikuchi, K. Sasano, et al. Two regulatory proteins that bind to the basic transcription element (bte), a gc box sequence in the promoter region of the rat p-4501a1 gene. *Embo J*, 11(10):3663–71, 1992.

[130] A. N. Imbalzano, H. Kwon, M. R. Green, and R. E. Kingston. Facilitated binding of tata-binding protein to nucleosomal dna. *Nature*, 370(6489):481–485, Aug 1994.

[131] T. Inoue, M. Wang, T. O. Ririe, J. S. Fernandes, and P. W. Sternberg. Transcriptional network underlying caenorhabditis elegans vulval development. *Proc Natl Acad Sci U S A*, 102(14):4972–4977, Apr 2005.

[132] IUPAC. Iupac-iub commission on biochemical nomenclature (cbn). abbreviations and symbols for nucleic acids, polynucleotides and their constituents. recommendations 1970. *Eur J Biochem*, 15(2):203–208, Aug 1970.

[133] A. Izenman. *Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning.* Springer New York, 2008.

[134] S. M. Jackson, J. Ericsson, R. Mantovani, and P. A. Edwards. Synergistic activation of transcription by nuclear factor y and sterol regulatory element binding protein. *J Lipid Res*, 39(4):767–776, Apr 1998.

[135] R. Jaenisch and R. Young. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell*, 132(4):567–582, Feb 2008.

[136] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, Oct 2000.

[137] J. Kaufmann and S. T. Smale. Direct recognition of initiator elements by a component of the transcription factor iid complex. *Genes Dev*, 8(7):821–829, Apr 1994.

[138] K. Kawakami, H. Takeda, N. Kawakami, M. Kobayashi, N. Matsuda, et al. A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. *Dev Cell*, 7(1):133–144, Jul 2004.

[139] M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *J Comput Biol*, 7(6):819–837, 2000.

[140] J. B. Kim, G. D. Spotts, Y. D. Halvorsen, H. M. Shih, T. Ellenberger, et al. Dual dna binding specificity of add1/srebp1 controlled by a single amino acid in the basic helix-loop-helix domain. *Mol Cell Biol*, 15(5):2582–2588, May 1995.

[141] T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, et al. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, Aug 2005.

[142] T.-M. Kim and P. J. Park. Advances in analysis of transcriptional regulatory networks. *Wiley Interdiscip Rev Syst Biol Med*, 3(1):21–35, 2011.

[143] K. Kimura, A. Wakamatsu, Y. Suzuki, T. Ota, T. Nishikawa, et al. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res*, 16(1):55–65, Jan 2006.

[144] T. Koide, T. Hayata, and K. W. Y. Cho. Xenopus as a model system to study transcriptional regulatory networks. *Proc Natl Acad Sci U S A*, 102(14):4943–4948, Apr 2005.

[145] R. Kothari and M. Dong. *Pattern Recognition From Classical to Modern Approaches*, chapter Decision trees for classification: a review and some new results, pages 169–184. World Scientific, 2001.

[146] R. Kothary, S. Clapoff, S. Darling, M. D. Perry, L. A. Moran, et al. Inducible expression of an hsp68-lacz hybrid gene in transgenic mice. *Development*, 105(4):707–714, Apr 1989.

[147] M. Koudritsky and E. Domany. Positional distribution of human transcription factor binding sites. *Nucleic Acids Res*, 36(21):6795–805, 2008.

[148] R. Krumlauf. Hox genes in vertebrate development. *Cell*, 78(2):191–201, Jul 1994.

[149] T. K. Kundu and M. R. Rao. Cpg islands in chromatin organization and gene expression. *J Biochem*, 125(2):217–222, Feb 1999.

[150] J. Kwik, S. Boyle, D. Fooksman, L. Margolis, M. P. Sheetz, et al. Membrane cholesterol, lateral mobility, and the phosphatidylinositol 4,5-bisphosphate-dependent organization of cell actin. *Proc Natl Acad Sci U S A*, 100(24):13964–13969, Nov 2003.

[151] H. Kwon, A. N. Imbalzano, P. A. Khavari, R. E. Kingston, and M. R. Green. Nucleosome disruption and enhancement of activator binding by a human sw1/snf complex. *Nature*, 370(6489):477–481, Aug 1994.

[152] P. A. Labosky and K. H. Kaestner. The winged helix transcription factor hfh2 is expressed in neural crest and spinal cord during mouse development. *Mech Dev*, 76(1-2):185–90, 1998.

[153] I. Ladunga. An overview of the computational analyses and discovery of transcription factor binding sites. *Methods Mol Biol*, 674:1–22, 2010.

[154] M. Lagha, J. D. Kormish, D. Rocancourt, M. Manceau, J. A. Epstein, et al. Pax3 regulation of fgf signaling affects the progression of embryonic progenitor cells into the myogenic program. *Genes Dev*, 22(13):1828–1837, Jul 2008.

[155] E. Lai, V. R. Prezioso, W. F. Tao, W. S. Chen, and J. Darnell, J. E. Hepatocyte nuclear factor 3 alpha belongs to a gene family in mammals that is homologous to the drosophila homeotic gene fork head. *Genes Dev*, 5(3):416–27, 1991.

[156] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[157] C. S. Lee, J. R. Friedman, J. T. Fulmer, and K. H. Kaestner. The initiation of liver development is dependent on foxa transcription factors. *Nature*, 435(7044):944–7, 2005.

[158] T. I. Lee and R. A. Young. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet*, 34:77–137, 2000.

[159] H. Lei, S. P. Oh, M. Okano, R. Juttermann, K. A. Goss, et al. De novo dna cytosine methyltransferase activities in mouse embryonic stem cells. *Development*, 122(10):3195–205, 1996.

[160] B. Lemon and R. Tjian. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev*, 14(20):2551–2569, Oct 2000.

[161] K. G. Leong and W.-Q. Gao. The notch pathway in prostate development and cancer. *Differentiation*, 76(6):699–716, Jul 2008.

[162] M. Levine. Transcriptional enhancers in animal development and evolution. *Curr Biol*, 20(17):R754–R763, Sep 2010.

[163] M. Levine and E. H. Davidson. Gene regulatory networks for development. *Proc Natl Acad Sci U S A*, 102(14):4936–4942, Apr 2005.

[164] M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–51, 2003.

[165] M. Lewitzky and S. Yamanaka. Reprogramming somatic cells towards pluripotency by defined factors. *Curr Opin Biotechnol*, 18(5):467–473, Oct 2007.

[166] Q. Li, D. Ritter, N. Yang, Z. Dong, H. Li, et al. A systematic approach to identify functional motifs within vertebrate developmental enhancers. *Dev Biol*, 337(2):484–495, Jan 2010.

[167] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[168] S. M. Lin, P. Du, W. Huber, and W. A. Kibbe. Model-based variance-stabilizing transformation for illumina microarray data. *Nucleic Acids Res*, 36(2):e11, Feb 2008.

[169] P. V. Loo and P. Marynen. Computational methods for the detection of cis-regulatory modules. *Brief Bioinform*, 10(5):509–524, Sep 2009.

[170] J. Lu, W. Lee, C. Jiang, and E. B. Keller. Start site selection by sp1 in the tata-less human ha-ras promoter. *J Biol Chem*, 269(7):5391–5402, Feb 1994.

[171] Z. Luan, Y. Zhang, A. Liu, Y. Man, L. Cheng, et al. A novel gtp-binding protein hgbp3 interacts with nik/hgk. *FEBS Lett*, 530(1-3):233–238, Oct 2002.

[172] M. Z. Ludwig, C. Bergman, N. H. Patel, and M. Kreitman. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403(6769):564–567, Feb 2000.

[173] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton. An overview of the structures of protein-dna complexes. *Genome Biol*, 1(1):REVIEWS001, 2000.

[174] V. J. Makeev, A. P. Lifanov, A. G. Nazina, and D. A. Papatsenko. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res*, 31(20):6016–26, 2003.

[175] S. Malik and R. G. Roeder. The metazoan mediator co-activator complex as an integrative hub for transcriptional regulation. *Nat Rev Genet*, 11(11):761–772, Nov 2010.

[176] T. K. Man and G. D. Stormo. Non-independence of mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (qumfra) assay. *Nucleic Acids Res*, 29(12):2471–2478, Jun 2001.

[177] M. Marin, A. Karis, P. Visser, F. Grosveld, and S. Philipsen. Transcription factor sp1 is essential for early embryonic development but dispensable for cell growth and differentiation. *Cell*, 89(4):619–28, 1997.

[178] R. Martone, G. Euskirchen, P. Bertone, S. Hartman, T. E. Royce, et al. Distribution of nf-kappab-binding sites across human chromosome 22. *Proc Natl Acad Sci U S A*, 100(21):12247–12252, Oct 2003.

[179] G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, 7:29–59, 2006.

[180] S. Matsuda, J. Kawamura-Tsuzuku, M. Ohsugi, M. Yoshida, M. Emi, et al. Tob, a novel protein that interacts with p185erbb2, is associated with anti-proliferative activity. *Oncogene*, 12(4):705–713, Feb 1996.

[181] J. S. Mattick. Rna regulation: a new genetics? *Nat Rev Genet*, 5(4):316–323, Apr 2004.

[182] J. S. Mattick, R. J. Taft, and G. J. Faulkner. A global view of genomic information–moving beyond the gene and the master regulator. *Trends Genet*, 26(1):21–28, Jan 2010.

[183] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, et al. Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–10, 2006.

[184] F. R. Maxfield and A. K. Menon. Intracellular sterol transport and distribution. *Curr Opin Cell Biol*, 18(4):379–385, Aug 2006.

[185] F. R. Maxfield and I. Tabas. Role of cholesterol and lipid organization in disease. *Nature*, 438(7068):612–621, Dec 2005.

[186] C. E. McPherson, E. Y. Shim, D. S. Friedman, and K. S. Zaret. An active tissue-specific enhancer and bound transcription factors existing in a precisely positioned nucleosomal array. *Cell*, 75(2):387–98, 1993.

[187] M. Megraw, F. Pereira, S. T. Jensen, U. Ohler, and A. G. Hatzigeorgiou. A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res*, 19(4):644–656, Apr 2009.

[188] D. M. Mellerick and M. Nirenberg. Dorsal-ventral patterning genes restrict nk-2 homeobox gene expression to the ventral half of the central nervous system of drosophila embryos. *Dev Biol*, 171(2):306–316, Oct 1995.

[189] S. Milgram. The small-world problem. *Psychol. Today*, 1:61–67, 1967.

[190] R. A. Miller, A. Galecki, and R. J. Shmookler-Reis. Interpretation, design, and analysis of gene array expression experiments. *J Gerontol A Biol Sci Med Sci*, 56(2):B52–B57, Feb 2001.

[191] J. Mingers. An empirical comparison of pruning methods for decision tree induction. *Mach. Learn.*, 4:227–243, November 1989.

[192] P. J. Mitchell and R. Tjian. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245(4916):371–378, Jul 1989.

[193] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[194] B. Modrek and C. Lee. A genomic view of alternative splicing. *Nat Genet*, 30(1):13–19, Jan 2002.

[195] R. H. Morley, K. Lachani, D. Keefe, M. J. Gilchrist, P. Flicek, et al. A gene regulatory network directed by zebrafish no tail accounts for its roles in mesoderm formation. *Proc Natl Acad Sci U S A*, 106(10):3829–3834, Mar 2009.

[196] A. Murrell, S. Heeson, and W. Reik. Interaction between differentially methylated regions partitions the imprinted genes igf2 and h19 into parent-specific chromatin loops. *Nat Genet*, 36(8):889–93, 2004.

[197] S. S. Myatt and E. W. Lam. The emerging roles of forkhead box (fox) proteins in cancer. *Nat Rev Cancer*, 7(11):847–59, 2007.

[198] S. Nakielny and G. Dreyfuss. Transport of proteins and rnas in and out of the nucleus. *Cell*, 99(7):677–690, Dec 1999.

[199] C. E. Nelson, B. M. Hersh, and S. B. Carroll. The regulatory content of intergenic dna shapes genome architecture. *Genome Biol*, 5(4):R25, 2004.

[200] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(2 Pt 2):026113, Feb 2004.

[201] M. A. Nobrega, I. Ovcharenko, V. Afzal, and E. M. Rubin. Scanning human gene deserts for long-range enhancers. *Science*, 302(5644):413, 2003.

[202] A. Ochoa-Espinosa and S. Small. Developmental mechanisms and cis-regulatory codes. *Curr Opin Genet Dev*, 16(2):165–170, Apr 2006.

[203] A. Ochoa-Espinosa, G. Yucel, L. Kaplan, A. Pare, N. Pura, et al. The role of binding site cluster strength in bicoid-dependent patterning in drosophila. *Proc Natl Acad Sci U S A*, 102(14):4960–4965, Apr 2005.

[204] S. Ogbourne and T. M. Antalis. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem J*, 331 ( Pt 1):1–14, Apr 1998.

[205] U. Ohler, G. chun Liao, H. Niemann, and G. M. Rubin. Computational analysis of core promoters in the drosophila genome. *Genome Biol*, 3(12):RESEARCH0087, 2002.

[206] U. Ohler and D. A. Wassarman. Promoting developmental transcription. *Development*, 137(1):15–26, Jan 2010.

[207] M. Okano, D. W. Bell, D. A. Haber, and E. Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–57, 1999.

[208] A. R. Oliphant, C. J. Brandl, and K. Struhl. Defining the sequence specificity of dna-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast gcn4 protein. *Mol Cell Biol*, 9(7):2944–2949, Jul 1989.

[209] P. Oliveri and E. H. Davidson. Gene regulatory network controlling embryonic specification in the sea urchin. *Curr Opin Genet Dev*, 14(4):351–60, 2004.

[210] E. N. Olson. Myod family: a paradigm for development? *Genes Dev*, 4(9):1454–1461, Sep 1990.

[211] G. Orphanides, T. Lagrange, and D. Reinberg. The general transcription factors of rna polymerase ii. *Genes Dev*, 10(21):2657–2683, Nov 1996.

[212] G. D. Paolo and P. D. Camilli. Phosphoinositides in cell regulation and membrane dynamics. *Nature*, 443(7112):651–657, Oct 2006.

[213] P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680, Oct 2009.

[214] P. J. Park, M. Pagano, and M. Bonetti. A nonparametric scoring algorithm for identifying informative genes from microarray data. In *Pac Symp Biocomput*, pages 52–63, 2001.

[215] P. Pavlidis, Q. Li, and W. S. Noble. The effect of replication on gene expression microarray experiments. *Bioinformatics*, 19(13):1620–1627, Sep 2003.

[216] L. A. Pennacchio, N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega, et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, 2006.

[217] N. Perreault, J. P. Katz, S. D. Sackett, and K. H. Kaestner. Foxl1 controls the wnt/beta-catenin pathway by modulating the expression of proteoglycans in the gut. *J Biol Chem*, 276(46):43328–33, 2001.

[218] C. L. Peterson and J. L. Workman. Promoter targeting and chromatin remodeling by the swi/snf complex. *Curr Opin Genet Dev*, 10(2):187–192, Apr 2000.

[219] L. J. Pike and J. M. Miller. Cholesterol depletion delocalizes phosphatidylinositol bisphosphate and inhibits hormone-stimulated phosphatidylinositol turnover. *J Biol Chem*, 273(35):22298–22304, Aug 1998.

[220] K. Plaimas, R. Eils, and R. König. Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst Biol*, 4:56, 2010.

[221] S. Powers and D. Mu. Genetic similarities between organogenesis and tumorigenesis of the lung. *Cell Cycle*, 7(2):200–204, Jan 2008.

[222] J. Quackenbush. Microarray data normalization and transformation. *Nat Genet*, 32 Suppl:496–501, Dec 2002.

[223] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, March 1986.

[224] J. R. Quinlan. Simplifying decision trees. *Int. J. Man-Mach. Stud.*, 27:221–234, September 1987.

[225] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[226] A. Rada-Iglesias, O. Wallerman, C. Koch, A. Ameur, S. Enroth, et al. Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Hum Mol Genet*, 14(22):3435–47, 2005.

[227] S. Rahmann, T. Muller, and M. Vingron. On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol*, 2:Article7, 2003.

[228] N. Rajewsky, M. Vergassola, U. Gaul, and E. D. Siggia. Computational detection of genomic cis-regulatory modules applied to body patterning in the early drosophila embryo. *BMC Bioinformatics*, 3:30, 2002.

[229] T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, et al. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–52, 2010.

[230] Z. Ravid, M. Bendayan, E. Delvin, A. T. Sane, M. Elchebly, et al. Modulation of intestinal cholesterol absorption by high glucose levels: impact on cholesterol transporters, regulatory enzymes, and transcription

factors. *Am J Physiol Gastrointest Liver Physiol*, 295(5):G873–G885, Nov 2008.

[231] B. D. Reed, A. E. Charos, A. M. Szekely, S. M. Weissman, and M. Snyder. Genome-wide occupancy of srebp1 and its partners nfy and sp1 reveals novel functional roles and combinatorial regulation of distinct classes of genes. *PLoS Genet*, 4(7):e1000133, Jul 2008.

[232] D. Reines, R. C. Conaway, and J. W. Conaway. Mechanism and regulation of transcriptional elongation by rna polymerase ii. *Curr Opin Cell Biol*, 11(3):342–346, Jun 1999.

[233] B. Ren and B. D. Dynlacht. Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors. *Methods Enzymol*, 376:304–315, 2004.

[234] J. J. Repa, G. Liang, J. Ou, Y. Bashmakov, J. M. Lobaccaro, et al. Regulation of mouse sterol regulatory element-binding protein-1c gene (srebp-1c) by oxysterol receptors, lxralpha and lxrbeta. *Genes Dev*, 14(22):2819–2830, Nov 2000.

[235] B. Rhead, D. Karolchik, R. M. Kuhn, A. S. Hinrichs, A. S. Zweig, et al. The ucsc genome browser database: update 2010. *Nucleic Acids Res*, 38(Database issue):D613–D619, Jan 2010.

[236] K. R. Rosenbloom, T. R. Dreszer, M. Pheasant, G. P. Barber, L. R. Meyer, et al. Encode whole-genome data in the ucsc genome browser. *Nucleic Acids Res*, 38(Database issue):D620–D625, Jan 2010.

[237] J. P. Rouault, R. Rimokh, C. Tessa, G. Paranhos, M. Ffrench, et al. Btg1, a member of a new family of antiproliferative genes. *EMBO J*, 11(4):1663–1670, Apr 1992.

[238] J. D. Saffer, S. P. Jackson, and M. B. Annarella. Developmental expression of sp1 in the mouse. *Mol Cell Biol*, 11(4):2189–99, 1991.

[239] N. J. Sakabe and M. A. Nobrega. Genome-wide maps of transcription regulatory elements. *Wiley Interdiscip Rev Syst Biol Med*, 2(4):422–437, 2010.

[240] Y. Sakakura, H. Shimano, H. Sone, A. Takahashi, N. Inoue, et al. Sterol regulatory element-binding proteins induce an entire pathway of cholesterol synthesis. *Biochem Biophys Res Commun*, 286(1):176–183, Aug 2001.

[241] H. B. Sanchez, L. Yieh, and T. F. Osborne. Cooperation by sterol regulatory element-binding protein and sp1 in sterol regulation of low density lipoprotein receptor gene. *J Biol Chem*, 270(3):1161–1169, Jan 1995.

[242] A. Sandelin, P. Bailey, S. Bruce, P. G. Engström, J. M. Klos, et al. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5(1):99, 2004.

[243] A. Sandelin, P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki, et al. Mammalian rna polymerase ii core promoters: insights from genome-wide studies. *Nat Rev Genet*, 8(6):424–436, Jun 2007.

[244] E. E. Schadt, C. Li, B. Ellis, and W. H. Wong. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl*, Suppl 37:120–125, 2001.

[245] E. E. Schadt, C. Li, C. Su, and W. H. Wong. Analyzing high-density oligonucleotide gene expression array data. *J Cell Biochem*, 80(2):192–202, Oct 2000.

[246] G. Schramm, N. Kannabiran, and R. König. Regulation patterns in signaling networks of cancer. *BMC Syst Biol*, 4:162, 2010.

[247] G. Schramm, S. Wiesberg, N. Diessl, A.-L. Kranz, V. Sagulenko, et al. Pathwave: discovering patterns of differentially regulated enzymes in metabolic pathways. *Bioinformatics*, 26(9):1225–1231, May 2010.

[248] J. Schug, W.-P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, et al. Promoter features related to tissue specificity as measured by shannon entropy. *Genome Biol*, 6(4):R33, 2005.

[249] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, 451(7178):535–540, Jan 2008.

[250] M. F. Shannon and S. Rao. Transcription. of chips and chips. *Science*, 296(5568):666–669, Apr 2002.

[251] J. T. Shin, J. R. Priest, I. Ovcharenko, A. Ronco, R. K. Moore, et al. Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. *Nucleic Acids Res*, 33(17):5437–5445, 2005.

[252] K. Simons and E. Ikonen. How cells handle cholesterol. *Science*, 290(5497):1721–1726, Dec 2000.

[253] H. Singh, K. L. Medina, and J. M. R. Pongubala. Contingent gene regulatory networks and b cell fate specification. *Proc Natl Acad Sci U S A*, 102(14):4949–4953, Apr 2005.

[254] S. T. Smale and D. Baltimore. The "initiator" as a transcription control element. *Cell*, 57(1):103–113, Apr 1989.

[255] S. T. Smale and J. T. Kadonaga. The rna polymerase ii core promoter. *Annu Rev Biochem*, 72:449–479, 2003.

[256] G. K. Smyth and T. Speed. Normalization of cDNA microarray data. *Methods*, 31(4):265–273, Dec 2003.

[257] G. K. Smyth, Y. H. Yang, and T. Speed. Statistical issues in cDNA microarray data analysis. *Methods Mol Biol*, 224:111–136, 2003.

[258] B. M. Spiegelman and R. Heinrich. Biological control through regulated transcriptional coactivators. *Cell*, 119(2):157–167, Oct 2004.

[259] R. Staden. Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci*, 5(2):89–96, Apr 1989.

[260] A. Stark. Learning the transcriptional regulatory code. *Mol Syst Biol*, 5:329, 2009.

[261] A. Stathopoulos and M. Levine. Genomic regulatory networks and animal development. *Dev Cell*, 9(4):449–462, Oct 2005.

[262] M. Steinfath, W. Wruck, H. Seidel, H. Lehrach, U. Radelof, et al. Automated image analysis for array hybridization experiments. *Bioinformatics*, 17(7):634–641, Jul 2001.

[263] R. Stoltenburg, C. Reinemann, and B. Strehlitz. Selex–a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol Eng*, 24(4):381–403, Oct 2007.

[264] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

[265] K. Struhl. Yeast transcriptional regulatory mechanisms. *Annu Rev Genet*, 29:651–674, 1995.

[266] D. Strumpf, C. A. Mao, Y. Yamanaka, A. Ralston, K. Chawengsaksophak, et al. Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. *Development*, 132(9):2093–102, 2005.

[267] W. Su, S. Jackson, R. Tjian, and H. Echols. Dna looping between sites for transcriptional activation: self-association of dna-bound sp1. *Genes Dev*, 5(5):820–6, 1991.

[268] B.-C. Suh and B. Hille. Regulation of ion channels by phosphatidylinositol 4,5-bisphosphate. *Curr Opin Neurobiol*, 15(3):370–378, Jun 2005.

[269] K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, Aug 2006.

[270] D. Tautz. Evolution of transcriptional regulation. *Curr Opin Genet Dev*, 10(5):575–579, Oct 2000.

[271] K. Tharakaraman, O. Bodenreider, D. Landsman, J. L. Spouge, and L. Marino-Ramirez. The biological function of some human transcription factor binding motifs varies with position relative to the transcription start site. *Nucleic Acids Res*, 36(8):2777–86, 2008.

[272] W. Thompson, M. J. Palumbo, W. W. Wasserman, J. S. Liu, and C. E. Lawrence. Decoding human regulatory circuits. *Genome Res*, 14(10A):1967–1974, Oct 2004.

[273] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. D. Moor, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–144, Jan 2005.

[274] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121, Apr 2001.

[275] G. Tuteja and K. H. Kaestner. Snapshot: forkhead transcription factors i. *Cell*, 130(6):1160, 2007.

[276] M. Uldry and B. Thorens. The slc2 family of facilitated hexose and polyol transporters. *Pflugers Arch*, 447(5):480–489, Feb 2004.

[277] A. Ungewickell, C. Hugge, M. Kisseleva, S.-C. Chang, J. Zou, et al. The identification and characterization of two phosphatidylinositol-4,5-bisphosphate 4-phosphatases. *Proc Natl Acad Sci U S A*, 102(52):18854–18859, Dec 2005.

[278] P. Van Loo, S. Aerts, B. Thienpont, B. D. Moor, Y. Moreau, et al. Moduleminer - improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol*, 9(4):R66, 2008.

[279] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, Jan 2002.

[280] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4):252–263, Apr. 2009.

[281] G. Varani. A cap for all occasions. *Structure*, 5(7):855–858, Jul 1997.

[282] S. Vardhanabhuti, J. Wang, and S. Hannenhalli. Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res*, 35(10):3203–13, 2007.

[283] T. Vavouri and G. Elgar. Prediction of cis-regulatory elements using binding site matrices–the successes, the failures and the reasons for both. *Curr Opin Genet Dev*, 15(4):395–402, Aug 2005.

[284] M. Vingron, A. Brazma, R. Coulson, J. van Helden, T. Manke, et al. Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biol*, 10(1):202, 2009.

[285] C. Virtanen and M. Takahashi. Muscling in on microarrays. *Appl Physiol Nutr Metab*, 33(1):124–129, Feb 2008.

[286] A. Visel, S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis, et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet*, 40(2):158–60, 2008.

[287] D. Vlieghe, A. Sandelin, P. J. D. Bleser, K. Vleminckx, W. W. Wasserman, et al. A new generation of jaspar, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*, 34(Database issue):D95–D97, Jan 2006.

[288] A. Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, 15(10):776–84, 1999.

[289] H. Wan, S. Dingle, Y. Xu, V. Besnard, K. H. Kaestner, et al. Compensatory roles of Foxa1 and Foxa2 during lung morphogenesis. *J Biol Chem*, 280(14):13809–16, 2005.

[290] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–287, Apr 2004.

[291] V. M. Weake and J. L. Workman. Inducible gene expression: diverse regulatory mechanisms. *Nat Rev Genet*, 11(6):426–437, Jun 2010.

[292] C.-L. Wei, Q. Wu, V. B. Vega, K. P. Chiu, P. Ng, et al. A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 124(1):207–219, Jan 2006.

[293] D. C. Weinstein, A. Ruiz i Altaba, W. S. Chen, P. Hoodless, V. R. Prezioso, et al. The winged-helix transcription factor hnf-3 beta is required for notochord development in the mouse embryo. *Cell*, 78(4):575–88, 1994.

[294] F. Westermann, D. Muth, A. Benner, T. Bauer, K.-O. Henrich, et al. Distinct transcriptional mycn/c-myc activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biology*, 9(10):R150, 2008.

[295] B. Wilczynski and E. E. M. Furlong. Dynamic crm occupancy reflects a temporal map of developmental progression. *Mol Syst Biol*, 6:–, June 2010.

[296] B. Wilczynski, T. R. Hvidsten, A. Kryshtafovych, J. Tiuryn, J. Komorowski, et al. Using local gene expression similarities to discover regulatory binding site modules. *BMC Bioinformatics*, 7:505, 2006.

[297] I. H. Witten and E. Frank. *Data mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2 edition, 2005.

[298] A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, et al. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, 3(1):e7, 2005.

[299] G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, et al. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20(9):1377–1419, Sep 2003.

[300] T. D. Wu. Analysing gene expression data from dna microarrays to identify candidate genes. *J Pathol*, 195(1):53–65, Sep 2001.

[301] H. Würtele and P. Chartrand. Genome-wide scanning of hoxb1-associated loci in mouse es cells using an open-ended chromosome conformation capture methodology. *Chromosome Res*, 14(5):477–495, 2006.

[302] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, et al. Systematic discovery of regulatory motifs in human promoters and 3' utrs by comparison of several mammals. *Nature*, 434(7031):338–45, 2005.

[303] H. Yamagishi, J. Maeda, T. Hu, J. McAnally, S. J. Conway, et al. Tbx1 is regulated by tissue-specific forkhead proteins through a common sonic hedgehog-responsive enhancer. *Genes Dev*, 17(2):269–81, 2003.

[304] K. D. Yokoyama, U. Ohler, and G. A. Wray. Measuring spatial preferences at fine-scale resolution identifies known and novel cis-regulatory element candidates and functional motif-pair relationships. *Nucleic Acids Res*, 37(13):e92, 2009.

[305] L. Zawel, K. P. Kumar, and D. Reinberg. Recycling of the general transcription factors during rna polymerase ii transcription. *Genes Dev*, 9(12):1479–1490, Jun 1995.

[306] J. Zeitlinger and A. Stark. Developmental gene regulation in the era of genomics. *Dev Biol*, 339(2):230–239, Mar 2010.

[307] J. Zhao, L. Hyman, and C. Moore. Formation of mrna 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mrna synthesis. *Microbiol Mol Biol Rev*, 63(2):405–445, Jun 1999.

[308] X. Zhu, M. Gerstein, and M. Snyder. Getting connected: analysis and principles of biological networks. *Genes Dev*, 21(9):1010–1024, May 2007.

[309] A. Zien, T. Aigner, R. Zimmer, and T. Lengauer. Centralization: a new method for the normalization of gene expression data. *Bioinformatics*, 17 Suppl 1:S323–S331, 2001.

[310] R. P. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E. E. Furlong. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70, 2009.

[311] R. P. Zinzen, K. Senger, M. Levine, and D. Papatsenko. Computational models for neurogenic gene expression in the drosophila embryo. *Curr Biol*, 16(13):1358–1365, Jul 2006.

[312] J. Zlatanova, S. H. Leuba, and K. van Holde. Chromatin fiber structure: morphology, molecular determinants, structural transitions. *Biophys J*, 74(5):2554–2566, May 1998.

[313] J. Zou, J. Marjanovic, M. V. Kisseleva, M. Wilson, and P. W. Majerus. Type I phosphatidylinositol-4,5-bisphosphate 4-phosphatase regulates stress-induced apoptosis. *Proc Natl Acad Sci U S A*, 104(43):16834–16839, Oct 2007.

# Appendix A

# Identification of spatio-temporal specific regulatory modules

**Table A.1. Grouping of transcription factors according to common PWMs.**

| Representative TF | TFs within the group | Binding preference |
|---|---|---|
| AFP1 | AFP1 | uniformly-distributed-BS |
| AIRE | AIRE | TSS-depleted BS |
| Alx-4 | Alx-4 | uniformly-distributed-BS |
| aMEF-2 | aMEF-2, MEF2A-isoform1, MEF-2DAB | TSS-depleted BS |
| AML1 | AML1, AML2, AML1b, AML1c, AML1a, PEBP2beta, AML3, PEBP2alpha, AML3-isoform2, AML3-isoform 1 | uniformly-distributed-BS |
| AP-2alphaA | AP-2alphaA, AP-2gamma, AP-2alphaB, AP-2beta | TSS-enriched-BS |
| AP-2rep | AP-2rep | |
| AP-3 (2) | AP-3 (2) | |
| AP-4 | AP-4 | TSS-enriched-BS |
| AR | AR | uniformly-distributed-BS |
| Arnt | Arnt, HIF-1, HIF-1alpha, AhR, Arnt (774 AA form), ARNT2, AhR:Arnt, HIF-1alpha-isoform1 | TSS-enriched-BS |

| Representative TF | TFs within the group | Binding preference |
|---|---|---|
| ATF3 | ATF3, ATF-4, ATF-2-xbb4, AP-1, c-Fos, c-Jun, Fra-1, JunB, JunD, Fra-2, JunB:Fra-2, JunB:Fr a -1, JunD:Fra-2, c-Jun:JunD, FosB, deltaFosB, ATF-1, ATF6, ATF, CREB, deltaCREB, ATF-a, 120-kDa CRE-binding protein, CREMalpha, ATF-2, ATFa-isoform1, ATF5 | TSS-enriched-BS |
| BCL-6 | BCL-6 | uniformly-distributed-BS |
| Blimp-1 | Blimp-1 | uniformly-distributed-BS |
| BRCA1 | BRCA1, BRCA1:USF2 | TSS-depleted-BS |
| CACCC-binding factor | CACCC-binding factor | TSS-enriched-BS |
| Cart-1 | Cart-1 | TSS-depleted-BS |
| CDC5L | CDC5L | uniformly-distributed-BS |
| CDP | CDP, CDP-isoform1 | uniformly-distributed-BS |
| CDX2 | CDX2, Cdx-1 | TSS-depleted-BS |
| C/EBPdelta | C/EBPdelta, C/EBPgamma, C/EBPalpha, CHOP-10, C/EBPbeta, C/EBPepsilon | TSS-depleted-BS |
| Chx10 | Chx10 | uniformly-distributed-BS |
| Clock:BMAL1 | Clock:BMAL1, Clock:BMAL2 | |
| c-Myb | c-Myb, c-Myb-isoform1 | uniformly-distributed-BS |
| CP2a | CP2a | TSS-enriched-BS |
| c-Rel | c-Rel | uniformly-distributed-BS |
| Crx | Crx, RX | uniformly-distributed-BS |
| DBP | DBP | TSS-depleted-BS |
| E2F-1 | E2F-1, E2F-1:DP-1, E2F:DP, E2F-2, E2F-3a, E2F-4, E2F-5, E2F:DP:E4, DP-1, E2F-7 | TSS-enriched-BS |
| E2F-1:DP-2 | E2F-1:DP-2 | TSS-enriched-BS |
| E2F-4:DP-1 | E2F-4:DP-1 | |
| E4BP4 | E4BP4 | uniformly-distributed-BS |
| E4F1 | E4F1 | uniformly-distributed-BS |
| Egr-1 | Egr-1, Egr-2, Egr-3, Egr-4 | TSS-enriched-BS |

| Representative TF | TFs within the group | Binding preference |
|---|---|---|
| Elk-1 | Elk-1, Elk-1-isoform1, GABP-alpha:GABP-beta, GABP-alpha, GABP-beta1, GABP-beta2, Tel-2 b, Tel-2a, Tel-2c, c-Ets-2, PU.1, PEA3, ELF-1, c-Ets-1, NERF-1a, Erg-1, SAP-1a, Ets-1 deltaVII, Fli-1, ERF, NERF-1b, NERF-2, NERF-2b, Erg-2, TCF, Spi-B, p38erg, Net, TEL1, p55, SA P-1b, p55erg, p49erg, Tel-2d, Tel-2e, Tel-2f, ELFR, Spi-B-isoform1 | TSS-enriched-BS |
| ERR1 | ERR1 | uniformly-distributed-BS |
| ER-alpha | ER-alpha, ER-alpha-L, ER-beta, ER-alpha:ER-beta | uniformly-distributed-BS |
| ETF | ETF | |
| FAC1 | FAC1, FAC1-xbb1 | TSS-depleted-BS |
| FBI-1 | FBI-1 | |
| FOXA1 | FOXA1, FOXA2, FOXA3 | TSS-depleted-BS |
| FOXC1 | FOXC1 | uniformly-distributed-BS |
| FOXD1 | FOXD1 | uniformly-distributed-BS |
| FOXI1 | FOXI1 | TSS-depleted-BS |
| FOXJ1a | FOXJ1a, FOXJ1b, FOXD3, FOXF2, FOXF1 | TSS-depleted-BS |
| FOXJ2 | FOXJ2 | TSS-depleted-BS |
| FOXL1 | FOXL1 | TSS-depleted-BS |
| FOXO1 | FOXO1 | TSS-depleted-BS |
| FOXO3a | FOXO3a, FOXO3A-1 | TSS-depleted-BS |
| FOXO4 | FOXO4 | TSS-depleted-BS |
| FOXN1 | FOXN1 | uniformly-distributed-BS |
| FOXP3 | FOXP3 | uniformly-distributed-BS |
| GATA-2 | GATA-2, GATA-3 isoform-1, GATA-4, GATA-6, GATA-1, GATA-1 isoform 1, GATA-5 | TSS-depleted-BS |
| GCMa | GCMa | uniformly-distributed-BS |
| GCNF-2 | GCNF-2, GCNF-1 | uniformly-distributed-BS |
| Gfi1b | Gfi1b, Gfi1 | uniformly-distributed-BS |
| GLI1 | GLI1, GLI3, GLI2alpha | uniformly-distributed-BS |
| GR-alpha | GR-alpha, GR-beta, GR, PR B, PR A | uniformly-distributed-BS |

| Representative TF | TFs within the group | Binding preference |
|---|---|---|
| HES-1 | HES-1 | uniformly-distributed-BS |
| HIC-1 | HIC-1 | TSS-enriched-BS |
| Hlf | Hlf | uniformly-distributed-BS |
| HMGI | HMGI, HMG-Y, HMGI-C | TSS-depleted-BS |
| HNF-1alpha-A | HNF-1alpha-A, HNF-1beta-A, HNF-1alpha-B, HNF-1alpha-C, HNF-1beta-B, HNF-1beta-C | TSS-depleted-BS |
| HOXA4 | HOXA4 | |
| HOXA7 | HOXA7 | |
| HOXA9B | HOXA9B, Meis-1 | uniformly-distributed-BS |
| HSF | HSF | uniformly-distributed-BS |
| HSF1-L | HSF1-L, HSF1-S | uniformly-distributed-BS |
| HSF2 | HSF2, HSF2A | uniformly-distributed-BS |
| Ikaros | Ikaros | |
| IPF1 | IPF1, IPF1:Pbx | uniformly-distributed-BS |
| IRF-7A | IRF-7A, IRF-1, IRF-8, IRF-2, IRF-3, IRF-4, IRF-5, IRF-7H, IRF-6, ISGF-3, IRF-9 | TSS-depleted-BS |
| Kid3 | Kid3 | |
| LBP-1 | LBP-1 | |
| LEF-1 | LEF-1, TCF-1, TCF-1A, TCF-1B, TCF-1C, TCF-1E, TCF-1F, TCF-1G | uniformly-distributed-BS |
| LF-A1 | LF-A1 | |
| LHX3a | LHX3a, LHX3b | TSS-depleted-BS |
| Lmo2 | Lmo2 | uniformly-distributed-BS |
| LRH-1 | LRH-1, LRH-1-xbb1, LRH-1-isoform1 | uniformly-distributed-BS |
| MAZR | MAZR | TSS-enriched-BS |
| MAZ | MAZ | TSS-enriched-BS |
| MIF-1 | MIF-1 | TSS-enriched-BS |

| Representative TF | TFs within the group | Binding preference |
|---|---|---|
| MRF-2-isoform1 | MRF-2-isoform1, USF2, USF2a, MyoD, Tal1-PP42, E12, E47, TFE3, TFEB-isoform1, MITF, MITF-M1 , DEC1, DEC2, HTF4, HTF4gamma, ITF-2, Tal1-PP22, SEF2-1B, ITF, ITF-1, Myogenin, Myf-5, Myf-6, MASH-1, INSAF, c-Myc, Max-isoform2, Max-isoform1, Max, USF1, USF2b, Mxi1, Mad1, deltaMax, Tal-2, HEN1, HEB1-p94, HEB1-p67, N-Myc, HAND1, HAND2, USF1:USF2 | |
| MTF-1 | MTF-1 | TSS-enriched-BS |
| MZF1B-C | MZF1B-C | TSS-enriched-BS |
| NCX | NCX | uniformly-distributed-BS |
| NF-1 | NF-1, CTF, CTF-1, CTF-2 | uniformly-distributed-BS |
| NF-AT1 | NF-AT1, NF-AT2, NF-AT4, NF-AT3, NF-AT1C | TSS-depleted-BS |
| NF-muE1 | NF-muE1 | uniformly-distributed-BS |
| NF-Y | NF-Y, NF-YB, NF-YA, NF-YC-3, NF-YA-L | TSS-enriched-BS |
| Nkx2-1 | Nkx2-1, Nkx2-1-isoform1 | uniformly-distributed-BS |
| Nkx2-2 | Nkx2-2 | uniformly-distributed-BS |
| Nkx2-5 | Nkx2-5 | uniformly-distributed-BS |
| Nkx3-1 | Nkx3-1, Nkx3-1 v1, Nkx3-1 v2, Nkx3-1 v3, Nkx3-1 v4 | TSS-depleted-BS |
| Nkx6-1 | Nkx6-1 | TSS-depleted-BS |
| Nkx6-2 | Nkx6-2 | TSS-depleted-BS |
| NRF-1 | NRF-1 | TSS-enriched-BS |
| Nrf2 | Nrf2, Nrf2-isoform1, LCR-F1, Bach1, NF-E2, NF-E2, p45, Bach2, Nrf1, MafG, MafK, Nrf1:MafK, Nrf2:MafK, Nrf3:MafK, MafG:MafG, Nrf1:MafG, Nrf2:MafG, MafF, MafB, Maf | uniformly-distributed-BS |
| OC-2 | OC-2, HNF-6alpha | TSS-depleted-BS |
| Otx1 | Otx1, Otx2 | uniformly-distributed-BS |
| p300 | p300 | TSS-enriched-BS |
| p53 | p53, p53-isoform1, p73alpha, DeltaNp63alpha, p63gamma, p73beta | uniformly-distributed-BS |
| Pax-1 | Pax-1 | uniformly-distributed-BS |
| Pax-5 | Pax-5, Pax-8, Pax-2 | uniformly-distributed-BS |
| Pax-3 | Pax-3 | uniformly-distributed-BS |

| Representative TF | TFs within the group | Binding preference |
|---|---|---|
| Pax-6 | Pax-6, Pax-6 / Pd-5a | uniformly-distributed-BS |
| Pbx1a | Pbx1a, Pbx1b, Pbx2, Pbx3a, Pbx3b, Pbx1:Prep1, Pbx1:HOXB1, Pbx, Pbx1b:Prep1, Pbx2:Prep1, Pb x1a:HOXC6, Pbx2:HOXC6, Pbx3a:HOXC6, Pbx1a:HOXB7, Pbx1a:HOXB8, Pbx1a:IPF1 | uniformly-distributed-BS |
| Pbx1 | Pbx1 | uniformly-distributed-BS |
| Pit-1B | Pit-1B | TSS-depleted-BS |
| PITX2 | PITX2, PITX2A | TSS-depleted-BS |
| PLZF | PLZF, PLZFB | TSS-depleted-BS |
| POU3F1 | POU3F1, POU3F2, POU3F2 (N-Oct-5a), POU3F2 (N-Oct-5b), POU2F1, Oct-1, POU2F2 (Oct-2.1), Oct-2, POU2F2B, oct-B2, oct-B3, Oct-2.1, POU5F1A, Octa-factor, octamer-bindin g factor, POU5F1B, OCA-B, POU4F1(l), POU5F1C | TSS-depleted-BS |
| PPARgamma1 | PPARgamma1, PPARgamma2, VDR, CAR, PXR-1A, FXR, FXR:RXR-alpha, LXR-alpha:RXR-alpha, LXR-b eta:RXR-alpha, HNF-4alpha2, RAR-alpha1, RAR-gamma, RAR-beta, T3R-alpha, T3R-beta1, T3R-alpha1, T3R-alpha2, RAR-beta2, RXR-beta, RXR-alpha, RXR-gamma, RAR-alpha, RAR-alpha:RXR -alpha, RAR-alpha:RXR-gamma, COUP-TF2, COUP-TF1, HNF-4gamma, HNF-4, HNF-4alpha, PPARalpha:RXR-alpha, LXR-alpha, LXR-beta, SXR:RXR-alpha, PXR-1A:RXR-alpha, PXR-1A:RXR-beta, C AR:RXR-alpha, HNF-4alpha1, HNF-4alpha4, HNF-4alpha3, HNF-4alpha7, PPARalpha, FXR-alpha | uniformly-distributed-BS |
| PR | PR | uniformly-distributed-BS |
| pRb:E2F-1:DP-1 | pRb:E2F-1:DP-1 | |
| PXR:RXR-alpha | PXR:RXR-alpha | |
| RBP-Jkappa | RBP-Jkappa | uniformly-distributed-BS |
| RelA-p65 | RelA-p65, p50, NF-kappaB, p52, NF-kappaB(-like), NF-kappaB1, p100, NF-kappaB2 (p49) | uniformly-distributed-BS |
| REST-form2 | REST-form2, REST-form1, REST | TSS-enriched-BS |
| RFX2 | RFX2, RFX3, RFX5, RFX1, RFX4, RFX1:RFX2, RFX1:RFX3, RFXANK, RFXAP, RFX5:RFXAP:RFXANK | uniformly-distributed-BS |
| RORalpha1 | RORalpha1, RORalpha2, RORalpha3, RORalpha | uniformly-distributed-BS |
| RP58 | RP58 | uniformly-distributed-BS |
| RREB-1 | RREB-1 | TSS-enriched-BS |

| Representative TF | TFs within the group | Binding preference |
|---|---|---|
| RSRFC4 | RSRFC4 | TSS-depleted-BS |
| SF-1 | SF-1 | uniformly-distributed-BS |
| Smad3 | Smad3, Smad4, Smad1, Smad2-L, Smad3:Smad4, Smad7, Smad6, Smad5, Smad2 (437 AA), Smad4delta3, Smad4delta6, Smad4delta5-6, Smad4delta4-6, Smad4delta4-7 | uniformly-distributed-BS |
| Sp1 | Sp1, Sp3-isoform1, Sp2, Sp3, Sp4 | TSS-enriched-BS |
| SREBP-2 | SREBP-2, SREBP-1a, SREBP-1b, SREBP-1c, SREBP-1 | uniformly-distributed-BS |
| SRF | SRF | uniformly-distributed-BS |
| SRY | SRY, Sox4, Sox9, Sox5, Sox11, Sox12, Sox20, Sox2, Sox3, Sox8, Sox10, Sox14, Sox18, Sox21 | uniformly-distributed-BS |
| STAT5B | STAT5B, STAT6, STAT1alpha, STAT1beta, STAT1, STAT3, STAT5A, STAT4, STAT2 | uniformly-distributed-BS |
| SZF1-1 | SZF1-1 | uniformly-distributed-BS |
| TBP | TBP, TFIID | TSS-enriched-BS |
| TBX5-L | TBX5-L | TSS-depleted-BS |
| TCF-4 | TCF-4 | |
| TEF-1 | TEF-1 | TSS-depleted-BS |
| TEF-xbb1 | TEF-xbb1 | TSS-depleted-BS |
| TFIIA | TFIIA, TFIIA-alpha/beta precursor (major), TFIIA-alpha/beta precursor (minor), TFIIA-gamma | TSS-enriched-BS |
| TFII-I | TFII-I | TSS-enriched-BS |
| TGIF | TGIF, TGIF-isoform2 | uniformly-distributed-BS |
| Topors-isoform1 | Topors-isoform1 | TSS-depleted-BS |
| WT1 | WT1, WT1 I -KTS, WT1 -KTS, WT1 I, WT1-del2, WT1 I-del2 | |
| XBP-1 | XBP-1 | uniformly-distributed-BS |
| YY1 | YY1 | TSS-enriched-BS |
| ZBRK1 | ZBRK1 | uniformly-distributed-BS |
| ZBTB7B | ZBTB7B | uniformly-distributed-BS |
| ZEB (1124 AA) | ZEB (1124 AA) | uniformly-distributed-BS |
| ZIC2 | ZIC2 | uniformly-distributed-BS |

| Representative TF | TFs within the group | Binding preference |
|---|---|---|
| ZID | ZID | uniformly-distributed-BS |
| ZNF219 | ZNF219 | TSS-enriched-BS |

# Appendix B

# Identification of novel putative SREBP target genes

**Table B.1. Differentially regulated genes in fibroblasts and HeLa cells under sterol-depletion.**

***A Genes up-regulated under sterol-depleted conditions***

| | | Fibroblasts | | HeLa cells | |
|---|---|---|---|---|---|
| Symbol | Name | FC[1] | pfp[2] | FC[1] | pfp[2] |
| HMGCR | 3-hydroxy-3-methylglutaryl-CoA reductase | 3.5239 | 0 | 4.3366 | 0 |
| FADS1 | fatty acid desaturase 1 | 3.1754 | 0 | 1.9445 | 0.0012 |
| INSIG1 | insulin induced gene 1 | 2.8316 | 0.0019 | 3.2167 | 0 |
| RGS2 | regulator of G-protein signaling 2, 24kDa | 2.4397 | 0.0021 | 1.6181 | 0.0281 |
| IDI1 | isopentenyl-diphosphate delta isomerase 1 | 2.2995 | 0.0028 | 2.1098 | 0.0013 |
| SCD | stearoyl-CoA desaturase (delta-9-desaturase) | 2.1175 | 0.0038 | 1.9835 | 0.0011 |
| LPIN1 | lipin 1 | 2.0214 | 0.004 | 1.7186 | 0.049 |
| DHCR7 | 7-dehydrocholesterol reductase | 1.9957 | 0.0043 | 1.883 | 0.0034 |
| ACSS2 | acyl-CoA synthetase short-chain family member 2 | 2.1836 | 0.0044 | 5.5355 | 0 |
| TRIB3 | tribbles homolog 3 (Drosophila) | 1.4702 | 0.0089 | 1.6387 | 0.0434 |
| ACLY | ATP citrate lyase | 1.9234 | 0.0095 | 1.6379 | 0.0178 |
| TMEM55B | transmembrane protein 55B | 1.7783 | 0.0133 | 2.0197 | 0.0017 |
| SQLE | squalene epoxidase | 1.7449 | 0.0157 | 2.4489 | 0 |
| FASN | fatty acid synthase | 1.7046 | 0.0185 | 1.8882 | 1.00E-04 |

[1] Fold change in fibroblasts or HeLa-cells

[2] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

| | | Fibroblasts | | HeLa cells | |
|---|---|---|---|---|---|
| Symbol | Name | FC[1] | pfp[2] | FC[1] | pfp[2] |
| ACAT2 | acetyl-CoA acetyltransferase 2 | 1.6689 | 0.0358 | 1.9446 | 9.00E-04 |
| SC4MOL | sterol-C4-methyl oxidase-like | 1.5652 | 0.0468 | 3.7526 | 0 |
| HES6 | hairy and enhancer of split 6 (Drosophila) | 3.7806 | 0 | | |
| BHLHE40 | basic helix-loop-helix family, member e40 | 2.616 | 3.00E-04 | | |
| RGS16 | regulator of G-protein signaling 16 | 2.9205 | 4.00E-04 | | |
| HMGCS1 | 3-hydroxy-3-methylglutaryl-CoA synthase 1 (soluble) | 3.1076 | 5.00E-04 | | |
| TMEM97 | transmembrane protein 97 | 2.6274 | 0.0014 | | |
| MVD | mevalonate (diphospho) decarboxylase | 2.3226 | 0.0023 | | |
| MAFB | v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian) | 2.3337 | 0.0024 | | |
| STC1 | stanniocalcin 1 | 2.3742 | 0.0028 | | |
| FABP3 | fatty acid binding protein 3, muscle and heart (mammary-derived growth inhibitor) | 2.3738 | 0.003 | | |
| KLF6 | Kruppel-like factor 6 | 3.1659 | 0.0031 | | |
| PIK3IP1 | phosphoinositide-3-kinase interacting protein 1 | 2.3102 | 0.0031 | | |
| PDGFRB | platelet-derived growth factor receptor, beta polypeptide | 2.3345 | 0.0035 | | |
| TM7SF2 | transmembrane 7 superfamily member 2 | 2.155 | 0.0036 | | |
| SLC20A1 | solute carrier family 20 (phosphate transporter), member 1 | 2.0112 | 0.004 | | |
| PFKFB4 | 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase<br>4 | 2.0521 | 0.004 | | |
| KIAA1199 | KIAA1199 | 2.4186 | 0.0041 | | |
| RGS4 | regulator of G-protein signaling 4 | 2.2216 | 0.0043 | | |
| PDGFRL | platelet-derived growth factor receptor-like | 2.2106 | 0.0043 | | |
| C17orf59 | chromosome 17 open reading frame 59 | 2.0269 | 0.0045 | | |
| PCDH18 | protocadherin 18 | 2.1986 | 0.0066 | | |
| MXRA5 | matrix-remodelling associated 5 | 2.1595 | 0.0086 | | |
| LDLR | low density lipoprotein receptor | 1.8743 | 0.0089 | | |
| GAS1 | growth arrest-specific 1 | 2.0526 | 0.0096 | | |

[1] Fold change in fibroblasts or HeLa-cells

[2] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

| | | Fibroblasts | | HeLa cells | |
|---|---|---|---|---|---|
| Symbol | Name | FC[1] | pfp[2] | FC[1] | pfp[2] |
| TP53INP2 | tumor protein p53 inducible nuclear protein 2 | 1.8091 | 0.0101 | | |
| KLF9 | Kruppel-like factor 9 | 1.7681 | 0.0117 | | |
| DBC1 | deleted in bladder cancer 1 | 1.906 | 0.0128 | | |
| SLC2A6 | solute carrier family 2 (facilitated glucose transporter), member 6 | 1.8013 | 0.0132 | | |
| TP53INP1 | tumor protein p53 inducible nuclear protein 1 | 1.7629 | 0.015 | | |
| KCNJ2 | potassium inwardly-rectifying channel, subfamily J, member 2 | 1.8601 | 0.0151 | | |
| ANGPTL2 | angiopoietin-like 2 | 1.7884 | 0.0151 | | |
| BCL10 | B-cell CLL/lymphoma 10 | 1.7948 | 0.0154 | | |
| PDGFRA | platelet-derived growth factor receptor, alpha polypeptide | 1.8534 | 0.0155 | | |
| HSD17B12 | hydroxysteroid (17-beta) dehydrogenase 12 | 1.7475 | 0.0155 | | |
| FBLN1 | fibulin 1 | 2.4985 | 0.0167 | | |
| SH3PXD2B | SH3 and PX domains 2B | 1.6868 | 0.0172 | | |
| C3orf54 | chromosome 3 open reading frame 54 | 1.7694 | 0.0175 | | |
| KLF13 | Kruppel-like factor 13 | 1.669 | 0.0183 | | |
| MNT | MAX binding protein | 1.678 | 0.0185 | | |
| ELOVL5 | ELOVL family member 5, elongation of long chain fatty acids (FEN1/Elo2, SUR4/Elo3-like, yeast) | 1.6902 | 0.0209 | | |
| TOB1 | transducer of ERBB2, 1 | 1.7352 | 0.0218 | | |
| MYO1D | myosin ID | 1.7072 | 0.0224 | | |
| ELOVL6 | ELOVL family member 6, elongation of long chain fatty acids (FEN1/Elo2, SUR4/Elo3-like, yeast) | 1.6842 | 0.0306 | | |
| INO80C | INO80 complex subunit C | 1.6071 | 0.0309 | | |
| DNAJB9 | DnaJ (Hsp40) homolog, subfamily B, member 9 | 1.6304 | 0.0309 | | |
| C13orf15 | chromosome 13 open reading frame 15 | 1.9172 | 0.0323 | | |
| C10orf58 | chromosome 10 open reading frame 58 | 1.5965 | 0.0328 | | |
| SELS | selenoprotein S | 2.3198 | 0.0329 | | |
| MFAP4 | microfibrillar-associated protein 4 | 1.9284 | 0.033 | | |
| ZCCHC14 | zinc finger, CCHC domain containing 14 | 1.6118 | 0.0349 | | |
| CKS2 | CDC28 protein kinase regulatory subunit 2 | 1.7607 | 0.0351 | | |

[1] Fold change in fibroblasts or HeLa-cells

[2] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

| | | Fibroblasts | | HeLa cells | |
|---|---|---|---|---|---|
| Symbol | Name | FC[1] | pfp[2] | FC[1] | pfp[2] |
| MYO10 | myosin X | 1.5884 | 0.0377 | | |
| PTGS2 | prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase) | 1.5625 | 0.038 | | |
| SLIT3 | slit homolog 3 (Drosophila) | 1.6017 | 0.038 | | |
| SAT1 | spermidine/spermine N1-acetyltransferase 1 | 1.4897 | 0.0381 | | |
| FRMD8 | FERM domain containing 8 | 1.6735 | 0.0392 | | |
| HIST2H2BE | histone cluster 2, H2be | 1.5447 | 0.0393 | | |
| SLC26A6 | solute carrier family 26, member 6 | 1.5668 | 0.0415 | | |
| ZC3H12A | zinc finger CCCH-type containing 12A | 1.5342 | 0.0415 | | |
| ITGA11 | integrin, alpha 11 | 1.6568 | 0.0417 | | |
| IER5L | immediate early response 5-like | 1.5482 | 0.0419 | | |
| AKR1C3 | aldo-keto reductase family 1, member C3 (3-alpha hydroxysteroid dehydrogenase, type II) | 1.8076 | 0.0437 | | |
| RNF150 | ring finger protein 150 | 1.6551 | 0.0438 | | |
| DDIT4 | DNA-damage-inducible transcript 4 | 1.1515 | 0.0444 | | |
| HCFC1R1 | host cell factor C1 regulator 1 (XPO1 dependent) | 1.5468 | 0.0458 | | |
| GP1BB | glycoprotein Ib (platelet), beta polypeptide | 1.5687 | 0.0462 | | |
| IL8 | interleukin 8 | 1.0995 | 0.047 | | |
| NMB | neuromedin B | 1.5495 | 0.0473 | | |
| SNAI1 | snail homolog 1 (Drosophila) | 1.5163 | 0.0474 | | |
| FBXO32 | F-box protein 32 | 1.1042 | 0.0484 | | |
| MSX1 | msh homeobox 1 | 1.4492 | 0.0485 | | |
| CXXC5 | CXXC finger protein 5 | 1.5027 | 0.0495 | | |
| CPSF1 | cleavage and polyadenylation specific factor 1, 160kDa | | | 2.5922 | 0 |
| CYP51A1 | cytochrome P450, family 51, subfamily A, polypeptide 1 | | | 2.3338 | 1.00E-04 |
| CCNG2 | cyclin G2 | | | 2.3832 | 1.00E-04 |
| NFE2 | nuclear factor (erythroid-derived 2), 45kDa | | | 2.0575 | 3.00E-04 |
| TMLHE | trimethyllysine hydroxylase, epsilon | | | 2.1874 | 3.00E-04 |
| NSDHL | NAD(P) dependent steroid dehydrogenase-like | | | 2.038 | 3.00E-04 |

[1] Fold change in fibroblasts or HeLa-cells

[2] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

| | | Fibroblasts | | HeLa cells | |
|---|---|---|---|---|---|
| Symbol | Name | FC[1] | pfp[2] | FC[1] | pfp[2] |
| SLCO2A1 | solute carrier organic anion transporter family, member 2A1 | | | 2.1154 | 3.00E-04 |
| FLVCR1 | feline leukemia virus subgroup C cellular receptor 1 | | | 2.0854 | 7.00E-04 |
| HSD17B7 | hydroxysteroid (17-beta) dehydrogenase 7 | | | 1.9727 | 0.0018 |
| FDFT1 | farnesyl-diphosphate farnesyltransferase 1 | | | 1.8615 | 0.0072 |
| TCP1 | t-complex 1 | | | 1.6476 | 0.0077 |
| EPDR1 | ependymin related protein 1 (zebrafish) | | | 1.6476 | 0.0077 |
| SNORA29 | small nucleolar RNA, H/ACA box 29 | | | 1.6476 | 0.0077 |
| GMNN | geminin, DNA replication inhibitor | | | 1.0727 | 0.0079 |
| C20orf20 | chromosome 20 open reading frame 20 | | | 1.0461 | 0.0095 |
| RASD1 | RAS, dexamethasone-induced 1 | | | 1.0021 | 0.0117 |
| PCYT2 | phosphate cytidylyltransferase 2, ethanolamine | | | 1.1331 | 0.017 |
| CDC16 | cell division cycle 16 homolog (S. cerevisiae) | | | 1.0945 | 0.0264 |
| EPR1 | effector cell peptidase receptor 1 (non-protein coding) | | | 1.0634 | 0.0275 |
| BIRC5 | baculoviral IAP repeat-containing 5 | | | 1.0634 | 0.0275 |
| FAM166B | family with sequence similarity 166, member B | | | 1.073 | 0.0286 |
| FADS2 | fatty acid desaturase 2 | | | 1.6762 | 0.0302 |
| FGFBP1 | fibroblast growth factor binding protein 1 | | | 1.6993 | 0.0318 |
| RORB | RAR-related orphan receptor B | | | 1.6022 | 0.0325 |
| RNF216L | ring finger protein 216-like | | | 1.0568 | 0.0333 |
| FAM189B | family with sequence similarity 189, member B | | | 1.6427 | 0.0364 |
| MVK | mevalonate kinase | | | 1.6572 | 0.0399 |
| SLC2A3P1 | solute carrier family 2 (facilitated glucose transporter), member 3 pseudogene 1 | | | 1.6067 | 0.0426 |
| STARD4 | StAR-related lipid transfer (START) domain containing 4 | | | 1.5488 | 0.0456 |

[1] Fold change in fibroblasts or HeLa-cells

[2] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

### *B Genes down-regulated under sterol-depleted conditions*

| | | Fibroblasts | | HeLa cells | |
|---|---|---|---|---|---|
| Symbol | Name | FC[1] | pfp[2] | FC[1] | pfp[2] |
| ID2 | inhibitor of DNA binding 2, dominant negative helix-loop-helix protein | 0.4843 | 0.0142 | 0.463 | 0 |
| CDC42EP3 | CDC42 effector protein (Rho GTPase binding) 3 | 0.614 | 0.0303 | 0.5798 | 0.0457 |
| CDC20 | cell division cycle 20 homolog (S. cerevisiae) | 0.3398 | 0 | | |
| ACTC1 | actin, alpha, cardiac muscle 1 | 0.4843 | 0.0167 | | |
| F3 | coagulation factor III (thromboplastin, tissue factor) | 0.4966 | 0.0192 | | |
| SPOCD1 | SPOC domain containing 1 | 0.4696 | 0.0194 | | |
| TUFT1 | tuftelin 1 | 0.5733 | 0.0202 | | |
| S100P | S100 calcium binding protein P | 0.4455 | 0.0207 | | |
| CRIP1 | cysteine-rich protein 1 (intestinal) | 0.5225 | 0.0208 | | |
| TSC22D2 | TSC22 domain family, member 2 | 0.5632 | 0.0209 | | |
| KRTAP1-5 | keratin associated protein 1-5 | 0.5769 | 0.0212 | | |
| NUSAP1 | nucleolar and spindle associated protein 1 | 0.5579 | 0.0215 | | |
| KRTAP1-1 | keratin associated protein 1-1 | 0.5745 | 0.0217 | | |
| NTF3 | neurotrophin 3 | 0.5795 | 0.0218 | | |
| GLIPR1 | GLI pathogenesis-related 1 | 0.5703 | 0.0218 | | |
| GAGE5 | G antigen 5 | 0.5047 | 0.0221 | | |
| PBK | PDZ binding kinase | 0.5954 | 0.0224 | | |
| CCNB2 | cyclin B2 | 0.5427 | 0.0226 | | |
| TRIP13 | thyroid hormone receptor interactor 13 | 0.5103 | 0.0226 | | |
| KLF2 | Kruppel-like factor 2 (lung) | 0.5312 | 0.0228 | | |
| PTTG1 | pituitary tumor-transforming 1 | 0.5504 | 0.0229 | | |
| KIF20A | kinesin family member 20A | 0.5582 | 0.0233 | | |
| LCE2B | late cornified envelope 2B | 0.5145 | 0.0233 | | |
| IQGAP3 | IQ motif containing GTPase activating protein 3 | 0.531 | 0.0236 | | |
| CENPF | centromere protein F, 350/400kDa (mitosin) | 0.5962 | 0.0238 | | |
| AXL | AXL receptor tyrosine kinase | 0.5541 | 0.0239 | | |
| TOP2A | topoisomerase (DNA) II alpha 170kDa | 0.4787 | 0.024 | | |
| PRC1 | protein regulator of cytokinesis 1 | 0.5529 | 0.0242 | | |
| GAGE4 | G antigen 4 | 0.5159 | 0.0244 | | |

[1] Fold change in fibroblasts or HeLa-cells
[2] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

| | | Fibroblasts | | HeLa cells | |
|---|---|---|---|---|---|
| Symbol | Name | FC[1] | pfp[2] | FC[1] | pfp[2] |
| TPM1 | tropomyosin 1 (alpha) | 0.4236 | 0.0251 | | |
| PMP22 | peripheral myelin protein 22 | 0.4008 | 0.0274 | | |
| NGF | nerve growth factor (beta polypeptide) | 0.5006 | 0.0293 | | |
| LCE3A | late cornified envelope 3A | 0.586 | 0.0304 | | |
| GAGE2C | G antigen 2C | 0.5241 | 0.0312 | | |
| KPRP | keratinocyte proline-rich protein | 0.5925 | 0.0317 | | |
| GAGE2E | G antigen 2E | 0.5355 | 0.0328 | | |
| ASPM | asp (abnormal spindle) homolog, microcephaly associated (Drosophila) | 0.6177 | 0.0329 | | |
| COL4A1 | collagen, type IV, alpha 1 | 0.6171 | 0.0332 | | |
| CEP55 | centrosomal protein 55kDa | 0.6256 | 0.0334 | | |
| OXTR | oxytocin receptor | 0.6168 | 0.0336 | | |
| DUSP1 | dual specificity phosphatase 1 | 0.5624 | 0.0336 | | |
| TIPARP | TCDD-inducible poly(ADP-ribose) polymerase | 0.5414 | 0.034 | | |
| TK1 | thymidine kinase 1, soluble | 0.626 | 0.0341 | | |
| CA12 | carbonic anhydrase XII | 0.6156 | 0.0345 | | |
| GAGE12I | G antigen 12I | 0.5524 | 0.0381 | | |
| CPS1 | carbamoyl-phosphate synthase 1, mitochondrial | 0.5324 | 0.0397 | | |
| GAGE6 | G antigen 6 | 0.527 | 0.04 | | |
| IL12A | interleukin 12A (natural killer cell stimulatory factor 1, cytotoxic lymphocyte maturation factor 1, p35) | 0.5829 | 0.0402 | | |
| PTX3 | pentraxin 3, long | 0.5853 | 0.0408 | | |
| HIST1H4C | histone cluster 1, H4c | 0.5847 | 0.0421 | | |
| RPL29 | ribosomal protein L29 | 0.6387 | 0.0428 | | |
| SKP2 | S-phase kinase-associated protein 2 (p45) | 0.5616 | 0.0443 | | |
| HERC5 | hect domain and RLD 5 | 0.5758 | 0.0457 | | |
| GADD45B | growth arrest and DNA-damage-inducible, beta | 0.6555 | 0.0465 | | |
| C4BPA | complement component 4 binding protein, alpha | 0.5507 | 0.0468 | | |
| DTL | denticleless homolog (Drosophila) | 0.659 | 0.0499 | | |
| ID1 | inhibitor of DNA binding 1, dominant negative helix-loop-helix protein | | | 0.4106 | 0 |
| LOC285556 | hypothetical protein LOC285556 | | | 0.6027 | 0 |

[1] Fold change in fibroblasts or HeLa-cells

[2] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

| | | Fibroblasts | | HeLa cells | |
|---|---|---|---|---|---|
| Symbol | Name | FC[1] | pfp[2] | FC[1] | pfp[2] |
| MCM5 | minichromosome maintenance complex component 5 | | | 0.9409 | 3.00E-04 |
| GBP1 | guanylate binding protein 1, interferon-inducible, 67kDa | | | 0.4885 | 4.00E-04 |
| PTGES3 | prostaglandin E synthase 3 (cytosolic) | | | 0.9244 | 5.00E-04 |
| LIMA1 | LIM domain and actin binding 1 | | | 0.556 | 0.0023 |
| IL7R | interleukin 7 receptor | | | 0.5309 | 0.0062 |
| CXCR2 | chemokine (C-X-C motif) receptor 2 | | | 0.6894 | 0.0117 |
| MORC4 | MORC family CW-type zinc finger 4 | | | 0.8871 | 0.0127 |
| MYPN | myopalladin | | | 0.5818 | 0.0132 |
| C6orf146 | chromosome 6 open reading frame 146 | | | 0.9768 | 0.0146 |
| MAML2 | mastermind-like 2 (Drosophila) | | | 0.5794 | 0.0157 |
| EDN2 | endothelin 2 | | | 0.5874 | 0.0195 |
| GBP3 | guanylate binding protein 3 | | | 0.5946 | 0.0276 |
| UBXN1 | UBX domain protein 1 | | | 0.7756 | 0.0442 |
| ARHGEF12 | Rho guanine nucleotide exchange factor (GEF) 12 | | | 0.8545 | 0.0445 |
| MRPS33 | mitochondrial ribosomal protein S33 | 0.9823 | 0.0454 | | |

[1] Fold change in fibroblasts or HeLa-cells

[2] Percentage of false positives (false discovery rate) in fibroblasts or HeLa-cells

**Table B.3. Comparison of /textitin silico predicted SREBP and NF-Y target genes with identified target genes from ENCODE [65] project.**

| | | SREBP | | NF-Y | |
|---|---|---|---|---|---|
| Symbol | Gene Name | Predicted* | ENCODE* | Predicted* | ENCODE* |
| HES6 | hairy and enhancer of split 6 (Drosophila) | x | x | x | x |
| CPSF1 | cleavage and polyadenylation specific factor 1, 160kDa | x | | x | x |
| CCNG2 | cyclin G2 | x | x | x | x |
| SLCO2A1 | solute carrier organic anion transporter family, member 2A1 | x | | | |
| BHLHE40 | basic helix-loop-helix family, member e40 | x | x | | x |
| FLVCR1 | feline leukemia virus subgroup C cellular receptor 1 | x | x | x | x |
| TMEM97 | transmembrane protein 97 | x | x | | x |

* x in columns 3 and 5 indicates predicted binding sites for SREBP and NF-Y and identified binding sites for SREBP and NF-Y from ENCODE in columns 4 and 6.

| | | SREBP | | NF-Y | |
|---|---|---|---|---|---|
| Symbol | Gene Name | Predicted* | ENCODE* | Predicted* | ENCODE* |
| HSD17B7 | hydroxysteroid (17-beta) dehydrogenase 7 | x | x | | x |
| INSIG1 | insulin induced gene 1 | x | | x | x |
| MVD | mevalonate (diphospho) decarboxylase | x | x | x | x |
| MAFB | v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian) | x | | x | |
| STC1 | stanniocalcin 1 | x | | x | x |
| IDI1 | isopentenyl-diphosphate delta isomerase 1 | x | x | x | x |
| FABP3 | fatty acid binding protein 3, muscle and heart (mammary-derived growth inhibitor) | x | | x | |
| KLF6 | Kruppel-like factor 6 | x | x | x | x |
| PDGFRB | platelet-derived growth factor receptor, beta polypeptide | x | | x | x |
| SCD | stearoyl-CoA desaturase (delta-9-desaturase) | x | x | | x |
| LPIN1 | lipin 1 | x | x | x | x |
| PFKFB4 | 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4 | x | | | |
| DHCR7 | 7-dehydrocholesterol reductase | x | x | x | x |
| RGS4 | regulator of G-protein signaling 4 | x | | x | |
| C17orf59 | chromosome 17 open reading frame 59 | x | x | | x |
| FDFT1 | farnesyl-diphosphate farnesyltransferase 1 | x | x | x | x |
| MXRA5 | matrix-remodelling associated 5 | x | | | |
| LDLR | low density lipoprotein receptor | x | x | | x |
| C20orf20 | chromosome 20 open reading frame 20 | x | | | x |
| GAS1 | growth arrest-specific 1 | x | | | |
| TP53INP2 | tumor protein p53 inducible nuclear protein 2 | x | x | x | x |
| RASD1 | RAS, dexamethasone-induced 1 | x | | | |
| DBC1 | deleted in bladder cancer 1 | x | | | |
| SLC2A6 | solute carrier family 2 (facilitated glucose transporter), member 6 | x | | | x |

* x in columns 3 and 5 indicates predicted binding sites for SREBP and NF-Y and identified binding sites for SREBP and NF-Y from ENCODE in columns 4 and 6.

| | | SREBP | | NF-Y | |
|---|---|---|---|---|---|
| Symbol | Gene Name | Predicted* | ENCODE* | Predicted* | ENCODE* |
| TMEM55B | transmembrane protein 55B | x | x | x | x |
| TP53INP1 | tumor protein p53 inducible nuclear protein 1 | x | x | x | x |
| KCNJ2 | potassium inwardly-rectifying channel, subfamily J, member 2 | x | | x | x |
| ANGPTL2 | angiopoietin-like 2 | x | | x | |
| HSD17B12 | hydroxysteroid (17-beta) dehydrogenase 12 | x | x | x | x |
| PDGFRA | platelet-derived growth factor receptor, alpha polypeptide | x | | | |
| SQLE | squalene epoxidase | x | x | x | x |
| FBLN1 | fibulin 1 | x | | | |
| PCYT2 | phosphate cytidylyltransferase 2, ethanolamine | x | | | x |
| C3orf54 | chromosome 3 open reading frame 54 | x | | | x |
| KLF13 | Kruppel-like factor 13 | x | | | x |
| FASN | fatty acid synthase | x | x | | x |
| MNT | MAX binding protein | x | | x | x |
| TOB1 | transducer of ERBB2, 1 | x | x | x | x |
| MYO1D | myosin ID | x | | | |
| EPR1 | effector cell peptidase receptor 1 (non-protein coding) | x | | | |
| BIRC5 | baculoviral IAP repeat-containing 5 | x | | x | |
| ELOVL6 | ELOVL family member 6, elongation of long chain fatty acids (FEN1/Elo2, SUR4/Elo3-like, yeast) | x | | | |
| DNAJB9 | DnaJ (Hsp40) homolog, subfamily B, member 9 | x | | | x |
| RORB | RAR-related orphan receptor B | x | | x | |
| ZCCHC14 | zinc finger, CCHC domain containing 14 | x | | x | x |
| FAM189B | family with sequence similarity 189, member B | x | x | | x |
| MYO10 | myosin X | x | | | |
| SLIT3 | slit homolog 3 (Drosophila) | x | | x | x |
| SAT1 | spermidine/spermine N1-acetyltransferase 1 | x | | x | x |
| FRMD8 | FERM domain containing 8 | x | x | x | x |

* x in columns 3 and 5 indicates predicted binding sites for SREBP and NF-Y and identified binding sites for SREBP and NF-Y from ENCODE in columns 4 and 6.

| | | SREBP | | NF-Y | |
|---|---|---|---|---|---|
| Symbol | Gene Name | Predicted* | ENCODE* | Predicted* | ENCODE* |
| MVK | mevalonate kinase | x | x | x | x |
| ZC3H12A | zinc finger CCCH-type containing 12A | x | x | | |
| SLC26A6 | solute carrier family 26, member 6 | x | x | x | x |
| IER5L | immediate early response 5-like | x | | x | x |
| SLC2A3P1 | solute carrier family 2 (facilitated glucose transporter), member 3 pseudogene 1 | x | | | |
| DDIT4 | DNA-damage-inducible transcript 4 | x | | x | x |
| HCFC1R1 | host cell factor C1 regulator 1 (XPO1 dependent) | x | x | | |
| GP1BB | glycoprotein Ib (platelet), beta polypeptide | x | | x | x |
| SNAI1 | snail homolog 1 (Drosophila) | x | x | | x |
| MSX1 | msh homeobox 1 | x | | x | x |
| CXXC5 | CXXC finger protein 5 | x | | x | |
| CDC20 | cell division cycle 20 homolog (S. cerevisiae) | x | | | x |
| LOC285556 | hypothetical protein LOC285556 | x | | | |
| PTGES3 | prostaglandin E synthase 3 (cytosolic) | x | | | |
| CXCR2 | chemokine (C-X-C motif) receptor 2 | x | | x | x |
| ID2 | inhibitor of DNA binding 2, dominant negative helix-loop-helix protein | x | x | | x |
| MAML2 | mastermind-like 2 (Drosophila) | x | | | |
| ACTC1 | actin, alpha, cardiac muscle 1 | x | | | |
| F3 | coagulation factor III (thromboplastin, tissue factor) | x | | x | |
| SPOCD1 | SPOC domain containing 1 | x | | | |
| EDN2 | endothelin 2 | x | | | |
| S100P | S100 calcium binding protein P | x | | | |
| CRIP1 | cysteine-rich protein 1 (intestinal) | x | | | |
| TSC22D2 | TSC22 domain family, member 2 | x | | x | |
| NTF3 | neurotrophin 3 | x | | | |
| TRIP13 | thyroid hormone receptor interactor 13 | x | | x | x |

* x in columns 3 and 5 indicates predicted binding sites for SREBP and NF-Y and identified binding sites for SREBP and NF-Y from ENCODE in columns 4 and 6.

| | | SREBP | | NF-Y | |
|---|---|---|---|---|---|
| Symbol | Gene Name | Predicted* | ENCODE* | Predicted* | ENCODE* |
| KLF2 | Kruppel-like factor 2 (lung) | x | | | |
| IQGAP3 | IQ motif containing GTPase activating protein 3 | x | | | |
| AXL | AXL receptor tyrosine kinase | x | | x | |
| PRC1 | protein regulator of cytokinesis 1 | x | x | | x |
| PMP22 | peripheral myelin protein 22 | x | | x | |
| GBP3 | guanylate binding protein 3 | x | | x | |
| CDC42EP3 | CDC42 effector protein (Rho GTPase binding) 3 | x | | | |
| ASPM | asp (abnormal spindle) homolog, microcephaly associated (Drosophila) | x | | x | x |
| CEP55 | centrosomal protein 55kDa | x | | x | x |
| DUSP1 | dual specificity phosphatase 1 | x | x | | |
| OXTR | oxytocin receptor | x | | | |
| TIPARP | TCDD-inducible poly(ADP-ribose) polymerase | x | x | | |
| CA12 | carbonic anhydrase XII | x | | x | x |
| RPL29 | ribosomal protein L29 | x | | | |
| UBXN1 | UBX domain protein 1 | x | | | |
| GADD45B | growth arrest and DNA-damage-inducible, beta | x | x | | x |

* x in columns 3 and 5 indicates predicted binding sites for SREBP and NF-Y and identified binding sites for SREBP and NF-Y from ENCODE in columns 4 and 6.

**Table B.4. Conservation of predicted SREBP binding sites in chimp and mouse.**

| Symbol | Gene Name | Predicted* | Chimp* | Mouse* |
|---|---|---|---|---|
| HES6 | hairy and enhancer of split 6 (Drosophila) | x | x | x |
| CPSF1 | cleavage and polyadenylation specific factor 1, 160kDa | x | x | x |
| CCNG2 | cyclin G2 | x | | x |
| SLCO2A1 | solute carrier organic anion transporter family, member 2A1 | x | x | x |
| BHLHE40 | basic helix-loop-helix family, member e40 | x | x | x |
| FLVCR1 | feline leukemia virus subgroup C cellular receptor 1 | x | | |
| TMEM97 | transmembrane protein 97 | x | x | x |

* x in column 3 indicates predicted binding sites for SREBP in human and in columns 4 and 5 conserved predicted SREBP binding sites in chimp and mouse, respectively.

| Symbol | Gene Name | Predicted* | Chimp* | Mouse* |
|---|---|---|---|---|
| HSD17B7 | hydroxysteroid (17-beta) dehydrogenase 7 | x | x | x |
| INSIG1 | insulin induced gene 1 | x | x | x |
| MVD | mevalonate (diphospho) decarboxylase | x | x | |
| MAFB | v-maf musculoaponeurotic fibrosarcoma oncogene homolog B (avian) | x | x | x |
| STC1 | stanniocalcin 1 | x | x | x |
| IDI1 | isopentenyl-diphosphate delta isomerase 1 | x | x | |
| FABP3 | fatty acid binding protein 3, muscle and heart (mammary-derived growth inhibitor) | x | x | x |
| KLF6 | Kruppel-like factor 6 | x | x | |
| PDGFRB | platelet-derived growth factor receptor, beta polypeptide | x | x | x |
| SCD | stearoyl-CoA desaturase (delta-9-desaturase) | x | x | x |
| LPIN1 | lipin 1 | x | x | x |
| PFKFB4 | 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 4 | x | x | x |
| DHCR7 | 7-dehydrocholesterol reductase | x | x | |
| RGS4 | regulator of G-protein signaling 4 | x | x | x |
| C17orf59 | chromosome 17 open reading frame 59 | x | x | |
| FDFT1 | farnesyl-diphosphate farnesyltransferase 1 | x | x | |
| MXRA5 | matrix-remodelling associated 5 | x | x | |
| LDLR | low density lipoprotein receptor | x | x | x |
| C20orf20 | chromosome 20 open reading frame 20 | x | x | x |
| GAS1 | growth arrest-specific 1 | x | x | |
| TP53INP2 | tumor protein p53 inducible nuclear protein 2 | x | x | x |
| RASD1 | RAS, dexamethasone-induced 1 | x | x | x |
| DBC1 | deleted in bladder cancer 1 | x | x | x |
| SLC2A6 | solute carrier family 2 (facilitated glucose transporter), member 6 | x | x | x |
| TMEM55B | transmembrane protein 55B | x | x | |
| TP53INP1 | tumor protein p53 inducible nuclear protein 1 | x | x | x |
| KCNJ2 | potassium inwardly-rectifying channel, subfamily J, member 2 | x | | |
| ANGPTL2 | angiopoietin-like 2 | x | x | |
| HSD17B12 | hydroxysteroid (17-beta) dehydrogenase 12 | x | x | x |
| PDGFRA | platelet-derived growth factor receptor, alpha polypeptide | x | x | x |

* x in column 3 indicates predicted binding sites for SREBP in human and in columns 4 and 5 conserved predicted SREBP binding sites in chimp and mouse, respectively.

| Symbol | Gene Name | Predicted* | Chimp* | Mouse* |
|---|---|---|---|---|
| SQLE | squalene epoxidase | x | x | |
| FBLN1 | fibulin 1 | x | x | |
| PCYT2 | phosphate cytidylyltransferase 2, ethanolamine | x | x | x |
| C3orf54 | chromosome 3 open reading frame 54 | x | x | x |
| KLF13 | Kruppel-like factor 13 | x | x | x |
| FASN | fatty acid synthase | x | x | x |
| MNT | MAX binding protein | x | x | x |
| TOB1 | transducer of ERBB2, 1 | x | x | x |
| MYO1D | myosin ID | x | x | |
| EPR1 | effector cell peptidase receptor 1 (non-protein coding) | x | x | |
| BIRC5 | baculoviral IAP repeat-containing 5 | x | x | x |
| ELOVL6 | ELOVL family member 6, elongation of long chain fatty acids (FEN1/Elo2, SUR4/Elo3-like, yeast) | x | x | x |
| DNAJB9 | DnaJ (Hsp40) homolog, subfamily B, member 9 | x | x | x |
| RORB | RAR-related orphan receptor B | x | x | x |
| ZCCHC14 | zinc finger, CCHC domain containing 14 | x | x | x |
| FAM189B | family with sequence similarity 189, member B | x | x | x |
| MYO10 | myosin X | x | | |
| SLIT3 | slit homolog 3 (Drosophila) | x | x | |
| SAT1 | spermidine/spermine N1-acetyltransferase 1 | x | x | x |
| FRMD8 | FERM domain containing 8 | x | x | x |
| MVK | mevalonate kinase | x | x | x |
| ZC3H12A | zinc finger CCCH-type containing 12A | x | x | x |
| SLC26A6 | solute carrier family 26, member 6 | x | x | x |
| IER5L | immediate early response 5-like | x | x | x |
| SLC2A3P1 | solute carrier family 2 (facilitated glucose transporter), member 3 pseudogene 1 | x | x | |
| DDIT4 | DNA-damage-inducible transcript 4 | x | x | x |
| HCFC1R1 | host cell factor C1 regulator 1 (XPO1 dependent) | x | x | x |
| GP1BB | glycoprotein Ib (platelet), beta polypeptide | x | x | x |
| SNAI1 | snail homolog 1 (Drosophila) | x | x | x |
| MSX1 | msh homeobox 1 | x | x | x |

* x in column 3 indicates predicted binding sites for SREBP in human and in columns 4 and 5 conserved predicted SREBP binding sites in chimp and mouse, respectively.

| Symbol | Gene Name | Predicted* | Chimp* | Mouse* |
|---|---|---|---|---|
| CXXC5 | CXXC finger protein 5 | x | x | x |
| CDC20 | cell division cycle 20 homolog (S. cerevisiae) | x | x | x |
| LOC285556 | hypothetical protein LOC285556 | x | x | x |
| PTGES3 | prostaglandin E synthase 3 (cytosolic) | x | x | |
| CXCR2 | chemokine (C-X-C motif) receptor 2 | x | x | |
| ID2 | inhibitor of DNA binding 2, dominant negative helix-loop-helix protein | x | x | x |
| MAML2 | mastermind-like 2 (Drosophila) | x | x | |
| ACTC1 | actin, alpha, cardiac muscle 1 | x | x | x |
| F3 | coagulation factor III (thromboplastin, tissue factor) | x | x | x |
| SPOCD1 | SPOC domain containing 1 | x | x | |
| EDN2 | endothelin 2 | x | x | x |
| S100P | S100 calcium binding protein P | x | x | |
| CRIP1 | cysteine-rich protein 1 (intestinal) | x | x | x |
| TSC22D2 | TSC22 domain family, member 2 | x | | x |
| NTF3 | neurotrophin 3 | x | x | x |
| TRIP13 | thyroid hormone receptor interactor 13 | x | x | x |
| KLF2 | Kruppel-like factor 2 (lung) | x | x | x |
| IQGAP3 | IQ motif containing GTPase activating protein 3 | x | x | x |
| AXL | AXL receptor tyrosine kinase | x | x | |
| PRC1 | protein regulator of cytokinesis 1 | x | x | |
| PMP22 | peripheral myelin protein 22 | x | x | |
| GBP3 | guanylate binding protein 3 | x | x | |
| CDC42EP3 | CDC42 effector protein (Rho GTPase binding) 3 | x | x | x |
| ASPM | asp (abnormal spindle) homolog, microcephaly associated (Drosophila) | x | x | x |
| CEP55 | centrosomal protein 55kDa | x | x | |
| DUSP1 | dual specificity phosphatase 1 | x | x | |
| OXTR | oxytocin receptor | x | x | x |
| TIPARP | TCDD-inducible poly(ADP-ribose) polymerase | x | x | x |
| CA12 | carbonic anhydrase XII | x | x | |
| RPL29 | ribosomal protein L29 | x | x | |
| UBXN1 | UBX domain protein 1 | x | x | |

* x in column 3 indicates predicted binding sites for SREBP in human and in columns 4 and 5 conserved predicted SREBP binding sites in chimp and mouse, respectively.

| Symbol | Gene Name | Predicted* | Chimp* | Mouse* |
|---|---|---|---|---|
| GADD45B | growth arrest and DNA-damage-inducible, beta | x | x | x |

* x in column 3 indicates predicted binding sites for SREBP in human and in columns 4 and 5 conserved predicted SREBP binding sites in chimp and mouse, respectively.

# Acknowledgments

Many people have contributed to this thesis and I would like to take the opportunity to thank them here.

I would like to thank Prof. Dr. Roland Eils for giving me the opportunity to work in the stimulating environment of his group, for his support, and for many scientific discussions. I am really grateful for the support PD Dr. Rainer König gave me during my time, for many motivating discussions, lots of ideas, and his continuous encouragement. I am also grateful for my second supervisor PD Dr. Stefan Wiemann for many inspiring ideas and discussions on my project and the support I received.

I would like to thank Prof. Dr. Ulrike Müller and Prof. Dr. Herbert Steinbeisser for agreeing to be my examiners on my defense.

A big thank you goes to former and current members of the Network modeling group, specifically Gunnar Schramm, Tobias Bauer, Richa Batra, Kitiporn Plaimas, Apichat Suratanee, Kannabiran Nandakumar, and Matthias Eidel, for the warm atmosphere and many (not always science related) discussions.

I also need to thank Karlheinz Großand Rolf Kabbe for an excellent infrastructure.

I am much obliged to Prof. Dr. Dr. med. Johannes Zschocke for initiating the project with Heiko Runz and many valuable discussions. Heiko Runz and Jessica Schilde deserve my gratitude for being excellent cooperation partners and for many thoughtful comments and critical discussions on our joint project. I would like to thank Jessica Schilde also for proof reading this thesis.

I would like to thank Daniela Richter for her support on graphic design software and for many interesting theories on life in Mystic Falls.

A special thank you goes to my grand-parents and especially my parents. Your believe in me, your support and love has given me the opportunity and strength to follow my own path and has carried (and still carries) me around the world.

My biggest thanks goes to Sebastian, for being my light, my rock, for believing in me, and for your endless support particularly during the last weeks.

# Erklärung

Ich versichere, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 10.02.2011

. . . . . . . . . . . . . . . . . . . .
(Anna-Lena Kranz)