International
Max Planck Research School
for Computational Biology
and Scientific Computing

Freie Universität Berlin

Prediction of transcription factor co-occurrence using rank based statistics

Alena van Bömmel

Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Betreuer: Prof. Dr. Martin Vingron

Berlin, July 2, 2014

Erstgutachter: Prof. Dr. Martin Vingron

Zweitgutachter: Prof. Dr. Anne-Laure Boulesteix

Tag der Disputation: 27. Februar 2015

Acknowledgements

To accomplish this thesis was only possible thanks to the support of many people, which I would like to aknowledge.

Foremost, I would like to thank my supervisor Martin Vingron for his scientific support, wise advice and for giving me the opportunity to be part of his group with exceptional scientists in a stimulating and friendly environment. I would like to thank my second reviewer Anne-Laure Boulesteix and the members of my PhD committee Sebastiaan Meijsing and Stefan Haas, for their valuable feedback. I also want to thank the International Max Planck Research School for Computational Biology and Scientific Computing for the funding I received during the time I was working on my PhD thesis and for funding my research stay in Melbourne. I am very thankful to both coordinators of the school, Hannes Luz and Kirsten Kelleher for their help and organization of all IMPRS events. I also thank Terry Speed for his valuable comments and scientific discussions during my time as a visiting scientist at the WEHI Institute in Melbourne. I wish to thank Morgane Thomas-Chollier for teaching me patiently all basics about transcription factors. I am grateful to my former office mates Akdes Serin, Jonathan Göke and Christian Rödelsperger for making my start in bioinformatics so smooth and for forming the most enjoyable PhD student office. I want to thank my more recent office mates Navodit Misra, Annalisa Marsico and Juliane Perner for the wonderful time in the shared office and for their helpfulness and exciting discussions. I wish to thank Peter Arndt, Sarah Behrens, Julia Lasserre, Ho-Ryun Chung, Mike Love and Anna Ramisch for their ideas and fruitful discussions that led to an improvement of my dissertation project. I am grateful to Alessandro Mammana and Matt Huska for their frequent help with scripts, inkscape and other programs. I want to thank Marius Tolzmann for his patient help with the system MARIUX and with all computer-related problems. Further, I wish to thank all former and current members of the Vingron department for creating a motivating environment and for all wonderful moments which we shared at parties, retreats and conferences. During my PhD time, I enjoyed every day at the institute, with all its scientific and social activities. I would like to thank Matt Huska, Mike Love, Juliane Perner and Florian van Bömmel for their proofreading which led to a significant improvement of my thesis.

Finally, my biggest gratitude goes to my family, my parents and my husband Florian for their inexhaustible support during my PhD time; in particular to Florian who had to learn a lot of statistics in the last years and who helped me with all important presentations and manuscripts.

Contents

List of Figures

List of Tables

1 Introduction

All living organisms are composed of cells, some of them of one cell, some of many millions. The higher organisms, such as mammals, consist of many millions cells which belong to hundreds of different cell types. Cells of different cell types may differ greatly in their morphology and function according to the tissue they form, see Figure 1.1. For example, the axons of the neuronal cells in human can be over one meter long (Figure 1.1a), the skeletal muscle cell can span tens of mm (Figure 1.1b), whereas the size of a white blood cell is about 7 μm in diameter (Figure 1.1c). Interestingly, the genetic information encoded in the nucleus of these cells is nearly identical. The reason for this is that all cells of an organism have the same origin from a single initial cell - the zygote, which is derived from the parental gametes. In every cell division, the genetic information of the zygote is passed to its daughter cells which is why all cells of an organism contain the same genetic information.

One of the key questions in molecular biology is how cells with the same genetic code differentiate to this large variety of cell types. Particularly, the differentiation of the cell is strongly controlled through the regulation of gene expression - a cellular mechanism when only specific part of the genetic information is active such that only specific gene products are generated. The main factors of the gene regulatory mechanisms are the cellular environment, DNA acces-

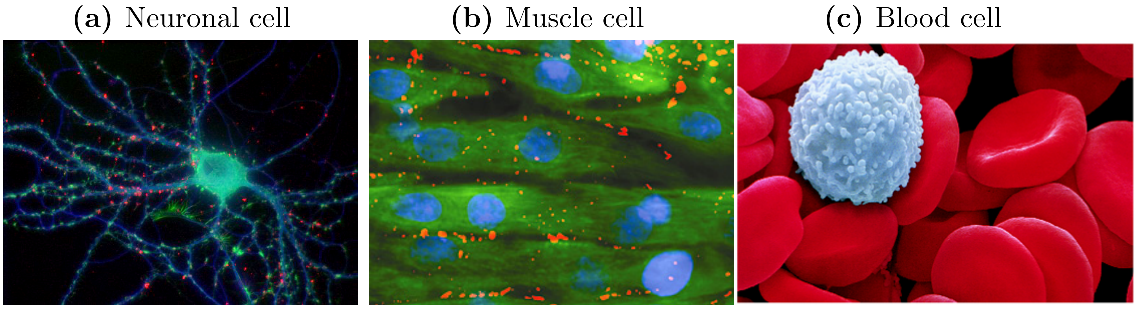


Figure 1.1: Human cells of three different cell types: (a) neuronal cell from www.wadsworth.org, (b) muscle cell from www.sciencemuseum.org.uk and (c) blood cells from www.just-health.net.

sibility and specific proteins controlling the transcription, called transcription factors (Coller and Kruglyak, 2008).

Transcription factors (TFs) recognize regulatory signals in the DNA sequence and bind to the DNA molecule to control the expression of their target genes. However, transcription factors usually do not act alone but they interact - directly or indirectly - with specific partner TFs. This combinatorial cooperation of TFs is critical for the regulation of gene transcription to achieve cell type specificity and developmental level of the cell (Remenyi et al., 2004; Vaquerizas et al., 2009). However, the experimental techniques that can detect the combinatorial cooperation of TFs on the DNA are very limited and they can assess only few proteins at the same time.

The aim of my thesis is to predict the cooperation of TF pairs for large number of various TFs on genomic regulatory regions using the sequence information and available information on binding specificity of those TFs. TFs are represented by ranked lists of the target sequences ordered by their binding affinities, thus several rank based statistics can be applied to detect significant associations between TF pairs.

1.1 Outline

Chapter 1: Introduction

In the first chapter, the basics of regulation of gene expression in eukaryotes are explained. In particular, TFs and the experimental and computational methods for measuring their binding affinity to the DNA are discussed. Furthermore, the importance of cooperative regulation of transcription factors is presented together with the possible approaches how to detect cooperativity.

Chapter 2: Rank based measures

In this chapter, the benefits of the scale and transformation free data representation via ranked lists are presented. Then, the most common rank based association measures for pairs of ranked lists are presented together with their applications to a simple example which depicts different properties of these measures.

Chapter 3: Prediction of transcription factor co-occurrence on human promoters

In this chapter, we show that transcription factors can be represented as a ranked list of genomic locations sorted by their binding affinity to the genomic sequence. Hence, pairs of co-occurring transcription factors can be predicted using an association measure for two ranked lists. The rank based association measures from previous chapter are applied to determine significant associations of transcription factor pairs on human promoters. Because strong associations between transcription factors might occur due to similarity of their binding motifs we introduce motif similarity as a possible confounding factor.

Chapter 4: Tissue-specific co-occurrence of transcription factors in human promoters

In this chapter, the concept of detecting transcription factor co-occurrence in tissue-specific manner is discussed. Here, we look at the association between two ranked lists (variables) stratified by tissue (third variable). Formally spoken, we arrive from classical 2-way contingency tables to 3-way contingency tables by introducing a third dimension (e.g. tissue) in the table. The best underlying null model for testing in a 3-way contingency table is discussed and evaluated on the distribution of p-values. Then, co-occurring TF pairs in various human tissues are predicted and validated with known protein-protein interactions and compared with other computational methods. Further, detailed analysis of co-occurring transcription factors in liver, skeletal muscle and hematopoietic stem cells is presented. The results of this part of my thesis were published in 2012 (Myšičková and Vingron, 2012).

Chapter 5: Cell-type specific co-occurrence of transcription factors in genomic regulatory regions

In this chapter, the challenges of investigating large number of genomic regions for transcription factor co-occurrence are discussed. Studying a large number of genomic regions increases dramatically the length of the ranked lists and with it the universe size of the corresponding contingency table. In accordance to these findings, we developed a new method based on ratios of p-values obtained from hypergeometric tests. This method is demonstrated on transcription factor binding affinity predictions to a large number of genomic regions of open chromatin defined by the DNase I hypersensitive sites in 64 cell types. In addition, definition, location and sequence properties of cell-type-specific accessible genomic regions derived from DNase I sequencing are considered. One part of this chapter is devoted to discussion of the optimal selection of the thresholds which define the top-ranked items.

Chapter 6: Summary

This chapter provides a brief summary of my thesis.

1.2 Transcription factors and their binding affinity

1.2.1 Transcription factors

Transcription factors are proteins which activate or repress the transcription of their target genes. They are able to bind to specific DNA sequences in order to regulate gene expression in *cis*. Such *cis*-regulatory sequences can be located proximal to the transcription start site (promoters) or can be located thousands of base pairs away from the transcription start site (TSS) of the gene being regulated (enhancers).

The protein family of transcription factors is very large and divergent. Several thousands of genes within the human genome encode for these specific proteins (Vaquerizas et al., 2009). One of the distinct characters of transcription factors is that they have a DNA-binding domain that recognizes a short specific DNA sequence and bind to the DNA. These short DNA sequences are usually different for distinct transcription factors and are called transcription factor binding sites (TFBS). TFBS can be identified with several experimental methods. The traditional approaches such as DNase footprinting assay (Galas and Schmitz, 1978) and the Electrophoretic Mobility Shift Assay (EMSA, Fried and Crothers, 1981; Garner and Revzin, 1981), are able to measure the binding preferences for only few sequences at a time. The recently developed methods apply the modern sequencing methods or microarray readouts: Chromatin Immunprecipitation assays combined with microarray (ChIP-chip, Ren et al., 2000) or with sequencing (ChIP-seq, Johnson et al., 2007), SELEX SAGE (Roulet et al., 2002) and Protein Binding Microarrays (PBMs, Bulyk, 2006). A comprehensive review of these methods was published by Xie et al. (2011).

1.2.2 Position Weight Matrix

Binding of transcription factors to the DNA is to some extent stochastic and depends on the biophysical properties of the DNA-binding domain and of the DNA sequence (Berg and von Hippel, 1987; Roider et al., 2007). As a consequence, every transcription factor is able to bind not only to a single DNA sequence but to a variety of DNA sequences that share a core structure. These patterns of transcription factor binding for a single factor can be represented with a position weight matrix (PWM), which summarizes the nucleotide counts of each position in the motif (Berg and von Hippel, 1987; Stormo et al., 1982). PWMs are usually derived from a set of aligned sequences that were experimentally found to be functionally related to transcription factor binding. A PWM has row entries for each symbol of the DNA alphabet (nucleotides A=adenine, C=cytosine, G=guanine and T=thymine) and column entries for each position in the pattern. In the first step, a position-specific frequency matrix (PSFM) is created. The cell entries of a PSFM are calculated as relative frequencies of each nucleotide at each position in the pattern. Formally, given a set S of N aligned sequences of length l, the entries of the PSFM denoted as $F(4 \times l)$, are as follows:

$$F_{k,j} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(S_{i,j} = k), \quad j \in (1, \ldots, l),\ k \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\} \tag{1.1}$$

where $\mathbb{1}$ is the indicator function taking the value of one if the condition is fulfilled and the value of zero else.

In the next step, the values of a PWM, denoted as $M(4 \times l)$, are calculated as log odds score of the corresponding PSFM entries and frequencies of a background model π_b as follows:

$$M_{k,j} = \log\left(\frac{F_{k,j}}{\pi_b(k)}\right), \quad j \in (1, \ldots, l),\ k \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}. \tag{1.2}$$

For the background model, one use a simple model with the uniform distribution over all nucleotides or a more sophisticated model reflecting the higher frequency of G and C in the regulatory regions; e.g. $\pi_b = (0.2, 0.3, 0.3, 0.2)$.

Usually, the number of experimental derived TFBS is very limited such that we can not observe all rare occurrences of some nucleotides at some positions. For this reason, the *pseudocounts* are applied in order to avoid PSFM entries having a value of 0 (resulting in PWM entries of $-\infty$). Then, to each each column of the PSFM a weighted background distribution is added to account for the non-observed TFBSs (Rahmann et al., 2003).

PWMs can be used to score genomic sequences being a functional binding site or being a random site. Since the PWM model assumes statistical independence of the matrix columns, the probabilities (or log odds scores) of the different positions can be treated as independent. Then, the likelihood score $L(S')$ of a new sequence S' to come from a functional binding site or from a background distribution can be calculated simply as the sum of the log odds scores in matrix M corresponding to nucleotides in sequence S'. If $L(S') > 0$, it is more likely that S' is a functional site than a random site, and if $L(S') < 0$, it is more likely that S' is a random site than a functional site. The sequence score can also be interpreted in a physical framework as the binding energy for that sequence.

A useful information of a PWM is its information content (IC) which gives an information how different a given PWM is from a background distribution. The IC is calculated from the frequencies in the PSFM matrix F as follows:

$$IC = -\sum_j \sum_k F_{k,j} \cdot \log_2(F_{k,j}). \tag{1.3}$$

Very popular is the visual representation of PWMs with sequence logos, which show the information content of each nucleotide at each position of the matrix. In such a graphic, a sequence logo shows stacked nucleotide symbols of heights proportional to their information content at the respective position (Schneider and Stephens, 1990). The information content for a matrix position j is calculated as a difference of the IC of the matrix column j and of the IC of a (uniform) background model:

$$\begin{aligned} IC(j) &= \sum_k \pi_b(k) \cdot log_2(\pi_b(k)) - \sum_k F_{k,j} \cdot \log_2(F_{k,j}) = \\ &= -2 - \sum_k F_{k,j} \cdot \log_2(F_{k,j}), \quad k \in \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}. \end{aligned} \tag{1.4}$$

The higher the preference of a nucleotide is in the PSFM at a given position, the higher is the IC at that position and the higher is the corresponding letter at that position in the sequence logo. Using the uniform background in the Eq. 1.4, the maximum information content at each position is 2 bits.

The PSFMs, PWMs and the DNA-binding sequences are stored in several databases such as TRANSFAC (Matys et al., 2006), JASPAR (Mathelier et al., 2014) and UniProbe (Newburger and Bulyk, 2009).

Example 1. *Figure 1.2 illustrates an example for calculating the PSFM, PWM and the motif logo for the nuclear receptor NR4A2. First, let us consider a set of $N = 13$ experimentally derived bound sequences of length $l = 8$ nucleotides as given in Figure 1.2a. Then the PSFM matrix with 4 rows and 8 columns can be easily calculated from the occurrences of the nucleotides at each position of the sequence, see Figure 1.2b. Using a simple background model with a uniform distribution: $\pi_b = (1/4, 1/4, 1/4, 1/4)^T$ and pseudocount weight of 0.9 for the PSFM and 0.1 for the background distribution gives the PWM shown in Figure 1.2c. The corresponding motif logo is shown in Figure 1.2d. Here, position 7 has an exclusive preference for one nucleotide only and has an information content of 2bits. Positions 2, 3 and 5 have a high preference for only one* *nucleotide and its information content is close to 2. On the other hand, positions with a non-specific distribution of nucleotides like position 1 and 8 have a small information content.*

Let us consider two new sequences: $S'_1 = \texttt{CAGGACAC}$ *and* $S'_2 = \texttt{TGGATTAT}$. *What is the likelihood that these sequences are functional binding sites or that they are just background sequences? To answer this question, we can calculate the likelihood score as a simple sum of the corresponding nucleotide entries on corresponding positions as:* $L(S'_1) = -0.79 + 1.15 + 1.15 + 0.98 - 0.39 + 1.15 + 1.22 + 0.78 = 5.25$; $L(S'_2) = -0.79 - 0.79 + 1.15 - 0.07 + 0.88 - 0.79 + 1.22 - 1.61 = -0.8$. *Thus the first sequence* $S'_1 = \texttt{CAGGACAC}$ *with* $L(S'_1) > 0$ *is more likely to be a functional binding site whereas the second sequence* $S'_2 = \texttt{TGGATTAT}$ *with* $L(S'_2) < 0$ *comes from a background distribution.*

1.2.3 Computational prediction of TF binding affinity

The binding affinity of a transcription factor to a DNA sequence can be estimated from the similarity of the binding motif (usually represented as a PWM) and the DNA sequence. However, genome wide prediction of the transcription factor binding sites is not trivial, since many binding sites can be found in the genome, approximately every 1000 bp. Thus, most of the predicted binding sites are false positive findings and only a minority of them are bound by the transcription factor *in vivo*. One reason for the small specificity of sequence based approaches is that they are cell type independent and do not have any information about the accessibility of the DNA in the cell type. Also, other factors such as co-factor binding and protein-protein interactions, have an influence on the cell-type-specific binding of transcription factors in the cell. Nevertheless, sequence based estimation of transcription factor binding affinities is a widely used approach to receive an information about the sequence of interest or to find the *cis*-regulatory sequences. Further, the theoretical approaches can be combined with experimental data which can limit the sequence space for searching of the binding sites (Pique-Regi et al., 2011).

Hit-based methods

The majority of the sequence based prediction methods are so called *hit-based* methods. Using a PWM (or PSFM) representation of the preferential transcription factor binding, a score based on the similarity to the binding motif is assigned to every site in the sequence of interest. Then, statistical approaches are used to find a significance of the hits compared to the background sequences and to define a threshold for a binary separation which report whether the site is a hit or not. As a result, one gets usually a list of sites

(a)

site	bound sequence
1	AAGGTCAC
2	AAGGTCAG
3	CAGAACAC
4	AAGGTCAC
5	AAGGCTAC
6	AAGGCCAG
7	AGGGTCAC
8	AATATCAC
9	TAGGTCAA
10	GAGAACAC
11	AAGGTCAA
12	GAGGTCAC
13	GAGGTCAA

(b) Position-specific frequency matrix

pos	1	2	3	4	5	6	7	8
A	0.62	0.92	0.00	0.23	0.15	0.00	1.00	0.23
C	0.08	0.00	0.00	0.00	0.15	0.92	0.00	0.62
G	0.23	0.08	0.92	0.77	0.00	0.00	0.00	0.15
T	0.08	0.00	0.08	0.00	0.69	0.08	0.00	0.00

(c) Position weight matrix with log odds scores

pos	1	2	3	4	5	6	7	8
A	0.78	1.15	-1.61	-0.07	-0.39	-1.61	1.22	-0.07
C	-0.79	-1.61	-1.61	-1.61	-0.39	1.15	-1.61	0.78
G	-0.07	-0.79	1.15	0.98	-1.61	-1.61	-1.61	-0.39
T	-0.79	-1.61	-0.79	-1.61	0.88	-0.79	-1.61	-1.61

(d) Sequence logo

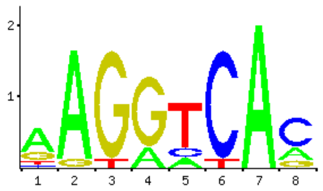


Figure 1.2: Experimentally derived bound sequences (a) with corresponding position-specific frequency matrix (PSFM) in (b), position weight matrix (PWM) in (c) and sequence logo (d).

which exceed the chosen threshold with their corresponding p-values reflecting the significance of the similarity. Some of the most used tools with integrated web services are RSAT (Thomas-Chollier et al., 2011), PAP (Chang et al., 2006), balanced method for false positive predictions (Rahmann et al., 2003) and binding energy estimation (Zhao et al., 2009).

Affinity based methods

The affinity-based approaches for transcription factor binding (Roider et al., 2007; Tanay, 2006) provides a measure of relative affinity of a transcription factor to a sequence of interest. Particularly, for our further analysis of co-occurrence of transcription factor binding, the *TRanscription factor Affinity Prediction (TRAP)* tool from Roider et al. is used because it has several advantages over the traditional hit-based methods. In comparison to the hit-based methods, TRAP does not require any threshold for binary separation, but integrate the contributions from individual sites in the sequence of interest to calculate the expected number of binding sites. More interestingly, TRAP provides a natural ranking of studied sequences with respect to the particular binding motif of interest.

TRAP is a probabilistic framework using a biophysical model of protein-DNA binding interaction inspired from the findings of Berg and von Hippel (1987). As a measure of relative affinity for a given PWM to a DNA sequence of interest, TRAP predicts the expected number of bound transcription factor molecules. This measure is calculated as the total contribution from all possible sites in the given DNA sequence.

1.3 Transcription factor cooperation

As other proteins in the organism, transcription factors function in cooperation with other transcription factors and proteins to regulate the spatial and temporal expression patterns of genes. The combinatorial regulation of gene expression increases the variability and flexibility of the regulatory mechanism, the organism can realize a large variety of transcriptional responses already with a small number of different transcription factors (Remenyi et al., 2004; Vaquerizas et al., 2009). Most of the transcription factors contain - in addition to the DNA-binding domain - a trans-activating or interaction domain, which allow the factor to interact with other co-factors. Transcription factors that build such complexes (called dimers for pairs of transcription factors) usually bind to the DNA in a close vicinity. Then their binding sites lie within a close distance on the DNA sequence and we can say that these transcription factors co-occur on the genomic sequence. Some of the interacting partners or dimers are well known and has been shown in various experiments (Chen et al., 1998, 2008a; Eriksson and Wrange, 1990; Wang et al., 2011). Some of the factors can interact with a factor of the same kind. These interaction are called homotypic interactions or homodimers. One of the well known examples is the glucocorticoid receptor (GR) which binds to the DNA as a homodimer (Eriksson and Wrange, 1990). Another example of a heterotypic interaction, e.g. factor interacting with a factor of another type, is the interaction of homeobox OCT4 and NANOG homeobox which was studied in embryonic stem cells (Chen et al., 2008a). Beyond that, DNA-binding transcription factors can interact with other transcription factors in an indirect way through other mediator proteins. These interactions are not easy to detect, neither with experimental nor with computational methods, since the interacting partners do not have to bind directly to the DNA.

1.3.1 Experimental methods to detect TF-TF interactions

The majority of the experimental methods which provide an evidence about interacting transcription factors are set up for observation of general protein-protein interactions (PPIs). In general, the detection of interactions between transcription factors is more challenging than the interactions between other proteins. Transcription factors are usually lowly expressed in the cell and have to be artificially expressed for the validation experiment. For some transcription factors, this process has very low efficiency, thus not for all interactions between TFs a testing is technically feasible. In addition, methods for detecting PPIs do not account for the potential transcriptional activity or expression of the transcription factors, thus they can provide only a partial information about the TF-TF interactions. In the following, the most popular approaches for detecting PPIs are presented.

Two-hybrid screening

In the two-hybrid screening (known as yeast two-hybrid system Y2H) an activation of downstream reporter gene by the binding of a specific transcription factor (usually GAL4) on an upstream sequence is tested. The specific transcription factor is split into two fragments, a DNA-binding domain and an activating domain. Then, two proteins which are screened for the PPI, called Bait and Prey, are fused with the DNA-binding domain (Bait) and with the activation domain (Prey), respectively. Usually, none of the two fused proteins is able to initiate the transcription of the reporter gene alone. Only if Bait and Prey interact, they gather the DNA-binding domain and the activating domain of the specific protein together and therewith initiate the transcription of the reporter gene. In the typical yeast two-hybrid system, which was introduced by Fields and Song (1989), separate Bait and Prey plasmids are introduced into the mutant yeast strain. Suzuki et al. (2001) developed a high-throughput assay for systematic analysis of PPIs based on the mammalian two-hybrid method (M2H), where Bait and Prey are transfected into mouse cells. Two-hybrid screening methods for transcription factors have a low sensitivity estimated to 25% and a moderate precision of 53% (Ravasi et al., 2010).

Protein crosslinking

Protein crossliniking is an approach where lysine residues of proteins form covalent crosslinks by oxidation. Further variations include other, less selective reagents which allow crosslinking independent of the sequence of the binding domain. Chemical crosslinking is often used to detect low-affinity or weak protein interactions (Berggård et al., 2007).

Mass spectroscopy

With mass spectroscopy (MS), introduced by Aebersold and Mann (2003), protein complexes can be identified directly. The basic principle of the MS method is to convert (unidentified) protein molecules in the condensed phase into ions using electrospray ionization (ESI) or matrix assisted laser desorption ionization (MALDI), which can be distinguished based on their mass-to-charge ratios. These mass spectra measurements then allow the determination of the polypeptide sequences using some of the large variety of algorithms (Aebersold and Mann, 2003). MS is not limited for binary protein interactions only, it can detect multi-protein complexes. The limiting step of the MS is the purification of protein complexes, which is not efficient for all proteins (or transcription factors) of interest (Shoemaker and Panchenko, 2007). To address this problem, Tandem Affinity Purification (TAP, Rigaut et al. (1999)) for rapid purification of protein complexes in natural conditions was developed.

Co-immunoprecipitation

In co-immunoprecipitation (co-IP), one protein of interest (target) is bound by a protein-specific antibody in a sample (usually cell lysate) and precipitated on a beaded support (Berggård et al., 2007). All proteins not precipitated on the beads are washed away, so that proteins bound to the target protein can be captured. These protein complexes can then be analyzed by Western blot or immuno-detection. Whereas the approaches mentioned above are used for screening of interactions between proteins, co-IP is a method to verify a specific interaction between two proteins.

1.3.2 Other experimental methods used for TF analysis

Recent large studies of functional elements in the genome (Bernstein et al., 2010; The ENCODE Project consortium, 2012) enable the analysis of regulatory mechanisms and transcription factor cooperation in cell-type-specific manner using a large collection of distinct experiments. For example, the Chromatin-Immunprecipitation approach combined with sequencing (ChIP-seq) determines genome wide binding information of a single transcription factor of interest in a given cell type. However, combining a large collection of such experiments for various transcription factors in different cell types can provide an information about the co-association patterns between different transcription factors (Gerstein et al., 2012). Furthermore, the DNase I digestion followed by sequencing (DNase-seq) experiments can measure genome wide chromatin accessibility in a cell type of interest (Boyle et al., 2008a). Binding of a transcription factor to the DNA protects the underlying sequence from cleavage by DNase I, which leaves so called 'footprint' recognizable in the sequencing experiment. This information can be combined with known binding preferences of various transcription factors to detect associative binding of these factors (Neph et al., 2012). In the following, we briefly introduce these two popular high-throughput methods.

Chromatin-immunoprecipitation assays

Chromatin immunoprecipitation (ChIP) is a widely used method that characterizes interaction of a protein and a DNA sequence *in vivo* (Johnson et al., 2007; Mardis, 2007). First, proteins directly binding DNA molecules are covalently crosslinked with the DNA using formaldehyde or other chemicals. Then, the chromosomal DNA is fragmented and the protein of interest with all crosslinked DNA fragments is immunoprecipitated using a protein-specific antibody. Finally, the bound DNA fragments can be labeled and hybridized to microarrays (ChIP-chip) or identified with massive paralell sequencing techniques (ChIP-seq).

Massive parallel sequencing (or next-generation sequencing, NGS) approaches use parallel platforms for sequencing of short DNA sequences (40-500 bp), which are called reads. First, DNA sequencing libraries are generated by clonal amplification. Then, the DNA is sequenced by synthesis, e.g. by the addition of nucleotides to the complementary strand. The sequencing of the DNA templates is processed in a massively parallel fashion resulting in many million short reads per instrument run. The obtained reads are then mapped against a reference genome. In order to obtain binding locations of the transcription factor, the mapped reads are analyzed with various algorithms to detect genomic regions enriched in reads, which correspond to loci in the genome bound by the transcription factor.

DNase I hypersensitivity assays

Large majority of genomic DNA is densely packed by wrapping around protein complexes called nucleosomes. Only a small part of the DNA is accessible for other proteins to bind, this regions have usually regulatory functions for transcriptions (e.g. promoters and enhancers). The accessible regions can be easily digested by the enzyme DNase I and are called DNase I hypersensitive sites (DHSs). A high-throughput approach which simultaneously identifies thousands of DHSs in an unbiased manner was developed by Crawford et al. (2006a,b). First, the chromatin is digested with a small amount of DNase I that preferentially cuts at a DNase I hypersensitive site. Then, a linker sequence is attached to the DNase I-digested ends which is then used for extraction of short adjacent DNA fragments. The DNA fragments can be labeled and hybridized on tiled arrays (DNase-chip, Crawford et al. (2006a)) or identified with next generation sequencing techniques (DNase-seq, Crawford et al. (2006b)).

In order to detect DHSs in DNase-seq experiments, obtained reads are first mapped against a reference genome. The hypersensitive sites in the genome are enriched in sequenced reads, thus various algorithms can be applied to detect these sites.

1.4 Conclusion

Cooperativity among transcription factors is essential for spatial and temporal gene expression of the cell. As discussed above, the experimental detection of co-regulating transcription factors is rather difficult, non-sensitive and unfeasible for some proteins. For this reason, computational and statistical methods combined with other experimental measurements can help to reveal more insight into cell-type-specific gene regulation. Our aim is to provide a statistical approach for prediction of co-occurring transcription factors on genomic regulatory regions which is based on a rank based representation of a transcription factor as a list of its target regions, in the following three chapters.

2 Rank based association measures

2.1 Ranked list representation of genomic data

Due to the rapid development of experimental techniques in molecular biology over the past few decades, large amounts of high-dimensional data are now available. Several of these experimental techniques are able to analyse thousands of items in one single experiment. For example, a typical ChIP-seq experiment measures binding of a protein of interest to hundreds of thousand genomic locations at a time or in a microarray study, mRNA expression profiles for tens of thousands of genes are generated from a collection of samples. Thus the challenge for scientists is now to integrate and prioritize massive amounts of information to gain insights into biological mechanisms (Aerts et al., 2006; Subramanian et al., 2005).

One of the most popular ways to integrate experimental data is to represent the results of each experiment as a prioritized or **ranked list** based on some (measurement derived) statistics or significance. In this case, all measured items are ordered by a chosen measure such that the items on the top of the list are the most relevant (informative) ones. For example, the genomic locations in a ChIP-seq study are ordered by the strength (or significance) of the protein binding. Or in a microarray study, the genes are ordered according to their differential expression between the classes.

A ranked list is a simple and natural way to represent data that is independent of scaling and of transformation. One of its advantages is that one can apply nonparametric statistics, where usually no assumptions on the underlying distribution are necessary. Due to the reliance on fewer assumptions, nonparametric methods are generally easier to use and more robust in comparison with parametric methods (Stuart et al., 2008).

A further advantage of the ranked list representation is that it simplifies the comparison of two ranked lists that might represent results of two different experiments based on different underlying scores. If the scores comes from two different experimental methods or from measurements provided in different laboratories, the direct comparison might be very problematic due to their differences in scaling, magnitudes etc. In these situations the comparison of two ranked lists using nonparametric tests for two populations or a large variety of other rank based association measures (see Section 2.3) can be applied. In many genomic studies, the results are based on **p-values** which assign a statistical significance to each studied item. To calculate the p-values, usually a statistical model with an underlying null hypothesis (and null distribution) is assumed and then the probabilities to obtain a result at random at least as extreme as the observed measurements are determined. Then the measurements that are the furthest from the null distribution and thereby with the smallest p-values are usually of interest. For example, in a microarray study, the p-value of a given gene assigns a probability that the differential expression between the classes can be obtained at random. However, although the p-values encode a lot of information, their comparison and interpretation can be difficult. One reason for this is our lack of knowledge about the underlying null distribution. Very often, we are not able to derive the true null model and as a consequence the significance assigned to some measurements might be overestimated or underestimated. Another reason for the difficulties with p-value interpretation is that most of the genomics studies perform a large number of simultaneous measurements (items) in a small number of samples. Then, for the identified significant measurements, one has to decide whether the findings are truly informative or just a consequence of multiple hypothesis testing or due to the variability introduced by measurement error. Although there are several methods correcting for multiple testing problem such as the Bonferroni correction (Miller, 1966; Simes, 1986) or Benjamini-Hochberg correction (Benjamini and Hochberg, 1995), there is no mathematical solution available for correcting the high intrinsic variability between individual measurements. Therefore, the ranked list representation (often based on the ordering according to p-values) is highly preferred in many genomics studies. For example, Subramanian et al. (2005) evaluated ranked lists of genes from microarray experiments for their biological interpretation, Boulesteix and Slawski (2009) used ranked lists of genes for aggregating results of different studies, Eden et al. (2007) discovered enriched sequence elements in ranked lists of sequences derived from ChIP-chip experiments and Tembe et al. (2009) proposed a framework to statistically compare ranked lists of candidates from different algorithms used in genome-wide association studies.

In most of the large-scale studies, the underlying rank information degrades with increasing ranks. In particular, the ranking at the bottom of the lists becomes more or less random. Thus, a researcher's interest usually focuses only on the top of the list, namely the **top-k items** with k smallest ranks. However, the identification of the best value for k is not straightforward and has to be set *a priori* or has to be derived from the data. The determination of the optimal threshold k was discussed in the literature by Schimek et al. (2012) and Hall and Schimek (2012). The selection of this parameter with an application on the transcription factor co-occurrence will be discussed more deeply in Sections 2.3.3 and 5.5.

The basic notation for ranked list representation is introduced in the next Section 2.2. Then, the most common rank based measures used for large experiments are discussed in Section 2.3. Namely, the following five association measures are described: Spearman's correlation (Section 2.3.1), Kendall's τ (Section 2.3.2), Fisher's exact test (Section 2.3.3), Restricted two-dimensional Kolmogorov-Smirnov score (Section 2.3.4) and Irreproducible discovery rate (Section 2.3.5). Further, the application of the rank based measures is demonstrated on a simple example with two ranked lists of p-values (see Example 2). The last section provides a short summary of this chapter.

2.2 Notation for ranked lists

In this section, we briefly summarize the mathematical notation of rankings and ranked lists which will be used throughout the whole thesis.

Let a set O of n objects be given and a quantity x which assigns some ordinal values to each object. Then $x(i)$ denotes the value of object i assigned by the quantity x, for all $i \in O$. The rank of object i according to quantity x, $r_x(i)$ is then defined as follows:

$$r_x(i) = \begin{cases} 1 + \sum\limits_{o \in O} \mathbb{1}[x(o) < x(i)] & \text{in case of increasing ordering} \\ 1 + \sum\limits_{o \in O} \mathbb{1}[x(o) > x(i)] & \text{in case of decreasing ordering,} \end{cases} \tag{2.1}$$

where $\mathbb{1}$ denotes the indicator function taking the value of one if the condition in the brackets is fulfilled and the value of zero else. In other words, rank of object i is defined as the number of objects with larger quantity x increased by one for sorting in an increasing order; or as the number of objects with smaller quantity x increased by one for sorting in a decreasing order. R_x stands for the complete ranked list of all n objects $o \in O$ ordered according to quantity x and $|R_x| = n$ denotes the length of the ranked list.

2.3 Association measures for two ranked lists

2.3.1 Spearman's correlation

Probably the most widely used rank based correlation measure is Spearman's correlation introduced by Spearman (1906). Already in 1906, Spearman recognized the "necessity of comparison by rank, as absolute measurements are not properly comparable with one another". Spearman's correlation coefficient is defined as Pearson's correlation coefficient between two ranked lists.

Formally, let $r_x(i)$ denote the rank of object i in the first ranked list R_x according to quantity x and $r_y(i)$ the rank of object i in the second ranked list R_y according to quantity y; and for their lengths: $|R_x| = |R_y| = n$. In case that there are no identical values within the scores x_i and no identical values within the scores y_i; so that there are no ties in the rankings, Spearman's correlation is defined as the normalized sum of squared differences between the ranks of the two lists subtracted from 1 (Kendall and Gibbons, 1990):

$$\rho_{Sp} = 1 - \frac{6\sum_{i=1}^{n}[r_x(i) - r_y(i)]^2}{n(n^2-1)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \,. \tag{2.2}$$

d_i denotes the ranking difference $d_i = r_x(i) - r_y(i),\ i = 1, \ldots, n$. The normalization and the subtraction from one is done to achieve the maximum correlation of 1 in case of identical rankings of both lists and minimum correlation of -1 for lists when the rankings are exactly the reverse of each other. Spearman's correlation ρ_{Sp} between two statistically independent random variables X and Y is expected to be zero.

The original definition of Spearman (1906) included absolute differences between the rankings instead of the squared differences as in Eq. (2.2). This coefficient, known as Spearman's 'footrule', is defined as follows:

$$R_{Sp} = 1 - \frac{3\sum_{i=1}^{n}|d_i|}{n^2-1} \tag{2.3}$$

Thus, the coefficient R_{Sp} does not fulfill the properties of a correlation coefficient. For two identical rankings, there holds $\sum_{i=1}^{n}|d_i| = 0$ and $R_{Sp} = 1$. However, for inverse rankings Spearman's footrule implies $R_{Sp} = -0.5$ for n odd and $R_{Sp} = -0.5(1 + \frac{3}{n^2-1})$ for n even. Moreover, R_{Sp} is less sensitive than ρ_{Sp}. For these reasons the use of

Spearman's correlation is preferred over Spearman's footrule (Kendall and Gibbons, 1990).

For some applications, measuring association between two truncated ranked lists after the top-k ranks might be of interest. In this case, we have to deal with incomplete rankings where some items with rank $r_x(i) \leq k$ in the first list will not be among the top-k items in the second list, so its ranks in the second list will be undefined. Similarly, some items ranked among the top-k in the second list might not be among the top-k items in the first list and its rank will be undefined in the truncated list. We denote the sets of these items as follows: $O_x = \{i : r_x(i) \leq k \ \& \ r_y(i) \text{ undefined}\}$ and $O_y = \{i : r_y(i) \leq k \ \& \ r_x(i) \text{ undefined}\}$; $\forall k \in \mathbb{N}$. We can modify Spearman's correlation for the incomplete rankings easily by setting: $r_y(i) := k + 1$, $\forall i \in O_x$ and $r_x(i) := k + 1$, $\forall i \in O_y$, as suggested by Schimek et al. (2012). Then, Spearman's correlation ρ_{Sp} can be calculated in the same manner as for the complete lists.

Example 2. *In the following example, we illustrate different behavior of the classical correlation coefficient and the rank based Spearman correlation coefficient on two different data samples which could be obtained from two different experiments studying 100 genes. We assume that one half of the genes is differentially expressed and should be reflected in the measurements with a significant p-value.*

Let us consider measurements x from the first experiment with corresponding p-values p_x arisen from a mixture of two uniform distributions. The first 50 p-values corresponding to a strong signal were sampled from the uniform distribution $U([0, 0.05])$. The next 50 p-values corresponding to a noise were sampled from the uniform distribution $U([0, 1])$, see histogram in Figure 2.1a. p_y-values of the second experiment with measurements y were simulated in such way that the first half is a quadratic function of the p-values from the strong signal from the first experiment and the second half was again randomly sampled from a uniform distribution $U([0, 1])$, see the distribution in Figure 2.1b. Thus the first 50 p-values obtained from both experiments have both small values smaller than 0.05 but the values in sample y are much smaller than the values in sample x. The remaining 50 values are randomly distributed between 0 and 1. The relationship between p_x and p_y is depicted in the scatterplot in Figure 2.1c. However, when we look at the rankings of the first 50 values, they are almost identical (see Figure 2.1d), which demonstrates the advantage of the rank based representation. The classical Pearson correlation coefficient between the p-values is $r = 0.55$, but the rank based Spearman correlation is much larger, namely $\rho_{Sp} = 0.83$.

Due to the nature of our simulated data an association measure for truncated lists could be of interest. Hence, we calculated Spearman's correlation for incomplete rankings for

different cutoffs of truncation k_1 and k_2 shown in Figure 2.2a. The maximal correlation $\rho_{Sp} = 1$ was obtained for the short lists of top-10 and top-20 objects. Very large correlations $\rho_{Sp} \in [0.95, 0.99]$ were obtained for combinations of top-30, top-40 and top-50 objects. In comparison, Pearson's correlations for the top-30, top-40 or top-50 were close to zero, $r \leq 0.03$, the highest correlation $r = 0.97$ was obtained for the top-20 objects, see Figure 2.2c.
This example shows that the rank based Spearman correlation can much better capture the fact that the ordering of the first 50% of objects is identical in both samples although their p-values might have different magnitudes.

2.3.2 Kendall's τ

Kendall's τ_K is defined as the number of adjacent pairwise exchanges required to convert one ranking to another normalized by the maximal number of pairwise comparisons between the ranks. Basically, this means counting the number of pairwise discordances between the two ranked lists. Let us first define the discordance between two ranked lists for a pair of objects i and j as:

$$d(i,j) = \mathbb{1}\left[(r_x(i) - r_x(j))(r_y(i) - r_y(j)) < 0\right] \ , \tag{2.4}$$

for $i, j = 1, \dots, n$. $\mathbb{1}$ is the indicator function taking the value of one if the condition in the brackets is fulfilled and the value of zero otherwise.
Then Kendall's τ is defined as:

$$\tau_K = \frac{\sum_{i,j \in \{1,\dots,n\}} d(i,j)}{\binom{n}{2}} \ . \tag{2.5}$$

The normalizing constant $\binom{n}{2} = n(n-1)/2$ is the number of all possible pairwise comparisons such that $\tau_k = 1$ in case of identical rankings R_x and R_y; and $\tau_k = -1$ if ranking R_x is the reverse of ranking R_y. If the random variables X and Y are statistically independent we expect $\tau_K = 0$. Compared to Spearman's correlation, Kendall's τ_K does not take into account the absolute rankings of the objects but the relative orderings only. One disadvantage of Kendall's τ to Spearman's correlation is the complexity of the calculation. Whereas the calculation of Spearmann's ρ_{Sp} is $\mathcal{O}(n)$ in time complexity, the most sophisticated algorithm for calculation of τ_K has complexity of $\mathcal{O}(n \log n)$.
Kendall's τ_K can be modified for comparison of incomplete (truncated) ranked lists as well. We will use one of the extensions suggested by (Fagin et al., 2003) and applied in (DeConde et al., 2006; Schimek et al., 2012) which we briefly review here. For any pair

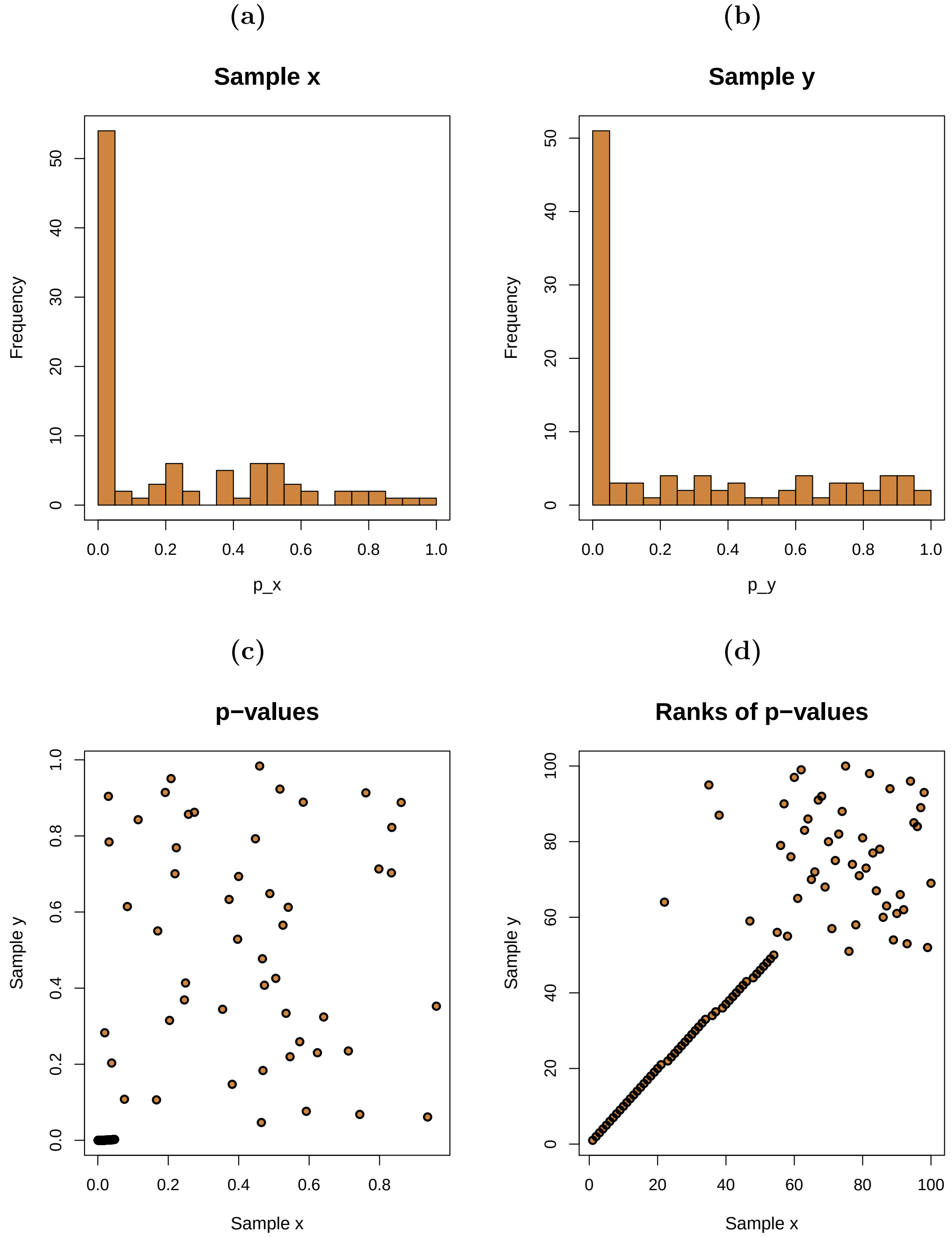


Figure 2.1: Simulated p-values of two experiments. (a) histogram of the p-values in the first experiment, (b) histogram of the p-values of the second experiment, (c) scatterplot of the p-values from both experiments (d) scatterplot of the ranks of the p-values from both experiments.

of objects $i,\ j \in O$, there are four possible cases:

1. Objects i and j appear in both top-k ranked lists R_x^k and R_y^k. Then the discordance is $d'(i,j) = \mathbb{1}\left[(r_x(i) - r_x(j))(r_y(i) - r_y(j)) < 0\right]$ as in the regular case.

2. Objects i and j appear in R_x^k and only object i (not object j) appears in R_y^k. Then the discordance $d'(i,j) = \mathbb{1}\left[r_x(i) < r_x(j)\right]$, because we infer that $r_y(i) < r_y(j)$ since object i appears in the top-k list R_y^k but object j does not.

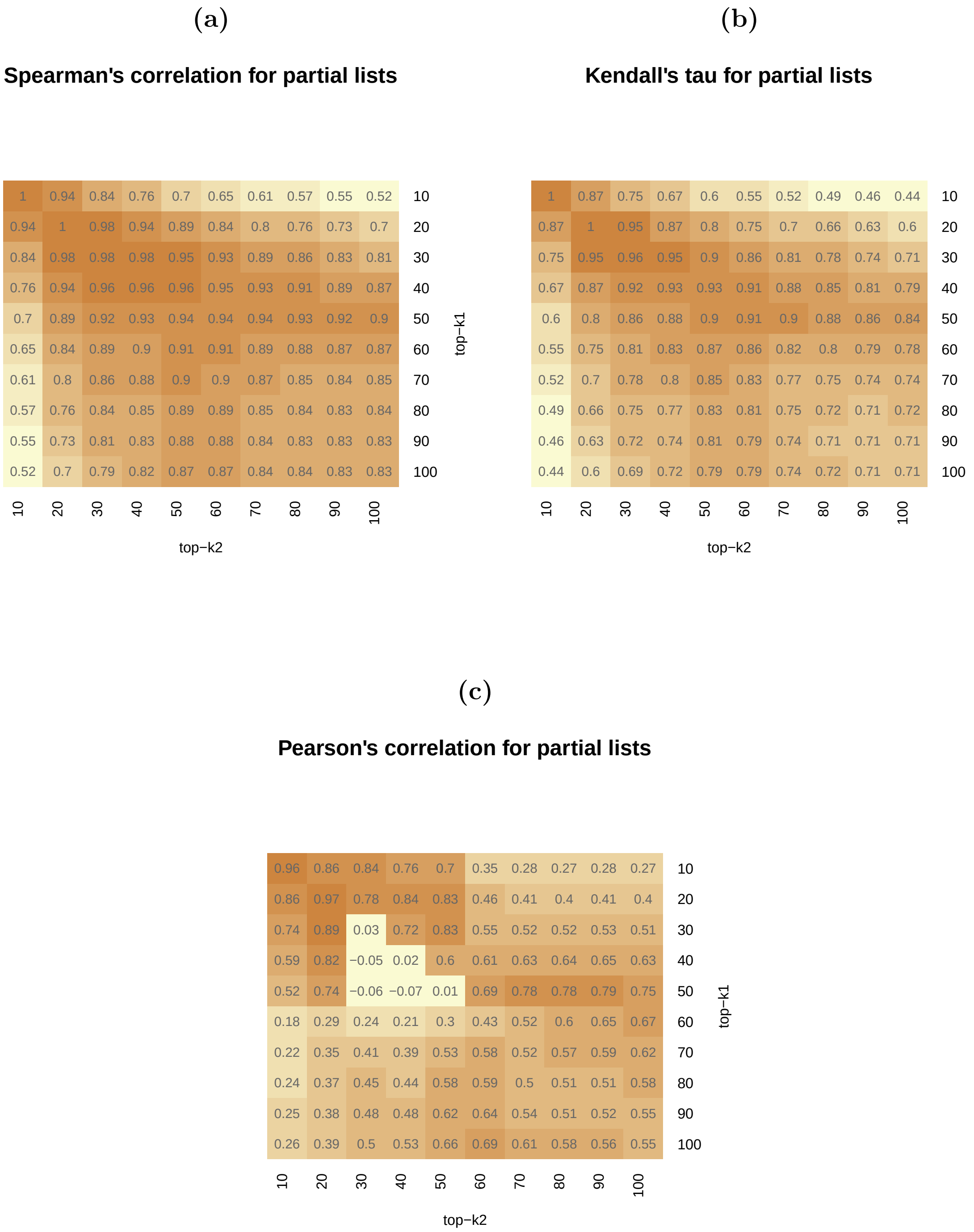
3. Object i appears only in R_x^k (and not in R_y^k) and object j appears only in R_y^k (and not in R_x^k). Then the discordance equals $d'(i,j) = 1$, since the ranking of the two objects disagree in the two lists.

4. Objects i and j both appear in R_x^k and none of them appear in R_y^k. For the missing information about the ordering of i and j in R_y, we set $d'(i,j) = \nu$, where ν is a predefined penalty $\nu \in [0,1]$. Fagin et al. suggest a neutral penalty $\nu = 0.5$ or an optimistic penalty $\nu = 0$ when the relative ranking of items ranked higher than k in one list is ignored.

Example 3. *Kendall's τ_K for the simulated data introduced in Example 2 is 0.71 which is much larger than the Pearson's correlation coefficient ($r = 0.55$) but smaller than Spearman's correlation coefficient $R_{Sp} = 0.83$.*
Kendall's τ_K for incomplete rankings with different cutoffs $k_1, k_2 \in \{10, 20, 30, \ldots, 100\}$ is shown in Figure 2.2b. Similarly to Spearman's correlation, the maximal value $\tau_K = 1$ was obtained for the short lists of top-10 and top-20 objects. Large values $\tau_K \in [0.95, 0.97]$ were obtained for combinations of top-20, top-30 and top-40 objects. With this example, we can conclude that Kendall's τ_K is another appropriate measure of rank based associations which can be adapted for partial ranked lists. Compared to Spearman's correlation it prefers slightly shorter truncated lists.

2.3.3 Fisher's exact test

Another possibility to measure an association between two ranked lists is to convert them into a contingency table and calculate Fisher's exact test for a statistical independence of the two underlying random variables (Fisher, 1935). Here, we have to split each ranked list into two parts: the top-k ranked items and the items ranked as $k+1$ or lower. Then we can construct a 2×2 contingency table as shown in Table 2.1. The entries n_{11}, n_{12}, n_{21} and n_{22} in the inner cells denote the number of objects in the corresponding categories.



(a)

Spearman's correlation for partial lists

top−k1 \ top−k2	10	20	30	40	50	60	70	80	90	100
10	1	0.94	0.84	0.76	0.7	0.65	0.61	0.57	0.55	0.52
20	0.94	1	0.98	0.94	0.89	0.84	0.8	0.76	0.73	0.7
30	0.84	0.98	0.98	0.98	0.95	0.93	0.89	0.86	0.83	0.81
40	0.76	0.94	0.96	0.96	0.96	0.95	0.93	0.91	0.89	0.87
50	0.7	0.89	0.92	0.93	0.94	0.94	0.94	0.93	0.92	0.9
60	0.65	0.84	0.89	0.9	0.91	0.91	0.89	0.88	0.87	0.87
70	0.61	0.8	0.86	0.88	0.9	0.9	0.87	0.85	0.84	0.85
80	0.57	0.76	0.84	0.85	0.89	0.89	0.85	0.84	0.83	0.84
90	0.55	0.73	0.81	0.83	0.88	0.88	0.84	0.83	0.83	0.83
100	0.52	0.7	0.79	0.82	0.87	0.87	0.84	0.84	0.83	0.83

(b)

Kendall's tau for partial lists

top−k1 \ top−k2	10	20	30	40	50	60	70	80	90	100
10	1	0.87	0.75	0.67	0.6	0.55	0.52	0.49	0.46	0.44
20	0.87	1	0.95	0.87	0.8	0.75	0.7	0.66	0.63	0.6
30	0.75	0.95	0.96	0.95	0.9	0.86	0.81	0.78	0.74	0.71
40	0.67	0.87	0.92	0.93	0.93	0.91	0.88	0.85	0.81	0.79
50	0.6	0.8	0.86	0.88	0.9	0.91	0.9	0.88	0.86	0.84
60	0.55	0.75	0.81	0.83	0.87	0.86	0.82	0.8	0.79	0.78
70	0.52	0.7	0.78	0.8	0.85	0.83	0.77	0.75	0.74	0.74
80	0.49	0.66	0.75	0.77	0.83	0.81	0.75	0.72	0.71	0.72
90	0.46	0.63	0.72	0.74	0.81	0.79	0.74	0.71	0.71	0.71
100	0.44	0.6	0.69	0.72	0.79	0.79	0.74	0.72	0.71	0.71

(c)

Pearson's correlation for partial lists

top−k1 \ top−k2	10	20	30	40	50	60	70	80	90	100
10	0.96	0.86	0.84	0.76	0.7	0.35	0.28	0.27	0.28	0.27
20	0.86	0.97	0.78	0.84	0.83	0.46	0.41	0.4	0.41	0.4
30	0.74	0.89	0.03	0.72	0.83	0.55	0.52	0.52	0.53	0.51
40	0.59	0.82	−0.05	0.02	0.6	0.61	0.63	0.64	0.65	0.63
50	0.52	0.74	−0.06	−0.07	0.01	0.69	0.78	0.78	0.79	0.75
60	0.18	0.29	0.24	0.21	0.3	0.43	0.52	0.6	0.65	0.67
70	0.22	0.35	0.41	0.39	0.53	0.58	0.52	0.57	0.59	0.62
80	0.24	0.37	0.45	0.44	0.58	0.59	0.5	0.51	0.51	0.58
90	0.25	0.38	0.48	0.48	0.62	0.64	0.54	0.51	0.52	0.55
100	0.26	0.39	0.5	0.53	0.66	0.69	0.61	0.58	0.56	0.55

Figure 2.2: (a) Spearman's correlation coefficient, (b) Kendall's τ_K and (c) Pearson's correlation coefficients for simulated data with various cutoffs k_1 (horizontal axis) and k_2 (vertical axis).

n_{j+} and n_{+k} denote the one-way marginal totals, defined as: $n_{j+} = \sum_{k=1}^{2} n_{jk}$,$\forall j \in \{1, 2\}$ as the row marginals and $n_{+k} = \sum_{j=1}^{2} n_{jk}$,$\forall k \in \{1, 2\}$ as the column marginals. For the two-way marginal, it holds: $n_{++} = \sum_{j=1}^{2}\sum_{k=1}^{2} n_{j,k} = n$.

Table 2.1: 2×2 contingency table

	$\lvert i\rvert : r_y(i) \leq k_2$	$\lvert i\rvert : r_y(i) > k_2$	$\sum$
$\lvert i\rvert : r_x(i) \leq k_1$	n_{11}	n_{12}	n_{1+}
$\lvert i\rvert : r_x(i) > k_1$	n_{21}	n_{22}	n_{2+}
$\sum$	n_{+1}	n_{+2}	n_{++}

Under the null hypothesis H_0 of independence of the ranked lists R_x and R_y, conditioning on the marginal totals, the counts in the 2×2 contingency table follow the hypergeometric distribution (Agresti, 2013; Fisher, 1935):

$$P(n_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}} = \frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{n_{11}!n_{12}!n_{21}!n_{22}!n!} \tag{2.6}$$

The H_0 of independence for 2×2 contingency tables is equivalent with $H_0 : \theta = 1$, where θ denotes the odds ratio. For our purpose we want to test for a positive association between the two variables (e.g. ranked lists), so we can formulate the alternative hypothesis as $H_a : \theta > 1$. Then the p-value of Fisher's exact test equals the probability $P(n_{11} \geq t_0)$, where t_0 stands for the observed value of n_{11}. In other words, the p-value is the sum of the null probabilities of the observed contingency table and of tables having more extreme (in this case of H_a, greater) counts n_{11} (Agresti, 2013; Fisher, 1935):

$$P(n_{11} \geq t_0) = \sum_{t=t_0}^{m} \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}} = 1 - \sum_{t=0}^{t_0-1} \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}} \ , \tag{2.7}$$

for $m = \min\{n_{1+}, n_{+1}\}$ as the maximal possible value of n_{11}.

The significance of the association between the two ranked lists R_x and R_y detected by the contingency table depends on the cutoffs k_1 and k_2 which partition the ranked objects into the top-ranked categories. To overcome this problem, Eden et al. (2007);

Roider et al. (2009) suggest a solution to iterate over a sequence of possible cutoffs and they assume that the smallest achieved p-value of tests with all possible combinations corresponds to the most meaningful detectable association between the two ranked lists. The resulting minimal p-value cannot be interpreted as the exact p-value due to the multiple testing complication. Eden et al. (2007) provide a novel algorithm for calculation of an exact p-value which can be determined by means of dynamic programming.

Example 4. *For the illustration of Fisher's exact test on two ranked lists we use the simulated data introduced in Example 2 . First, we define a sequence of all possible cutoffs which we want to test. So let us investigate all pairwise combinations of cutoffs* $k_1,\ k_2 \in \{5, 10, 15, \dots, 75, 80\}$*. Then, for each combination of* k_1 *and* k_2 *we construct a* 2×2 *contingency table and calculate the significance of the number of shared top-ranked objects. For a better demonstration, we show the contingency table for* $k_1 = k_2 = 40$ *in Table 2.2. The* p*-value of this contingency table is* $P(n_{11} \geq 37) = 2.47 \cdot 10^{-20}$ *which is highly significant.*

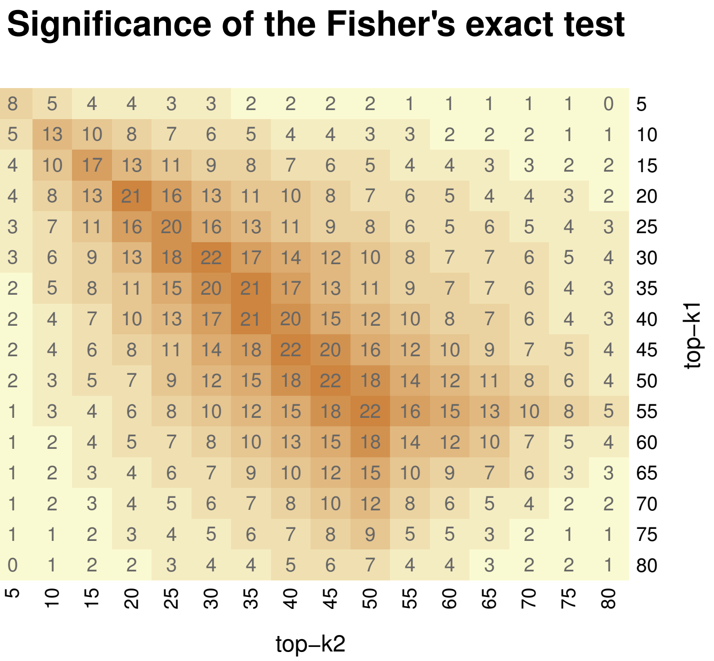
Table 2.2: 2×2 contingency table for simulated data and cutoffs $k_1 = k_2 = 40$.

	$\lvert i\rvert : r_y(i) \leq 40$	$\lvert i\rvert : r_y(i) > 40$	$\sum$
$\lvert i\rvert : r_x(i) \leq 40$	37	3	$n_{1+} = 40$
$\lvert i\rvert : r_x(i) > 40$	3	57	$n_{2+} = 60$
$\sum$	$n_{+1} = 40$	$n_{+2} = 60$	$n_{++} = 100$

However, to find the best combination of cutoffs with the smallest p*-value, we have to investigate all 256 combinations of* k_1 *and* k_2*. The significance (as* $-\log_{10} p$*-value) for all combinations is shown in the Figure 2.3. The maximal significance* $\log_{10}(3.45 \cdot 10^{-23} = 22.46$ *was achieved for the combinations* $k_1 = 50;\ k_2 = 45$ *and* $k_1 = 55;\ k_2 = 50$*. These values lie very close to the true values used in the simulation that were* $k_1 = k_2 = 50$*. Thus, Fisher's exact test assigns the maximal significance to the correct values and herewith performs better than Spearman's correlation or Kendall's* τ_K*. Note that the significance decreases rapidly for* $k_1, k_2 > 55$ *where the ranking of both samples becomes random.*

2.3.4 Restricted two-dimensional Kolmogorov-Smirnov score

Another measure of similarity between two ranked lists was proposed by Ni and Vingron (2012). The restricted two-dimensional Kolmogorov-Smirnov score (R2KS), puts the



5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	top–k1
8	5	4	4	3	3	2	2	2	2	1	1	1	1	1	0	5
5	13	10	8	7	6	5	4	4	3	3	2	2	2	1	1	10
4	10	17	13	11	9	8	7	6	5	4	4	3	3	2	2	15
4	8	13	21	16	13	11	10	8	7	6	5	4	4	3	2	20
3	7	11	16	20	16	13	11	9	8	6	5	6	5	4	3	25
3	6	9	13	18	22	17	14	12	10	8	7	7	6	5	4	30
2	5	8	11	15	20	21	17	13	11	9	7	7	6	4	3	35
2	4	7	10	13	17	21	20	15	12	10	8	7	6	4	3	40
2	4	6	8	11	14	18	22	20	16	12	10	9	7	5	4	45
2	3	5	7	9	12	15	18	22	18	14	12	11	8	6	4	50
1	3	4	6	8	10	12	15	18	22	16	15	13	10	8	5	55
1	2	4	5	7	8	10	13	15	18	14	12	10	7	5	4	60
1	2	3	4	6	7	9	10	12	15	10	9	7	6	3	3	65
1	2	3	4	5	6	7	8	10	12	8	6	5	4	2	2	70
1	1	2	3	4	5	6	7	8	9	5	5	3	2	1	1	75
0	1	2	2	3	4	4	5	6	7	4	4	3	2	2	1	80

Figure 2.3: Significance of Fisher's exact tests for simulated data with various cutoffs k_1 (horizontal axis) and k_2 (vertical axis). Rounded $-\log_{10}(p\text{-values})$ for each underlying contingency table with the corresponding k_1 and k_2 cutoffs are shown in the matrix.

main emphasis on finding the objects at the top of the lists, while the objects with lower rankings should have a weaker influence. The motivation of the R2KS is in the graphical illustration of the two ranked lists in a scatter plot. In case that the two ranked lists are unrelated of each other and rankings are complete and without ties, the ranks of one lists should be uniformly distributed throughout the other list. If the two ranked lists have an association among the ranks on the top of the lists one would see a concentration of dots in the lower-left rectangle of the scatterplot. The R2KS measure quantifies this effect by comparing the observed density of the dots (ranked objects) in the lower-left rectangle with the expected density. Similarly to the (one-dimensional) Kolmogorov-Smirnov statistic, the final score is the score in the point with the maximal difference between the observed and expected density. Hence, this approach can be understood as a special version of a two-dimensional Kolmogorov-Smirnov statistic. Formally, for two ranked lists R_x and R_y with $|R_x| = |R_y| = n$, let us define $A_{i,j}$ as the rectangle $[1,\ldots,i] \times [1,\ldots,j]$, i and j are the ranks in R_x and R_y, respectively. Let U denote the whole space $[1,\ldots n] \times [1,\ldots n]$. Then for each combination $i \times j$, $R_{i,j}$ is defined as:

$$R_{i,j} = \frac{\text{number of objects in } A_{i,j}}{n} - \frac{i \times j}{n \times n}\,, \tag{2.8}$$

and in a weighted version:

$$R'_{i,j} = \frac{\sum_{l \in A_{i,j}} w_l}{\sum_{l \in U} w_l} - \frac{i \times j}{n \times n} \text{ with weights: } w_l = \frac{h \cdot (h+1)}{2} \ , \tag{2.9}$$

where h is the distance of the object to a cutoff k defining the most informative top-ranked objects. The R2KS score R^* is then defined as:

$$R^* = \max_{i,j}(R_{i,j}) \text{ or in the weighted version: } R^* = \max_{i,j}(R'_{i,j}) \ . \tag{2.10}$$

Ni and Vingron (2012) give a simple dynamic programming algorithm for computation of the R2KS score and show a stable distribution of the normalized score $\sqrt{n}R^*$ for simulated datasets.

Example 5. *The unweighted R2KS score for the simulated lists introduced in Example 2 was relatively small: $R^* = 2.32$. The reason for this might be the short lists in the simulated data set since the R2KS is designed for large samples.*

2.3.5 Irreproducible discovery rate

Li et al. (2011a) assess the reproducibility of ranked objects and consistency of the top-ranked objects in long ranked lists over a small number of different experiments (e.g. replicates). Li et al. define reproducibility as the extent to which the ranks of the signals are no longer consistent across replicates in decreasing significance. The loss of consistency of rankings is visualized via a copula-based graphical tool, which enables an empirical inspection of the possible consistency breakdown without any prior model assumption or without predefined thresholds. Then a copula mixture model is used to quantify the reproducibility by classifying the signals into a reproducible and irreproducible groups. For the copula-based graphical tool, let us first define $\Psi_n(k_1, k_2)$, the proportion of the pairs ranked both among the top-k_1% objects of R_x and among the top-k_2% objects of R_y:

$$\Psi_n(k_1, k_2) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left[(r_x(i) \leq \lceil nk_1 \rceil, r_y(i) \leq \lceil nk_2 \rceil \right] \ , \quad 0 < k_1 \leq 1, \ 0 < k_2 < 1 \ , \tag{2.11}$$

where $\lceil nk_1 \rceil$ denotes the smallest integer greater or equal nk_1. $\Psi_n(k_1, k_2)$ is an empirical survival copula and for simplicity we use $\Psi_n(k) = \Psi_n(k_1, k_2)$ for $k = k_1 = k_2$. Then the change of consistency with the decrease of significance can be visualized when plotting $\Psi_n(k)$ vs. k for $0 \leq k \leq 1$, which is called a correspondence curve. One can also plot the derivative $\Psi'_n(k)$ of $\Psi_n(k)$ which shows the change of the correspondence curve. From the properties of the survival copula, the correspondence curve and the change of the correspondence curve have the following patterns in the three extreme cases:

1. perfectly correlated ranks $R_x = R_y$: all points lie on a straight line $\Psi_n = t$; all points lie on a horizontal line $\Psi'_n = 1$, see Figure 2.4a.

2. independent ranks $R_x \perp R_y$: all points lie on a parabola $\Psi_n = k^2$; all points lie on a straight line $\Psi'_n = 2t$, see Figure 2.4b.

3. perfectly correlated ranks for the top-k_0 objects and independent for the higher ranked objects: the points lie on the curve:

$$\Psi_n(k) = \begin{cases} k & \text{if } k \leq k_0 \\ \frac{k^2-2kk_0+k_0}{1-k_0} & \text{if } k > k_0 \end{cases}$$

   or on the derivative curve:

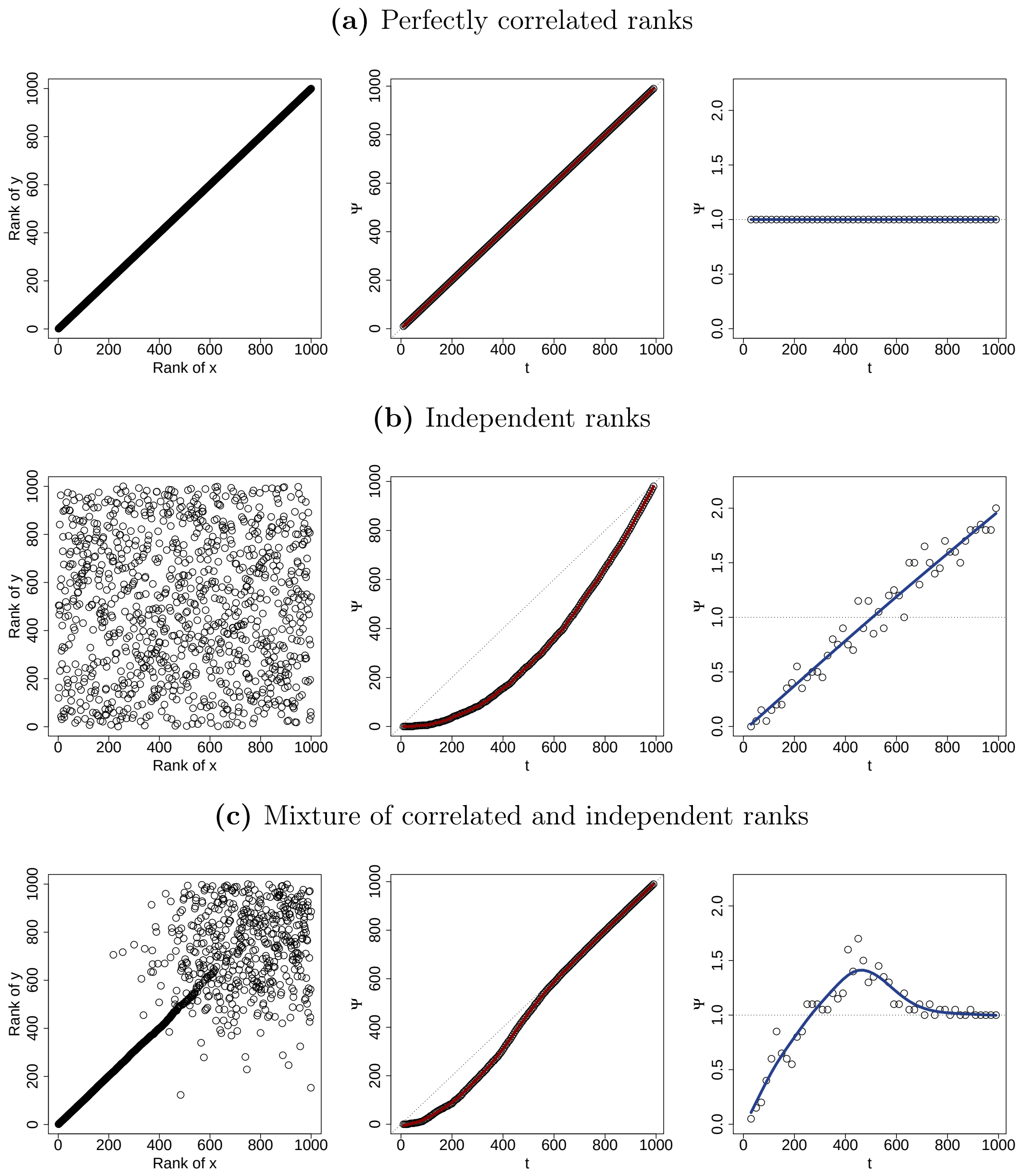
$$\Psi'_n(k) = \begin{cases} 1 & \text{if } k \leq k_0 \\ \frac{2(k-k_0)}{1-k_0} & \text{if } k > k_0 \end{cases}$$

   Both functions are depicted in Figure 2.4c.

These properties enable the user to detect the level of the positive association between the two ranked lists and to derive the consistency breakdown from the graphical visualisation.

For the statistical modeling of the dependence structure of the two ranked lists, Li et al. suggest a semiparametric copula model, in which the marginal distributions (usually unknown) are estimated nonparametrically by the ranks and the associations are modeled parametrically. A simple parametric model for the associations should be able to distinguish the objects with consistent rankings over the lists from the noisy, inconsistent ones. We assume that the assignment of ranks (or of the original scores) consists of a mixture of two processes: the genuine process generating reproducible ranks (or scores) and a spurious process generating random ranks (scores). Then we can assume that the dependence structure between the replicates is different in the reproducible group and in the less reproducible group.

Namely, the dependence between the replicates in the reproducible group is modeled with a bivariate Gaussian distribution with the correlation coefficient $\rho > 0$, whereas the dependence between the replicates in the less reproducible group is modeled with



(a) Perfectly correlated ranks

(b) Independent ranks

(c) Mixture of correlated and independent ranks

Figure 2.4: Correspondence curve and the change of the correspondence curve for three different data associations. From left to right: scatterplot of the ranks, correspondence curve Ψ in estimated data points with a smoothed curve (red line), change of the correspondence curve Ψ′ in estimated data points with a smoothed curve (blue line). (a) Perfectly correlated ranks: data from multivariate Gaussian distribution with correlation matrix $\Sigma = \left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right)$ (b) independent ranks: data from multivariate Gaussian distribution with correlation matrix $\Sigma = \left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$ (c) perfectly correlated ranks for the top 500 objects and independent ranks for the higher ranked objects: data from a mixture of multivariate Gaussian with a perfect correlation and no correlation.

a bivariate Gaussian distribution with $\rho = 0$. Then the underlying latent variables, representing the unobserved biological replicates, follow a bivariate Gaussian distribution conditioned on an unobserved Bernoulli-distributed variable indicating whether the rank was assigned by the genuine process or by the spurious process. Thus the copula mixture model is parametrized by the probability of the Bernoulli-variable being 1 (e.g. rank assigned by the genuine process), by the parameters of the Gaussian distribution of the reproducible group (mean, covariance matrix) and by the continuous marginal distribution of the original scores in the two groups. The parameters can be estimated with the 'pseudo-likelihood' approach.

The *local irreproducible discovery rate (idr)* which corresponds to the posterior probability that a rank is irreproducible in a pair of replicates, can be calculated from the copula density function. The *irreproducible discovery rate (IDR)* is then defined in analogy to the false discovery rate (FDR, Benjamini and Hochberg (1995)), as expected irreproducible discovery rate for observations that are as irreproducible or more irreproducible than the given observations. This corresponds to a procedure giving in case of n tests (e.g. ranked objects in the list), an expected rate of irreproducible replicates no greater than predefined level $\alpha \cdot n$, where $\alpha \in [0, 1]$.

Example 6. *We used the simulated data introduced in Example 2 and estimated the expected irreproducible discovery rate IDR for all 100 objects. Then we selected objects with $IDR \leq 0.05$ which were the first 50 ranked objects and 2 other objects from the bottom part of the lists. Thus, the first 50 objects were correctly identified as irreproducible (because of a high rank correlation). Further we estimated the correspondence curve Ψ and its derivative Ψ' to find the point where the consistency of ranks is lost. The IDR tool identifies the correct point at the position 50 and the consistency loss point is visible in the correspondence curve Ψ as the decline from the diagonal line at point 50, see Figure 2.5. Similarly, the derivative of the correspondence curve Ψ' declines from the horizontal line between points 40 and 50.*

2.4 Conclusion

Due to the rapid development of high-throughput approaches in molecular genomics, one of the main challenges for scientists now is the integration of high-dimensional data to gain insights into biological mechanisms. One of the popular data transformation techniques, which is independent of scaling and transformation, is **ranked list data representation** in which the measured items are sorted by a chosen quantity such that the most informative items are placed on the top of the list.

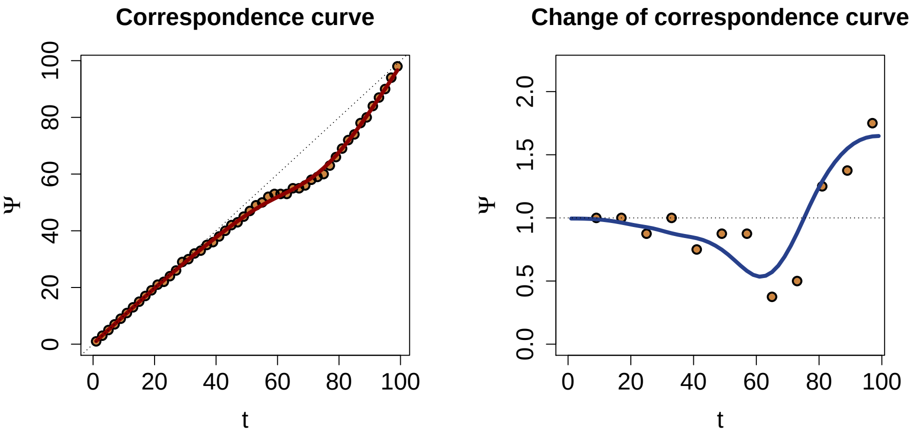


Figure 2.5: Correspondence curve Ψ (left) and the derivative of correspondence curve Ψ′ (right) for the simulated data from Example 2. Points are the estimated values in selected points, solid lines show the fitted splines.

One of the advantages of ranked list data representation is the simplicity of comparison of two ranked lists. To determine the strength of an association between two ranked lists, a large variety of nonparametric, rank based measures can be applied. In addition, most of the measures can be extended to truncated ranked lists, where the focus is on the k top-ranked items. In this chapter, we discussed five of the most common **rank based measures** applied on large lists: Spearman's and Kendall's correlation (ρ_{Sp} and τ_K, respectively), Fisher's exact test, Restricted two-dimensional Kolmogorov-Smirnov score (R2KS) and Irreproducible discovery rate (IDR). The technical details of these measures were discussed in Section 2.3.

Furthermore, their application on a simple example with two ranked lists based on p-values from two different experiments was demonstrated. In the example, two samples of p-values assigned to 100 genes were simulated. 50 genes were differentially expressed, with significant p-values. However, the significant p-values in one sample were quadratically smaller such that a direct comparison of the p-values was difficult but their ranking was almost identical, see Figure 2.1.

We calculated all five rank based association measures for these two simulated lists and further we tried to identify the true significant genes using all five measures. We could show that the rank based correlations (ρ_{Sp} and τ_K) outperform the conventional Pearson's correlation coefficient r for the simulated example.

Additionally, Fisher's exact test and IDR almost perfectly identified the true 50 significant genes in both lists. Spearman's and Kendall's correlation together with R2KS underestimated the number of significant genes (cutoff of 20 and 3 genes, respectively). One of the reasons for the good performance of Fisher's exact test and IDR is that they are designed for the detection of the most reproducible (or most associated) subsamples, whereas Spearman's and Kendall's correlation are designed for comparison of full lists and were additionally adapted for comparing incomplete lists. R2KS puts the main emphasis on finding items at the top of the lists by comparing the density of the dots in a scatterplot of the top-ranked items to the density of the dots in scatterplot with low-ranked items. Surprisingly, R2KS considerably underestimates the number of significant genes in this simulated example. The reason for this might be the small size of the simulated samples since R2KS was constructed for analyzing large data sets (thousands of genes). IDR was developed in such way that it identifies reproducible items, e.g. items with similar rankings over different experiments and therefore is able to find the significant genes in the simulated example. Fisher's exact test divides the two lists into top-ranked part and lower-ranked part and then calculates the significance of the shared top-ranked items. Thus with all possible partitions of the lists to be tested, it is possible to detect the best partition with the highest significance. The disadvantage of repeated Fisher's exact test approach is the exponentially increasing number of required tests with the length of the ranked lists. On the other hand, one advantage of the Fisher's exact test is that it can be relatively easily extended for three or more variables by constructing multiway contingency tables. This extension together with an application on predicting tissue-specific transcription factor co-occurrence will be discussed in Chapter 4.

3 Prediction of transcription factor co-occurrence on human promoters

3.1 Motivation

In this chapter, we apply the rank based association measures described in Section 2.3 to identify pairs of transcription factors (TFs) with highly associated ranked lists of their target genes. To do so, we represent the **TF** as a **ranked list of its target genes** based on the predicted binding affinity of the TF to the promoter sequences of the target genes. Thus, if we assume that two co-occurring TFs share a significantly higher number of target genes as compared with randomly selected TFs, then the TF pairs with highly associated ranked lists represent the predicted **co-occurring TFs**. Namely, if two TFs bind to the same promoter regions together more frequently than by chance they very likely act together to direct the expression of their (shared) target genes. To evaluate the significance of the shared target genes, we apply the five rank based association measures introduced in Section 2.3.

In the next section, the representation of a transcription factor by a ranked list of its target genes is introduced. The confounding factor of **motif similarity** is presented in Section 3.3. In Section 3.4, the rank based measures are applied to pairs of TFs represented by ranked lists of their target genes. The distribution of these measures together with their relation to motif similarity is described. Section 3.5 describes the predicted co-occurring transcription factors and compares the obtained results using different association measures. The last Section summarizes our findings.

3.2 Representing transcription factors by ranked lists of their target genes

For our approach, we represent each transcription factor of interest as a ranked list of its target genes (e.g. promoters) ordered by the estimated binding affinity of the transcription factor to the promoter sequences. To define such ranked lists one needs to:

1. define the promoter sequences of all genes of the studied organism (e.g. human)
2. define the binding preferences of the studied TFs using position weight matrices (PWMs)
3. predict the binding affinity of the TFs using their PWMs to the promoter sequences in such a way that they are comparable

To accomplish step 1., we define the promoter sequences as windows up to 500 basepairs (bp) upstream of the transcription start site (TSS) for all catalogued genes in the *hg19 Ensembl* assembly of the human genome from genome.ucsc.edu, which results in a list of 42 380 promoters.

For step 2., we use the JASPAR CORE Vertebrata database (Bryne et al., 2008) of 130 PWMs corresponding to the most studied mammalian TFs. This results in total of $130 * 129/2 = 8385$ TF pairs that are studied for their association.

For step 3., we choose the TRAP approach from Roider et al. (2007) to predict the binding affinity of the 130 TFs to the promoter sequences. Unlike the hit-based binding affinity prediction models TRAP avoids the artificial separation between binding sites and non-binding sites in the studied sequence but rather computes the binding probability of a given TF (e.g. PWM) to each possible site in the sequence. These probabilities are then summed over the whole sequence such that we obtain a single value of the binding affinity of the particular PWM to the particular promoter sequence, see Section 1.2.3. An example of the distributions of the TRAP scores for the promoter sequences of four different TFs is shown in Figure 3.1. Note that the x-axis of the histograms vary from the interval $[0; 0.8]$ to the interval $[0; 3.5]$. The ranges of the scores are not directly comparable, thus the rank based representation of the TFs is suitable. Therefore, the promoter sequences (e.g. genes) are sorted in decreasing order by the binding affinity separately for each TF such that the genes with high binding affinity are placed at the top of the list.

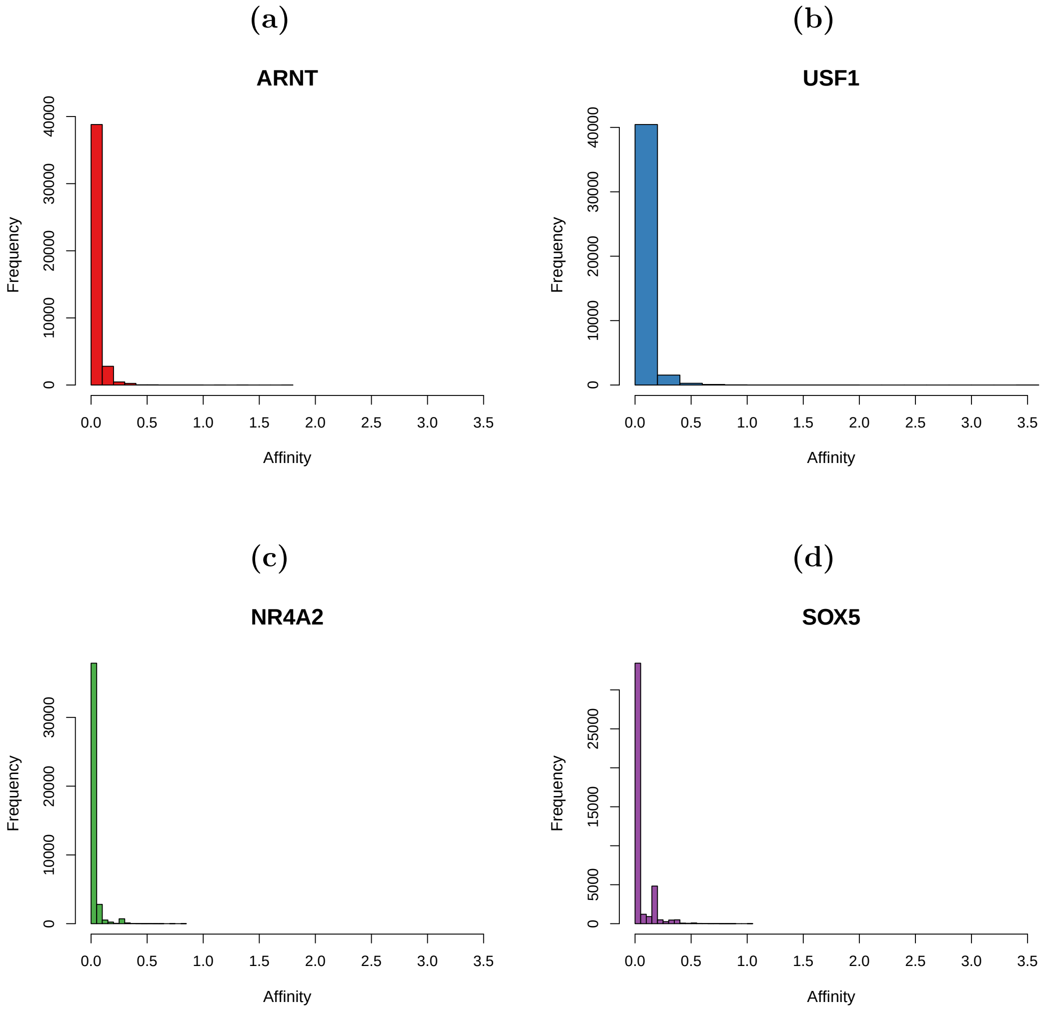


Figure 3.1: Distribution of TRAP binding affinities for four transcription factors (a) ARNT, (b) USF1, (c) NR4A2 and (d) SOX5.

3.3 Confounding factor: motif similarity

Let us consider two TFs with very similar PWMs, graphically represented with motif logos as shown in Figure 3.2 for the TFs ARNT and USF1. Very likely, the ranked lists of their target genes will be very similar merely due to their motif similarity and not necessarily due to their real co-occurrence at the genes' promoters (Pape et al., 2009). To eliminate the selection of highly associated TF pairs with similar ranked lists as a result of the similarity of their motifs, we include a confounding factor into the analysis, a measure of motif similarity. We choose the motif similarity measure MOSTA developed by Pape et al. (2008) which is defined as the maximum of log-odds ratios of the overlap probability and the independent probability of hits of two motifs over all possible configurations on both strands of the DNA sequence. Formally, $S_d(M_i, M_j)$ denotes the log odds ratio of the independent hits and joint hits of motifs M_i and M_j with distance d and is defined as follows:

$$S_d(M_i, M_j) = \log \left(\frac{P(\gamma_k^{M_i} = 1) \cdot P(\gamma_k^{M_j} = 1)}{P(\gamma_k^{M_i} = 1,\ \gamma_{k+d}^{M_j} = 1)} \right) , \tag{3.1}$$

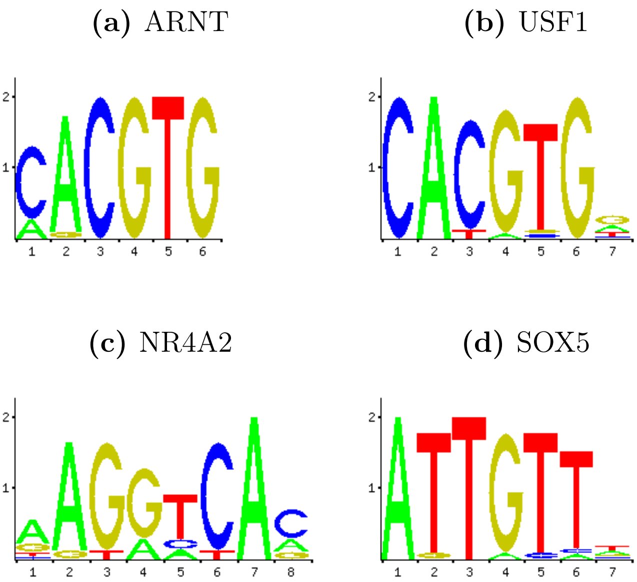
where $\gamma_k^{M_i}$ is an indicator random variable for a hit of motif M_i at starting position k. MOSTA score $S^{\max}$ is then defined as the maximum over all distances between the two motifs and over all possible configurations on both strands on the DNA sequence:

$$\begin{aligned} S^{\max}(M_i, M_j) &= \max \Big[\max_d S_d(M_i, M_j), \max_d S_d(M_j, M_i), \\ &\quad \max_d S_d(M_i', M_j), \max_d S_d(M_j', M_i) \Big] \ . \end{aligned} \tag{3.2}$$

Here, the parameter d denotes the nucleotide position w.r.t. the first motif, M_i' assigns motif M_i on the reverse complement strand.

The MOSTA scores were calculated for all 8385 TF pairs in our set and the distribution of the MOSTA similarity scores is shown in Figure 3.3. In our analysis, we focus on TF pairs with small motif similarity defined by a threshold of the 90%-quantile of the empirical distribution of MOSTA scores of the 130 JASPAR motifs. We choose the quantile as a threshold for non-similar matrices to avoid TF pairs with very similar or almost identical binding motifs. This threshold is highlighted with a vertical red line in Figure 3.3.

To overcome the problem of similar ranked lists based only on the similarity of the underlying binding motifs another approach could be applied. As an alternative approach of predicting the binding affinity of a particular TF to a DNA sequence the hit-based



(a) ARNT **(b)** USF1

(c) NR4A2 **(d)** SOX5

Figure 3.2: Four transcription factor motifs with various motif similarities. ARNT (a) and USF1 (b) have very similar motifs ($S^{\max} = 7.31$), ARNT (a) and NR4A2 (c) motifs have medium similarity ($S^{\max} = 3.60$) and ARNT (a) and SOX5 (d) have very different motifs ($S^{\max} = -0.18$).

method could be used, which predicts the exact binding sites usually based on some significance threshold. Then the TF can be represented as a ranked lists of promoter sequences ranked by both the significance of the binding sites and the number of the binding sites in the sequence. However, this ranking might not be as straightforward as the ranking based on the TRAP scores. One would have to decidet the importance of two quantities: shall a promoter with a large number of less significant binding sites be ranked higher than a promoter with few, but highly significant binding sites? To avoid the high similarity of the ranked lists due to their motif similarity, one would have to control for an overlap of the predicted binding sites for each pair of TFs such that the motif will not overlap by more than $p\%$ of the length of the motif or simply by more than x basepairs. Thus, this alternative method requires another kind of thresholding - for binding site prediction and for the overlap control - which might be even more complex. For this reason, we choose from our perspective the simpler method using TRAP and MOSTA scores.

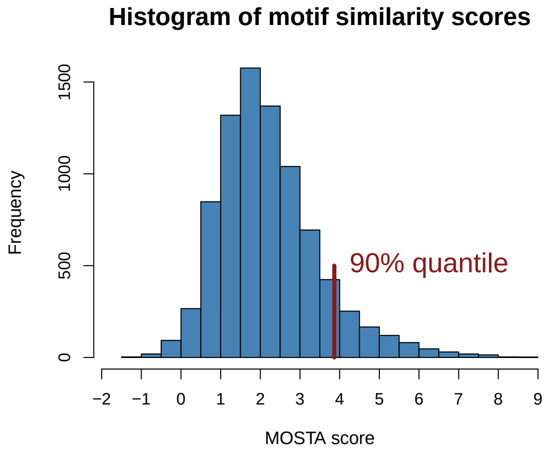


Figure 3.3: Histogram of MOSTA motif similarity values for all transcription factor pairs from JASPAR database. The 90%−quantile at the value 3.9 is highlighted with the red line.

3.4 Rank based measures applied on ranked lists of target genes

Here, we apply the rank based methods introduced in Chapter 2 to all TF pairs in the JASPAR database to predict transcription factor co-occurrence in human promoters. To do so, we calculated the rank-based association measures for all pairs of transcription factors represented as ranked lists of promoters.

Let us first investigate the distribution of the association measures for all TF pairs, as shown in the left column of Figure 3.4. The top two panels correspond to the rank correlations with similar distribution which have a bell-shaped density in the interval $[-1, 1]$ with highest density at 0.2 (Spearman correlation) and at 0.1 (Kendall's τ). Both of the distributions are slightly negative skewed which might be caused by the signal from TF pairs with highly similar ranked lists of promoters.

The third panel shows a histogram of the minimal p-values of Fisher's exact test, to define the top-ranked genes, all possible combination of cutoffs $k_1, k_2 \in \{10, 20, \dots, 90, 100, 200,\text{-} \dots, 2000\}$ were used. For a random data, one expects the p-values to be uniformly distributed, however our histogram has many small p-values in the interval $[0, 0.05]$. This phenomenon has two explanations. First, the overrepresented small p-values come from overrepresented TF pairs with highly similar ranked lists, e.g. from a real biological signal. Second, the overrepresentation of the small p-values is due to the way how we choose the p-values. Namely, we selected the minimal p-values obtained from Fisher's exact test for contingency tables which were based on different cutoffs.

The forth panel from the top shows the distribution of R2KS scores with the majority of measurements in the interval $[0, 20]$ and a maximum at 42. The bottom panel shows the distribution of the IDR scores. Here, we calculated for each TF pair the proportion of highly reproducible promoters with an IDR score $\leq 10^{-10}$. The majority of TF pairs share a small portion ($< 15\%$) of highly reproducible promoters, however there is a group of approximately 500 completely reproducible TF pairs (with proportion one).

As expected, Spearman's rank correlation and Kendall's τ give very similar results, the Pearson's correlation coefficient between them is one. Yet, the correlation of the rank correlations and R2KS score is very high too ($r = 0.9$). Similarly, the correlation of the rank correlations and the significance of Fisher's exact test (defined as negative logarithm of p-value) is relative high, $r = 0.65$. Surprisingly, IDR proportion has a negative correlation to all other measures in a range of -0.7 with R2KS to -0.3 with the significance of Fisher's exact test.

Next we studied the relationship between the rank based association measures and the confounding factor motif similarity. The right column in Figure 3.4 shows smooth scatterplots of these measures and the MOSTA motif similarity score introduced in Section 3.3. The density of the data points is represented with the dark color in all smooth scatterplots. Fisher's p-values were transformed to negative logarithms with base 10 with predefined minimal value of 10^{-50} to avoid negative infinity values for $\log(0)$.

As expected, TF pairs with very similar motifs ($S^{\max} \in [6; 8]$) have high rank correlations, both Spearman's and Kendall's, highly significant p-values and large R2KS values. Surprisingly, a large group of nonsimilar TF pairs with $S^{\max} \leq 3$ are completely reproducible, with IDR proportions equal to one. All measures are positively correlated ($r \in [0.36, 0.41]$) to the motif similarity score with the exception of the IDR proportions which are negatively correlated ($r = -0.16$).

We also identified known protein-protein interactions (PPIs) that are present in public databases (Chatraryamontri et al., 2013; Ravasi et al., 2010), denoted as red crosses in the scatterplots in Figure 3.4. However, the majority of these known interactions correspond to TF pairs with relatively small correlations ($\rho_{Sp} \leq 0.5, \tau_K \leq 0.4$) or low significance ($p \geq 10^{-3}$), small R2KS (R2KS ≤ 20) and IDR proportions (IDR ≤ 0.1). Only a small group of non-similar interacting TFs have large significant values according to Fisher's test, these are protein-protein interactions mainly involving the factor SP1 (SP1:E2F1, SP1:MYC, SP1:YY1) and ETS1:NFKB. Three known PPIs (CEBPA:ESR1, CEBPA:MYC, ESR1:FOXO3) have large IDR value and small motif similarity. However, these pairs do not overlap with the significant ones from Fisher's test.

3.5 Predicted co-occurring TF pairs on promoters

To define significantly co-occurring TFs we select TF pairs with the largest association scores (larger than the 99%-quantile of the corresponding measure) with nonsimilar motifs ($S^{\max} < 3.9$). With this criterion, we detected between 29 (with Spearman's correlation) and 428 (with IDR) significant TF pairs. Consistency among the first four methods (Spearman's correlation, Kendall's τ, R2KS and Fisher's exact test) is very high. 26 out of 29 TF pairs found with Spearman's correlation are found by all other methods and all 29 TF pairs are also found with Kendall's τ and R2KS. Kendall's τ detects one more TF pair, which is found with R2KS as well. R2KS identified 6 more significant TF pairs, 2 of which were also identified by Fisher's exact test. Findings for these four association measures are summarized in the Venn diagram of the overlapping significant TF pairs, shown in Figure 3.5.

Surprisingly, there was no agreement found among the predicted 428 TF pairs by the IDR proportion and all other rank based measures. We have investigated the binding affinity scores and rankings of these TF pairs with a large proportion of promoters with small IDR values. The majority of these TF pairs show very surprising relationship of *mutual exclusivity* of their binding affinities, see scatterplots of 8 TF pairs in Appendix, Figure A.1. Since we are interested in TF pairs with similarly top-ranked promoters, TF pairs predicted with the IDR score were not considered for further analysis.

Out of the TF pairs which were significant by at least one of the four association measures (Spearman's correlation, Kendall's τ, R2KS and Fisher's exact test) an interaction network was derived, see Figure 3.6. Here, transcription factors correspond to the nodes of the network and the significant associations to the edges. The network has a total of 102 edges among 47 nodes and consists of two separated subnetworks and a single TF pair. The width of the edges corresponds to the number of association measures supporting the edge (e.g. 1, 2, 3 or all 4). The first subnetwork is dominated by helicase-like transcription factor (HLTF), which has helicase activity and regulates the transcription of its target genes by altering chromatin structure (Maglott et al., 2011). Other highly connected TFs are members of forkhead box family (FOXL1, FOXQ1) or members of homeobox family (hepatic nuclear factors HNF1A, HNF1B, NKX3-1, NKX2-5, LHX3). The enriched functions of these transcription factors are 'organismal and cellular development' (functional analysis conducted with Ingenuity ® Systems (IPA)). The second subnetwork is dominated by the transcription factor AP2 (TFAP2) and by the transcription factor SP1. Both of them are general transcription factors which regulate a large number of genes and are involved in many cellular processes such as cell differentiation, cell growth and apoptosis (Maglott et al., 2011). Most of the co-occurring

TFs with TFAP2A and SP1 have a role in cellular differentiation and cell cycle progression (Ingenuity ® Systems, IPA; Maglott et al., 2011). 8 of 47 factors (17%) were found in other studies as promoter-centric or promoter-specific TFs (Neph et al., 2012; Whitfield et al., 2012), thus their frequent occurrence on general promoter regions is expected. Among our predicted co-occurring TF pairs, 4 TF pairs were found in experimental databases (Chatraryamontri et al., 2013) as directly interacting PPIs: SP1:YY1, SP1:MYC, SP1:ETS1, TFAP2A:YY1; see red edges in Figure 3.6.

3.6 Conclusion

In this chapter, we first introduced the representation of **transcription factors** as **ranked lists of their target genes**. With the usage of an affinity-based model for prediction of transcription factor binding to genomic sequences, we are able to construct for each transcription factor a ranked list of its target genes ordered by binding affinity. With this representation, we have shown that rank based association measures can be applied for prediction of co-occurring transcription factors on human promoter sequences, when corrected for their motif similarity. Interestingly, a large proportion of the significant transcription factors pairs are **consistent** for **four different association measures**: Spearman's correlation, Kendall's τ_K, R2KS and Fisher's exact test. The results derived with the irreproducibility discovery rate (IDR) do not have any agreement with the remaining association measures. The reason for this divergence might be the different application of the IDR which was not constructed to directly compare two ranked lists and to give a single association score to a pair of ranked lists. IDR was rather designed for assessing the reproducibility of scores for single items with replicative measurements, represented as ranked lists.

The **significant transcription factor pairs** which are supported by four different measures build a highly connected subnetwork with 13 transcription factors and 3 single transcription factor pairs with the promoter-specific transcription factor TFAP2A. Most of the factors, which were found to have co-occurring significant partners, have known functions related to the general differentiation of the cell, cell growth and apoptosis. Furthermore, 17% of the factors are known promoter-specific regulators which tend to bind to general promoters rather than to cell-type-specific distal regulatory regions. The co-occurrence of transcription factors on the tissue-specific and cell-type-specific regulatory regions is discussed in the following two chapters.

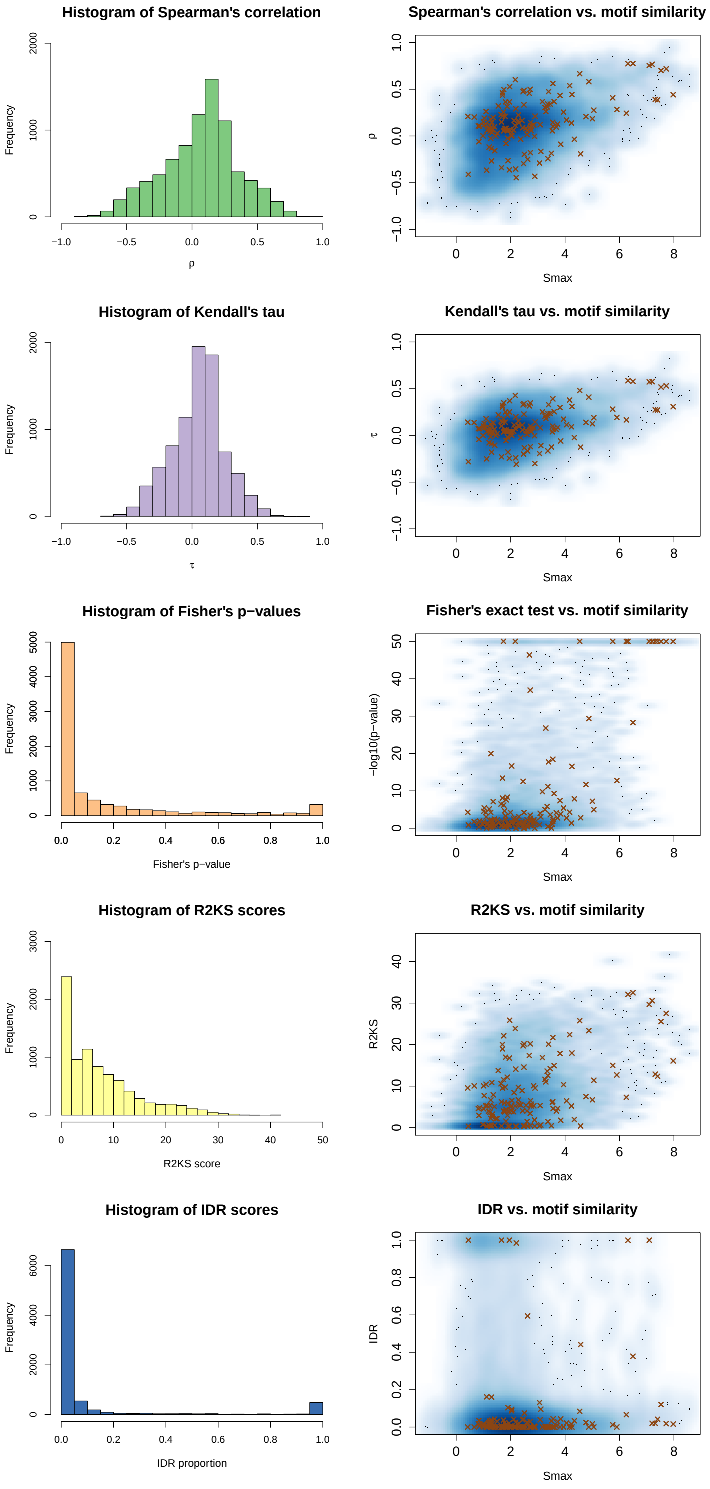


Figure 3.4: Histograms of association measures (left) and relation of association measures to motif similarity (right) for all TF pairs. From top to bottom: Spearman's correlation, Kendall's τ, minimal p-values of Fisher's exact test, R2KS measure and proportion of genes with IDR $\leq 10^{-10}$. Dark color in scatterplots indicates high density of data points, brown crosses highlight known PPIs.

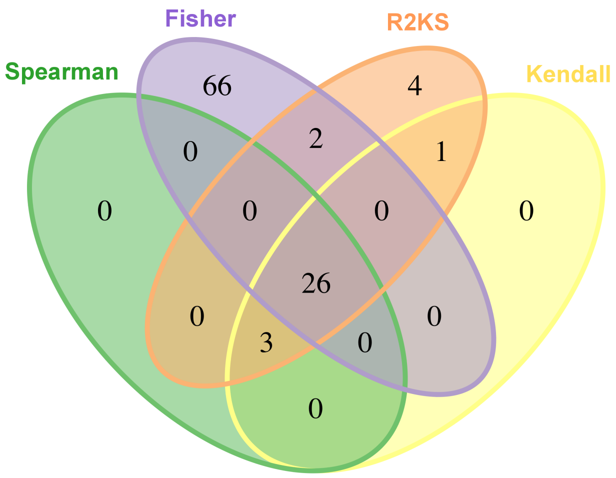


Figure 3.5: Venn diagram of predicted co-occurring TF pairs by four different methods.

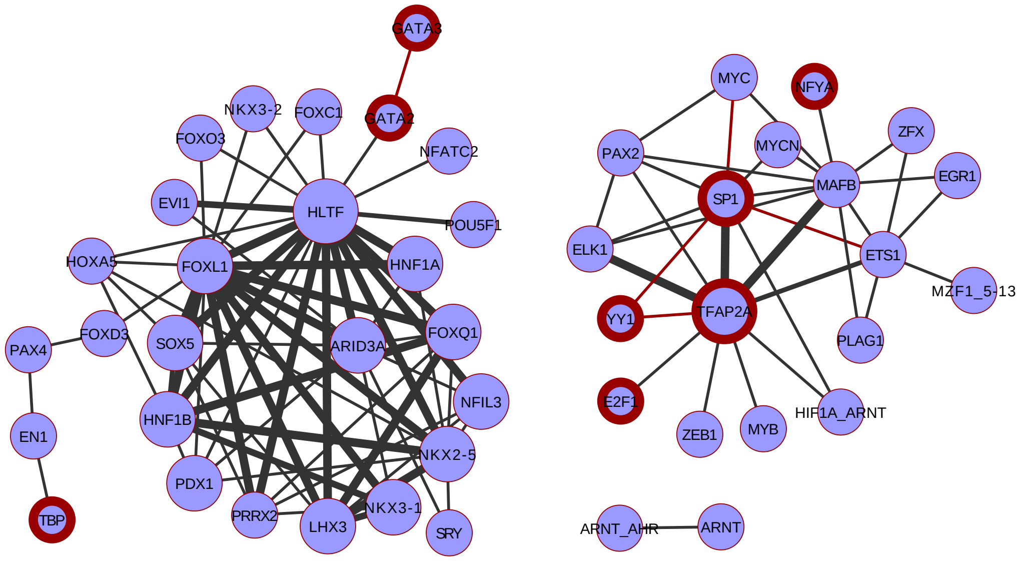


Figure 3.6: Network of significant TF pairs derived using four different association measures. The width of the edges corresponds to the number of association measures (e.g. 1, 2, 3 or 4) supporting the edge, red edges are known PPIs in databases, TFs with red border are known promoter-specific factors.

4 Prediction of transcription factor co-occurrence on tissue-specific promoters

4.1 Motivation

In Chapter 3, we showed that the prediction of transcription factor co-occurrence on promoter sequences using rank based association measures gives plausible results. However, even of larger interest is the **tissue-specific gene regulation**. The tissue-specific gene expression is regulated by an interplay of various transcription factors (Remenyi et al., 2004). The key question is how different combinations of these factors in different tissues influence the expression of their target genes. In particular, we want to investigate the co-occurrence of transcription factors in various tissues. And specifically, with the tissue-specific analysis of co-occurring transcription factors we want to answer the following questions:

- Are the main players different in different tissues?
- Or, are there some 'stable' players in most of the tissues and only their partners are changing?
- Or, are there distinct sets of cooperative factors in different tissues or tissue groups which define the tissue specificity?

There are many studies which investigate the regulatory networks in various tissues. One group of previous studies (Klein and Vingron, 2007; Smith et al., 2007; Yu et al., 2006) is based on common features in the promoter sequences of genes that are overexpressed in the tissue of interest. Another study (Hu and Gallo, 2010) analysed the functional conservation of various transcription factor binding sites in mouse and human to detect synergistic factors in functional pathways. Although these studies make plausible findings, they usually require large amounts of information apart from the DNA sequences, such as the conservation across species, gene expression measurements in various conditions or genes belonging to a functional group or pathway.

In the following, we present a method for the statistical prediction of **tissue-specific transcription factor co-occurrence**. To identify co-occurring TFs, we combine the predicted binding affinities of all possible pairs of TFs on their target genes and the information about the tissue-specificity of these genes. To determine the significance of the overlap of tissue-specific top-ranked target genes for pairs of different TFs we apply a **3-way contingency table test**. Our approach is based on the same assumptions as the general approach discussed in Chapter 3:

1. Two interacting TFs are expected to share a significant number of their target genes in comparison with two random target sets.
2. The list of target genes of a single TF can be represented by a ranked gene list based on the binding affinity of the TF to the promoter sequences.

To our knowledge, this is the first method which is able to predict transcription factor co-occurrence based only on the promoter sequence, its tissue-specificity information and TF-binding motifs.

In this chapter, we first present the 3-way contingency table, the types of **underlying independence models** and the construction of the test statistic (Section 4.2). In Section 4.3, the selection of the best underlying null model is discussed and evaluated on the distribution of p-values. Then, the co-occurring TF pairs in human tissues are predicted and validated by known protein-protein interactions (Section 4.4). Further, the predicted TF pairs in selected well-studied tissues (liver, muscle and hematopoietic stem cells) are discussed in more detail in Section 4.5. A comparison with different computational methods predicting tissue-specific co-occurring TFs is conducted in Section 4.6. Finally, the last Section 4.7 concludes the whole chapter.

4.2 Testing in 3-way contingency tables

By definition, the association of two ranked lists (or two random variables) partitioned into two categories can be depicted by 2-way contingency tables as described in Section 2.3.3. In the application to transcription factor co-occurrence prediction, the two variables are the ranked lists of binding affinities to promoter sequences of two transcription factors. In case of the *tissue-specific* co-occurrence, we additionally stratify by tissue. By introducing a third dimension in the contingency table leads to a 3-way contingency table. For this purpose, let us introduce the binary variable Z_t, which indicates the specific (e.g. overexpressed) states of each of the genes in a particular tissue t:

$$Z_t(i) = \begin{cases} 1, & \text{if gene } i \text{ is expressed specifically in tissue } t \\ 0 & \text{otherwise .} \end{cases} \tag{4.1}$$

Equivalently to Chapter 3, the ranked lists R_x and R_y are the lists of genes sorted according to the binding affinities of the first and second TF, respectively, to their promoter sequences. We can then easily define the corresponding binary variables X and Y indicating the genes ranked among the top-k_1 in R_x and among the top-k_2 in R_y, respectively. Formally, for each gene i, $i = 1, \dots, n$:

$$\begin{aligned} X_{k_1}(i) &= \begin{cases} 1 & r_x(i) \leq k_1, \ k_1 \in \{1, \dots, n\} \\ 0 & \text{otherwise} \end{cases} \\ Y_{k_2}(i) &= \begin{cases} 1 & r_y(i) \leq k_2, \ k_2 \in \{1, \dots, n\} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{4.2}$$

A graphic illustration of the setting of these three variables is shown in Figure 4.1. All n genes are shown as dots, blue ones indicate tissue-specific genes, where $Z_t(i) = 1$. The green set highlights the top-50 ranked target genes of the first TF with $X_{k_1=50}(i) = 1$ and the red set highlights the top-50 ranked target genes of the second TF with $Y_{k_2=50}(i) = 1$. The corresponding $2 \times 2 \times 2$ contingency table (with general count notation) is shown in Table 4.1, with the color coding of the random variables being identical with the color coding in Figure 4.1.

Table 4.1: $2 \times 2 \times 2$ contingency table for shared genes among the top-k_1 and top-k_2 ranked target genes of two different TFs and tissue-specific genes.

	tissue-specific		not tissue-specific		
	$\lvert i \rvert : r_y(i) \leq k_2$	$\lvert i \rvert : r_y(i) > k_2$	$\lvert i \rvert : r_y(i) \leq k_2$	$\lvert i \rvert : r_y(i) > k_2$	$\sum$
$\lvert i \rvert : r_x(i) \leq k_1$	n_{111}	n_{121}	n_{112}	n_{122}	n_{1++}
$\lvert i \rvert : r_x(i) > k_1$	n_{211}	n_{221}	n_{212}	n_{222}	n_{2++}
$\sum$	n_{+11}	n_{+21}	n_{+12}	n_{+22}	$n_{+++} = n$

Basically, we want to test for each tissue whether the number of genes in the intersection of all three variables, e.g. $n_{111} := \sum_i \mathbb{1}[X_{k_1}(i) = 1, Y_{k_2}(i) = 1, Z_t(i) = 1]$, is larger than

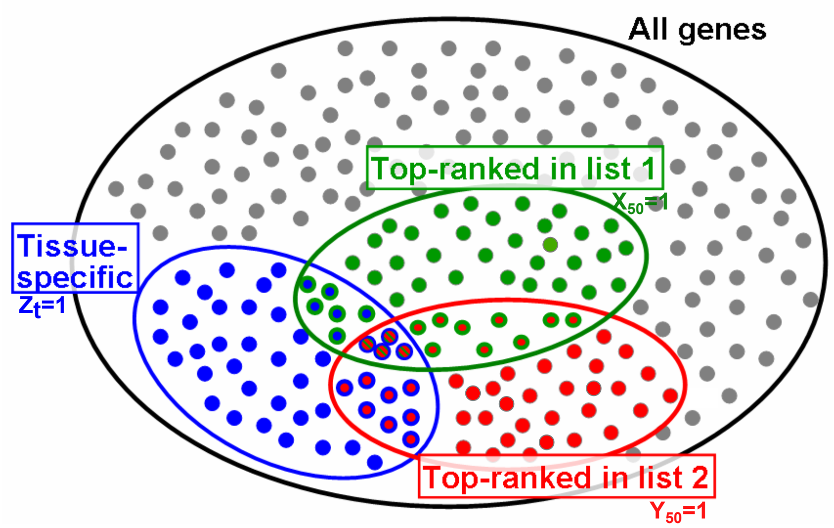


Figure 4.1: Venn diagram of the setting for independence tests in 3-way contingency tables. Grey dots indicate all human genes, blue dots are genes known to be specific for a selected tissue. Green and red sets denote the top-ranked target genes of the first and second TF, respectively.

expected by chance. This is in analogy to the null hypothesis of Fisher's exact test for 2×2 contingency tables, which was discussed in Section 2.3.3.

4.2.1 Notation for 3-way contingency tables

Before applying statistical testing in the 3-way contingency tables, we define some basic terminology and notation used in the analysis of multiway-contingency tables, based on Agresti (2013). Each cell of the contingency table has its probability π_{jkl} where in general $j = 1, \dots, J$; $k = 1, \dots, K$ and $l = 1, \dots, L$ with numbers of categories J, K, L of the three random variables X, Y and Z, respectively. In our case, we deal only with binary categories, thus $J = K = L = 2$. The only constraint on the cell probabilities is the total probability sum: $\sum_j \sum_k \sum_l \pi_{jkl} = 1$.

Then, the expected frequencies for each cell are $\mu_{jkl} = n\pi_{jkl}$ and the observed counts in each cell are denoted by n_{jkl} as in Table 4.1, with $n = \sum_j \sum_k \sum_l n_{jkl}$ being the total number of observations. The two-way marginals for the observed counts and for the cell probabilities are defined in analogy to the two-dimensional tables as:

$$\begin{aligned} n_{jk+} &= \sum_l n_{jkl} & \pi_{jk+} &= \sum_l \pi_{jkl}, \\ n_{j+l} &= \sum_k n_{jkl} & \pi_{j+l} &= \sum_k \pi_{jkl}, \\ n_{+kl} &= \sum_j n_{jkl} & \pi_{+kl} &= \sum_j \pi_{jkl}, \end{aligned} \tag{4.3}$$

and equally, the one-way marginals as:

$$\begin{aligned} n_{j++} &= \sum_k \sum_l n_{jkl} & \pi_{j++} &= \sum_k \sum_l \pi_{jkl}, \\ n_{+k+} &= \sum_j \sum_l n_{jkl} & \pi_{+k+} &= \sum_j \sum_l \pi_{jkl}, \\ n_{++l} &= \sum_j \sum_k n_{jkl} & \pi_{++l} &= \sum_j \sum_k \pi_{jkl} \; . \end{aligned} \tag{4.4}$$

4.2.2 Types of independence

With more variables in the contingency table, there are more potential hypotheses to test. Thus in a 3-way contingency table one can test four types of independence that are described below for binary random variables X, Y and Z. For further analysis, it is useful to express the independence models with a loglinear representation, which specifies the joint distribution among the random variables X, Y and Z that are cross-classified to form the table. In the following, we use the formulation from Agresti (2013, chap. 9).

Mutual independence

$\mathrm{H}_0 : X \perp\!\!\!\perp Y \perp\!\!\!\perp Z$

Under the mutual independence of all three variables, no relationship among X, Y and Z is assumed, so any pair of variables is independent. The cell probabilities in the contingency table under this null hypothesis are independent of the two-way marginals, such that:

$$\pi_{jkl} = \pi_{j++}\pi_{+k+}\pi_{++l} \quad \forall j, k, l. \tag{4.5}$$

For expected frequencies $\{\mu_{jkl}\}$, the mutual independence model has loglinear form:

$$\log \mu_{jkl} = \lambda + \lambda_j^X + \lambda_k^Y + \lambda_l^Z \quad \forall j, k, l, \tag{4.6}$$

with a constant λ, a row effect λ_j^X and column effects λ_k^Y and λ_l^Z. For the identifiability, some constraints on the effects are required, usually we set: $\sum_j \lambda_j^X = \sum_k \lambda_k^Y = \sum_l \lambda_l^Z = 0$.

Partial or joint independence

$\mathrm{H}_0 : (X, Y) \perp\!\!\!\perp Z$
Under the joint independence, one assumes that there is no relationship between the joint variable (X, Y) and variable Z, but a relationship between X and Y is possible although not required. The expected cell probabilities in the contingency table correspond then to an ordinary two-way independence between Z and a variable composed of the $J \cdot K$ combinations of categories of X and Y. This can be expressed as the following relation of the cell probabilities:

$$\pi_{jkl} = \pi_{jk+}\pi_{++l} \quad \forall j, k, l. \tag{4.7}$$

The loglinear model for the expected frequencies is:

$$\log \mu_{jkl} = \lambda + \lambda_j^X + \lambda_k^Y + \lambda_l^Z + \lambda_{jk}^{XY} \quad \forall j, k, l. \tag{4.8}$$

with a constant λ, a row effect λ_j^X and column effects λ_k^Y, λ_l^Z and with an association term λ_{jk}^{XY} reflecting the deviation from the mutual independence. The association term λ_{jk}^{XY} is affiliated with the conditionally dependent variables X and Y. An equivalent model can be constructed for the case of joint independence of variable X and joint variable (Y, Z) or variable Y and joint variable(X, Z), respectively.

Conditional independence

$\mathrm{H}_0 : (X \perp\!\!\!\perp Y)|Z$
Here, one assumes that any relationship between variables X and Y can be explained by variable Z. Thus, the independence of X and Y holds for each partial table with fixed category of Z. In terms of the conditional cell probability $\pi_{jk|l} := P(X = j, Y = k|Z = l)$ it holds:

$$\pi_{jk|l} = \pi_{j+|l}\pi_{+k|l} \quad \forall j, k, l. \tag{4.9}$$

For the cell probabilities over the entire table, there holds:

$$\pi_{jkl} = \frac{\pi_{j+l}\pi_{+kl}}{\pi_{++l}} \quad \forall j, k, l. \tag{4.10}$$

The conditional independence of X and Y given Z has loglinear representation

$$\log \mu_{jkl} = \lambda + \lambda_j^X + \lambda_k^Y + \lambda_l^Z + \lambda_{jl}^{XZ} + \lambda_{kl}^{YZ} \quad \forall j,k,l, \tag{4.11}$$

with a constant λ, a row effect λ_j^X and column effects λ_k^Y and λ_l^Z and with association terms λ_{jl}^{XZ} and λ_{kl}^{YZ} referring to conditionally dependent variables X and Z, and Y and Z, respectively. An equivalent model arises for conditional independence of variables X and Z given Y or of variables Y and Z given X.

No-three-factor interaction

$\mathrm{H}_0 : (XY, YZ, XZ)$
Under the no-three-factor interaction, no association among the three variables is assumed, which means in the log linear representation: $\lambda_{jkl}^{XYZ} = 0$. But, this model permits all three pairs of variables to be conditionally dependent, thus the loglinear model form is easy to derive using the association terms:

$$\log \mu_{jkl} = \lambda + \lambda_j^X + \lambda_k^Y + \lambda_l^Z + \lambda_{jk}^{XY} + \lambda_{jl}^{XZ} + \lambda_{kl}^{YZ} \quad \forall j,k,l, \tag{4.12}$$

with all three pairs of association terms λ_{jk}^{XY}, λ_{jl}^{XZ} and λ_{kl}^{YZ}. For this independence model, no closed form solution in terms of margins $\{\pi_{jkl}\}$ is available.

All these four types of independence with corresponding cell probabilities are summarized in Table 4.2. The mutual independence of all three variables is the strongest condition and it implies partial independence of any one variable from the other two. Conditional independence is a weaker condition than mutual or partial independence. If variable Z is jointly independent of variable X and Y, it implies that X and Z are conditionally independent given Y and Y and Z are conditionally independent given X.

4.2.3 χ^2 Goodness-of-fit test

To test whether an observed 3-way contingency table is consistent with a particular null hypothesis one commonly uses a goodness-of-fit test. The goodness-of-fit test compares the deviation of the observed cell counts n_{jkl} from the fitted expected counts $\widehat{\mu}_{jkl}$ based on the underlying null hypothesis. The fitted values are derived from the cell probabilities by setting $\pi_{jkl} = \mu_{jkl}/n$ and using the maximum likelihood (ML) estimates for the expected cell counts. For illustration, in case of the *mutual independence* Eq. 4.5 holds for the cell probabilities: $\pi_{jkl} = \pi_{j++}\pi_{+k+}\pi_{++l}$ and thus for the expected

Table 4.2: Independence models for 3-way contingency tables, its probabilistic form and fitted values

Interpretation	Symbol	Cell probabilities	Fitted values
Mutual independence Eq. 4.5	$X \perp\!\!\!\perp Y \perp\!\!\!\perp Z$	$\pi_{jkl} = \pi_{j++}\pi_{+j+}\pi_{++k}$	$\widehat{\mu}_{jkl} = \frac{\mu_{j++}\mu_{+k+}\mu_{++l}}{n^2}$
Partial independence Eq. 4.7	$(X,Y) \perp\!\!\!\perp Z$	$\pi_{jkl} = \pi_{jk+}\pi_{++l}$	$\widehat{\mu}_{jkl} = \frac{n_{jk+}n_{++l}}{n}$
Conditional independence Eq. 4.9	$(X \perp\!\!\!\perp Y)\|Z$	$\pi_{jkl} = \frac{\pi_{j+l}\pi_{+kl}}{\pi_{++l}}$	$\widehat{\mu}_{jkl} = \frac{n_{j+l}n_{+kl}}{n_{++l}}$
No-three-factor interaction Eq. 4.12	(XY, YZ, XZ)	no closed form	iterative methods

cell counts $\mu_{jkl} = \mu_{j++}\mu_{+k+}\mu_{++l}/n^2$. The ML estimates of the expected cell count marginals are derived from the loglinear models and are equal to the observed count marginals (Agresti, 2013, Chap. 9.6). So the fitted values for the mutual independence are: $\widehat{\mu}_{jkl} = n_{j++}n_{+k+}n_{++l}/n^2$ for all j, k, l.

For the no-three-factor interaction hypothesis, there is no explicit formulation of the cell probabilities in terms of marginal probabilities. For this reason, this independence model does not have a direct estimate for the fitted values $\widehat{\mu}_{jkl}$ and the ML estimation requires numerical iterative methods like Newton-Raphson method (Agresti, 2013, Chap. 4.6) or Iterative proportional fitting (Bishop, 1969; Deming and Stephan, 1940). The fitted values for all four independence models are summarized in Table 4.2.

After estimating the fitted cell counts, the goodness-of-fit statistics for a particular null hypothesis can be calculated. It is defined as the log-likelihood ratio of the model under the null hypothesis H_0 and of the saturated model based on the sample counts in the contingency table:

$$-2\log\frac{\mathrm{ML}_{\mathrm{H}_0}}{\mathrm{ML}_{\mathrm{Data}}} = 2\sum_{j,k,l} n_{jkl}\log\left(\frac{n_{jkl}}{\widehat{\mu}_{jkl}}\right) =: D(n,\widehat{\mu}) \tag{4.13}$$

This test statistic corresponds to the deviance measure $D(n, \widehat{\mu})$ and for large expected frequencies μ_{jkl} it follows approximately a χ^2-distribution under the particular null hypothesis H_0. The degrees of freedom equals the difference between the number of parameters in the general case ($df = JKL - 1$) and the number of parameters under the null hypothesis. So in case of binary categories of all three variables (e.g. $J = K = L = 2$), the *mutual* independence model has $df = JKL - J - K - L + 2 = 4$, the *partial* independence model has $df = (L-1)(JK-1) = 3$, the *conditional* independence has $df = L(J-1)(K-1) = 2$ and the *no 3-way interaction* has $df = (J-1)(K-1)(L-1) = 1$ degree of freedom.

4.3 Selection of the underlying null model

Before we predict the tissue-specific co-occurrence of transcription factors, the null model underlying the majority of the 3-way contingency tables for all TF pairs should be investigated. The choice of a proper null model is enormously important since the underlying model influence dramatically the derived p-values and therewith the obtained results. Thus, our strategy is to fit all possible null models to all TF pairs and then investigate the obtained distributions of the test statistics or p-values.

For random data, one expects a uniform distribution of p-values in the interval $[0, 1]$. Thus for a mixture of random data and of a biological signal, one supposes a slight enrichment of small p-values reflecting the non-randomness in the biological signal (Robins et al., 2000). Hence, we search for the model which is closest to such distribution.

First, we derived 3-way contingency tables for all TF pairs in all tissues using a threshold of $k_1 = k_2 = 1000$ top-ranked genes for both transcription factors in each pair. Then, we fitted all four models listed in Section 4.2.2, in our case namely

1. mutual independence of all three variables
2. joint independence of variables related to the transcription factors (X and Y) and the variable related to tissue-specificity (Z)
3. conditional independence of the two transcription-factor-related variables (X, Y) given the tissue-specificity variable Z
4. no-three-factor interaction among the 3 variables X, Y and Z

For each independence model, we calculated the deviance measure $D(n, \widehat{\mu})$ from Eq. 4.13 and the corresponding p-values for each TF pair. Then, the distributions of p-values in all four models were examined.

The histograms of the p-values in four selected tissues (muscle, liver, kidney and heart) are shown in Figure 4.2; each row panel corresponds to one independence model, each column panel to one of the selected tissues. The first row shows the p-value distribution for the mutual independence; whereas half of the p-values are distributed uniformly in the interval $[0, 1]$, half of the p-values show highly significant values $p \leq 0.05$. This property is consistent over all four tissues. Very similar behavior can be observed in the conditional independence model (third row in Figure 4.2). Here as well, half of the p-values are significant (≤ 0.05). On the other hand, the no-three-factor interaction model shows very different behavior. Only a very small part of the p-values (roughly 2%) are significant at the level of 0.05. Besides, the distribution of the p-values under this hypothesis follows a moderate bell curve instead of the expected flat uniformity. The model of partial independence has the closest distribution to a uniform distribution. The histograms in the second row show a moderate signal, roughly 20% of the p-values are significant and a majority of the p-values are uniformly distributed between 0 and 1.

From our findings described above we conclude that, the model of **partial independence** fits best the underlying 3-way contingency tables for all different TF pairs and studied tissues. The distribution of p-values consistently over all tissues, follows a uniform distribution on the $[0, 1]$ interval with a moderate enrichment of small p-values reflecting TF pairs with a large number of shared top-ranked tissue-specific genes. The distribution of p-values under the mutual and conditional independence is indeed close to the uniform distribution, however nearly one half of p-values is significant at level of 0.05. Usually, we do not expect that nearly half of the data show divergent behavior from the null model. For this reasons, we chose the model of partial independence for our further analysis performed in the following sections. Then, TF pairs with a significant p-value in a given tissue have strongly associated a TF-related joint variable (X, Y) and a tissue-related variable Z.

4.4 Prediction of tissue-specific transcription factor co-occurrence

4.4.1 Overview of the method

Our method for detecting tissue-specific transcription factor co-occurrence is summarized in Figure 4.3. First, for each TF separately, all promoter sequences (e.g. genes) are ranked by the binding affinity of the particular TF to these sequences. Second, for

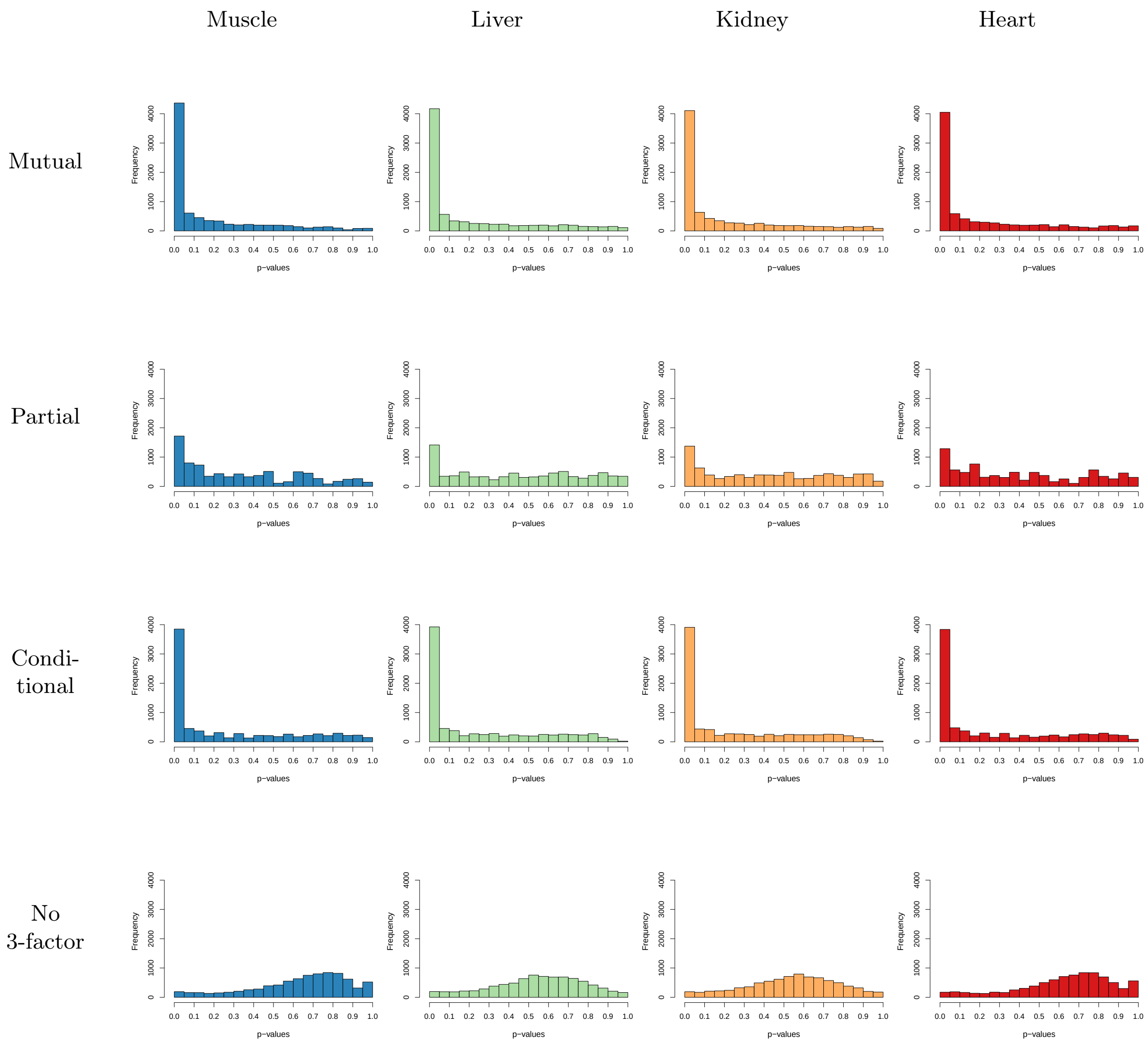


Figure 4.2: Histograms of p-values for four types of independence. From top to bottom in each row: mutual independence, partial independence, conditional independence and no-three-factor interaction in four different tissues: (a) muscle, (b) liver (c) kidney and (d) heart.

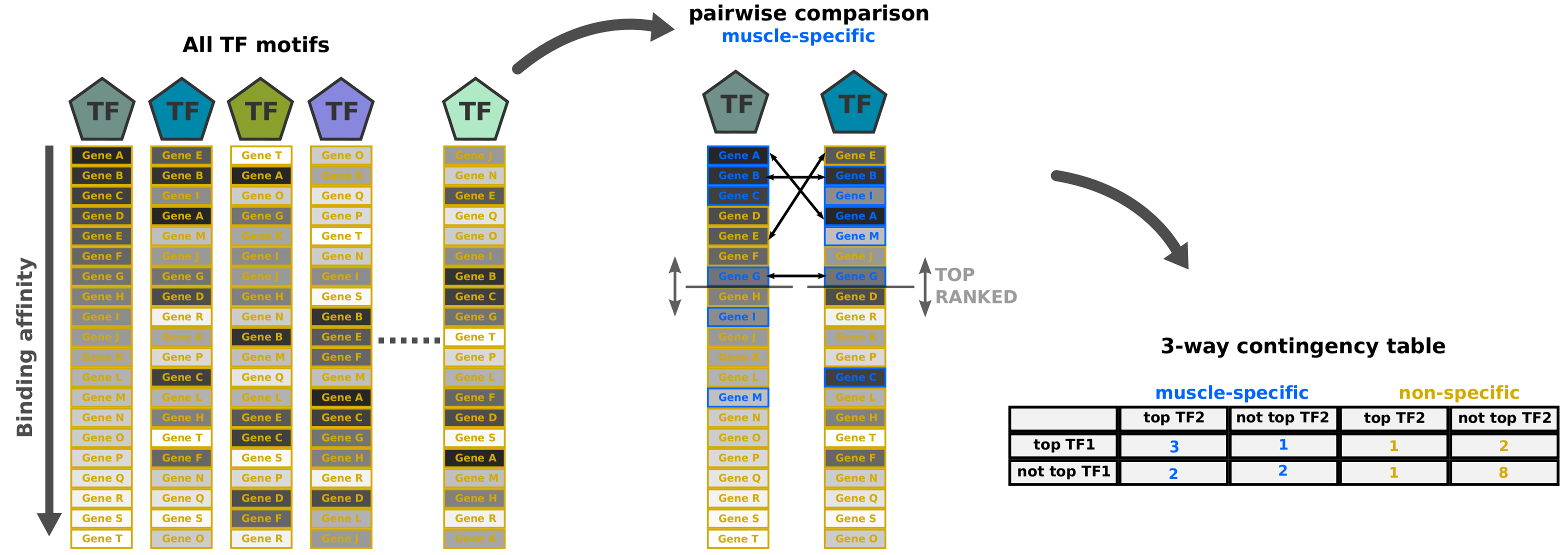
each particular tissue of interest, the tissue-specific genes are marked in the lists (e.g. the muscle-specific genes highlighted in blue in Figure 4.3). Next, all possible pairs of TFs are created and for each TF pair and each tissue, a 2 × 2 × 2-contingency table is derived. Then, the independence model of interest can be tested and a p-value can be assigned to the table. The most significant pairs for each tissue are marked as candidates of co-occurring transcription factors in the corresponding tissue and a network out of these TF pairs is created.

To define the tissue-specific genes, we made use of the data published by Yu et al. (2006) in 30 human tissues and data from Gupta et al. (2005) for 4 homogenous cell lines. Both datasets are based on expression enrichments values for expression sequencing tags (ESTs). Yu et al. (2006) calculates for each gene and each tissue the expression enrichment ratios between the observed and the expected number of ESTs. Assuming a binomial distribution of the expression enrichment ratios, the corresponding p-value for each gene and tissue can be derived. The tissue-specific genes are then defined as those with large expression enrichment ratios ($EE > 5$) and small p-values ($p < 10^{-3.5}$).

Gupta et al. (2005) evaluate the tissue specificity of a given gene by computing a p-value which reflects the overrepresentation of ESTs from a tissue among all ESTs of a given EST cluster. For our analysis, only EST clusters with p-value $< 10^{-6}$ in at least one of the tissue categories were utilized.

The number of tissue-specific genes varies from 58 for uterus to 1409 for lymphocyte. These are relatively small numbers in comparison with the total number of promoters (42 380) listed in the *hg19* assembly of the human genome from genome.ucsc.edu.

For the definition of tissue-specific genes, we chose data based on the ESTs analysis. The ESTs are short sequence tags connected to the TSS of each gene, thus it is easy to compare the expression measurements over all genes and various tissues. As an alternative, one could use direct approaches such as gene expression microarrays measuring the change of expression between two conditions or direct RNA-sequencing of tissue samples.



	muscle-specific		non-specific	
	top TF2	not top TF2	top TF2	not top TF2
top TF1	3	1	1	2
not top TF1	2	2	1	8

Figure 4.3: Overview of the method for the detection of tissue-specific co-occurrence of transcription factors on promoters. First, for each TF in the database, promoter sequences are ranked by the binding affinity. Then for the particular tissue of interest (e.g. skeletal muscle), the tissue-specific genes are identified (highlighted in blue). For all possible pairs of TFs, the corresponding 3-way contingency table is derived and the independence model of interest is tested.

4.4.2 Predicted co-occurring TF pairs

For the prediction of the tissue-specific co-occurring transcription factors, we need to specify the thresholds k_1 and k_2 defining the top-ranked genes for each transcription factor in the pair. To keep the balance between the relevance of the biological information and statistical significance we fix the thresholds of the top-target genes for both transcription factors in the pair to $k_1 = k_2 = 1000$. We can assume that an average transcription factor regulates 1000 or more genes in a specific tissue (Chua et al., 2006). However, we do not want to set the top-ranked genes threshold too large to avoid a large number of possible false positive results. With increasing thresholds k_1 and k_2 the chance to obtain a significant result increases, because increasing thresholds $(k_1, k_2 \longrightarrow n)$ more and more genes fall into the intersection of variables X and Y. For these reasons, we chose the cutoff for defining the top-ranked genes for all ranked lists of target genes to $k_1 = k_2 = 1000$.

Taking the most significant TF pairs, we identify 594 TF pairs in 4 specific cell lines (with p-value $\leq 10^{-11}$) and 409 TF pairs in 12 human tissues (with p-value $\leq 10^{-6}$). For the reasons discussed in Section 3.3, we focus on TF pairs with nonsimilar motifs (e.g. with motif similarity $S^{\max} <$ 90%-quantile$(S^{\max}) = 3.9$). The majority (869; 86.6%) of the significant TF pairs are between TFs with nonsimilar motifs. This points to a strong association between the TF pairs that is not due to a high similarity of their binding motifs.

Tissues with the highest number of identified TF pairs are retinal pigmented epithelium (259), lymphocyte (181) and liver (106). 181 TF pairs are significant simultaneously in two or more different tissues. 61 out of these are common in kidney and liver and 43 TF pairs are common in hematopoietic stem cells and lymphocytes. In both cases, there are tissues or cell lines with related molecular functions.

In 18 tissues, we did not find any significant TF pairs with p-value smaller or equal 10^{-6}. We searched then for TF pairs with slightly larger p-value $\in (10^{-6}, 10^{-5}]$ and found additional 58 interactions, 17 of them in another 6 tissues. No significant TF pairs with p-value $\leq 10^{-5}$ were found in the following 8 tissues: bone marrow, mammary gland, ovary, prostate, skin, soft tissue, thymus and uterus.

Altogether, we conducted 8835 tests in each studied tissue. With the choice of a relative strict p-value threshold of 10^{-5} we limit the number of possible false positives such that an additional multiple testing correction was not necessary. The number of tissue-specific TFs and the number of co-occurring TF pairs including the three most significant TF pairs for each of the 22 tissues are summarized in Table 4.3.

Table 4.3: Summary of the predicted tissue-specific TF pairs. For each of the 22 human tissues the table lists the 3 most significant TF pairs, the number of transcription factors in each tissue-specific network and the central nodes of the network.

Tissue	♯ interactions (nonsimilar)	♯ factors	Top three interactions (nonsimilar)	Hubs
Bladder [a]	3 (3)	3	ELK1:NFYA,ELK1:NOBOX, NFYA:NOBOX	–
Blood	6 (4)	5	SPI1:ARID3A, SPIB:ARID3A, SPI1:CTCF	ARID3A, SPI1, SPIB
Bone	24 (24)	25	TBP:TFAP2A, TBP:EWSR1-FLI1, TBP:NOBOX	TBP
Brain	25 (17)	20	SP1:SOX10, SP1:ESR2, SP1:REST	MZF1-1-4, SP1
Cervix	40 (30)	24	ZFP423:ZFX, ELK1:ZFX, MIZF:ZFX	ZFX, KLF4, ZFP423
Eye [a]	4 (2)	6	T:HNF1B, SP1:TAL1-TCF3	SP1
Heart [a]	6 (5)	7	MEF2A:MAFB, MEF2A:NFKB1, MEF2A:REST	MEF2A
Kidney	95 (87)	64	GATA1:HNF1A, HNF1A:ARID3A, TP53:HNF1B	HNF1A, HNF1B
Liver	106 (99)	67	HNF1A:HNF1B, HNF1A:HNF4A, HNF1A:CEBPA	HNF1A, HNF1B
Lymph node	64 (57)	65	SPI1:MZF1-1-4, SPI1:MYF, SPI1:FOXQ1	SPI1
Muscle	41 (38)	40	MEF2A:ZFP423, MEF2A:NHLH1, MEF2A:NFIC	MEF2A, TBP
Pancreas[a]	14 (13)	15	AR:TAL1-GATA1, MZF1:TAL1-GATA1, E2F1:TAL1-GATA1	TAL1-GATA1
Placenta [a]	2 (2)	4	RREB1:PDX1, ESRRB:POU5F1	–
PNS [a]	1 (0)	2	ELK4:REL	–
Small intestine [a]	1 (1)	2	NFYA:TBP	–
Stomach [a]	7 (6)	8	EWSR1-FLI1:PLAG1, MYC:PLAG1, PAX6:PLAG1	PLAG1
Testis [a]	16 (14)	19	FOXC1:HOXA5, ARNT-AHR:NOBOX, ARNT-AHR:NKX2-5	ARNT-AHR
Tongue [a]	12 (11)	16	NFKB1:NFIL3, NFKB1:TFAP2A, GATA3:NKX3-1	NFKB1
Adipose [b]	104 (90)	46	MZF1:NFYA, NFYA:MYB, NFYA:TBP	NFYA, MZF1
Lymphocyte [b]	181 (156)	107	ELK1:CEBPA, ELK1:FOXA2, ELK1:POU5F1	NFYA, ELK1,GABPA
HSC [b]	50 (41)	36	ELK1:NFYA, NFYA:GABPA, ELK1:EGR1	NFYA, ELK1
Retinal pigm. epithelium [b]	259 (219)	116	ARNT-AHR:CREB1, ARNT-AHR:NFYA, CREB1:BRCA1	CREB1, NFYA, PAX2

[a] Network predicted with $p \leq 10^{-5}$

[b] network predicted with $p < 10^{-10}$

4.4.3 Evaluation of the predicted TF pairs by known protein-protein interactions

One of the possibilities to evaluate the biological relevance of the predicted co-occurring transcription factors is to compare the co-occurring TF pairs with known interactions from protein-protein-interaction (PPI) databases. Namely, when the two transcription factors bind to the same promoter within relatively small distance they might physically interact and together regulate the transcription of their target gene.

We compared our significant TF pairs to the BioGRID database of PPIs (Stark et al., 2011) derived from various experiments including yeast-two-hybrid (Y2H) screens, mass spectrometry and others. Further, we included the interactions between human and mouse transcription factors validated by Ravasi et al. (2010) with mammalian-two-hybrid (M2H) screens.

For the evaluation, we calculate the ratio of the PPIs from the databases in the set of our candidates and of the total number of our candidates. Then, 4.2% of predicted tissue-specific co-occurring TF pairs are also physically interacting proteins, this corresponds to 1.8-fold enrichment compared to a random set of TF pairs and the corresponding p-value of the Fisher's exact test is $p = 8.4 \cdot 10^{-4}$.

Further, we also calculate this ratio of known PPIs among the significant TF pairs separately for each tissue, the enrichment is visualised in the barplot in Figure 4.4. In the tissues *eye,blood, bone* and *brain* the percentage of known PPIs is more than 7-fold higher than in a randomly chosen set of TF pairs. A moderate enrichment of known PPIs can be observed in *tongue, lymph node, hematopoietic stem cell, cervix, muscle, adipose, kidney, retinal pigment epithelial, liver* and *lymphocyte*. However, there are 8 tissues (*bladder, pancreas, stomach, testis, heart, placenta, peripheral nervous system* and *small intestine*) where no PPIs were found in the database.

One reason for the absence of known PPIs in some groups of significant tissue-specific TF pairs may be the incompleteness of the experimental databases. Usually, there are preferred groups of proteins and TFs which are well studied and easy-to-prepare in the laboratory. Such proteins have more experimentally validated interaction partners than the less studied proteins. Furthermore, there are many proteins and TFs for which the validation experiment cannot be performed due to technical difficulties. On top of that, the experimental techniques used for the PPI validation have very low sensitivity and precision (Berggård et al. (2007); Ravasi et al. (2010), see Section 1.3.1), such that many interactions cannot be detected.

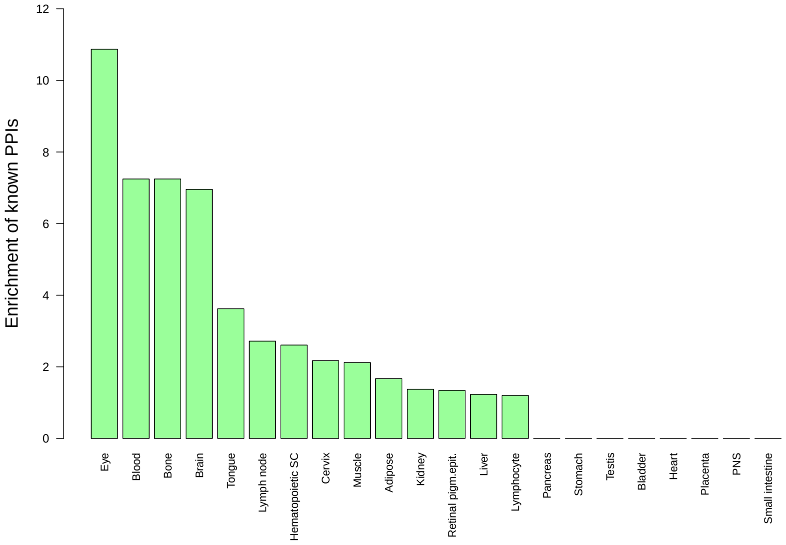


Figure 4.4: Enrichment of known protein-protein interactions among significant TF pairs in 22 tissues.

4.5 Predicted co-occurring TF pairs in selected tissues

In this section we present our predictions of co-occurring TFs in three well-studied homogenous human tissues. Since a lot of information is provided in the literature for these tissues, we are able to validate our predictions indirectly with related experimental data.

4.5.1 Predicted TF pairs in liver

In analogy to the analysis of the association measures of two ranked lists in Section 3.5, we investigate the relationship between the p-values of the 3-way contingency table test and the PWM similarity measure, see Figure 4.5. Due to the stratification by tissue the distribution of the data points changes in comparison with the general case shown in Figure 3.4. Now, the highly significant TF pairs do not localize in the upper right corner corresponding to the highly similar PWMs. On the contrary, there is a group of highly significant TF pairs of non similar binding motifs shown as a cloud of significant TF pairs with motif similarity smaller than 4, see Figure 4.5. However, looking at known protein-protein interactions in the scatterplot, the majority of them do not have significant p-values in liver (see red crosses in Figure 4.5). The following 5 TF pairs constitute the exception: HNF1A:HNF1B, HNF1A:HNF4A,HNF1B:CREB1, HNF1A:CEBPA and

HNF1A:STAT3 (see circled crosses in Figure 4.5). These interacting proteins are known regulators in liver which in turn suggests that the predicted interactions are real.

With a threshold of p-value $\leq 10^{-6}$, we define 106 significant TF pairs in liver. The network created out of all these TF pairs is shown in Figure 4.6. Here, solid edges indicate 98 TF pairs with low motif similarity $S^{\max} < q_{90\%}(S^{\max}) = 3.9$, the remaining 8 edges (dashed lines) are between TFs with highly similar motifs.

We have further investigated the 67 TFs (nodes) in our predicted network. We found support in the literature (Heinemeyer et al., 1998; Ingenuity ® Systems, IPA; Matys et al., 2006) for 9 (13.4%) TFs in the network (CEBPA, HNF1A, HNF1B, HNF4A, NR2F1, NFKB1, POU5F1, RELA and RXRA) to be known transcriptional regulators in liver (see red nodes in Figure 4.6). Further, Krivan and Wasserman (2001) defines 4 TFs (HNF3, HNF1, HNF4 and CEBP) as critical regulators in liver, where the last 3 of these factors as well as the connections between them are found in our liver regulatory network too. We identified identical central regulators (HNF1A, HNF4A) in the network according to Odom et al. (2004), these two central hubs have the highest number of TF partners in our predicted network too. Further, HNF1A and HNF4A were identified by the experimental work of Ravasi et al. (2010) as *specifier* in liver, e.g. transcription factors with high specific expression in liver. In addition, we conducted an expression analysis of the TFs in the predicted liver network. The majority (59.7%) of nodes (light green nodes in Figure 4.6) were shown experimentally to be expressed in liver tissue (Matys et al., 2006; Parkinson et al., 2011; Ravasi et al., 2010).

3 of the significant TF pairs (HNF1A:HNF1B, HNF1A:HNF4A, HNF1A:CEBPA; red edges in the network in Figure 4.6) are known interacting proteins in the PPI databases mentioned in Section 4.4.3. Next, we investigated the presence of a common known interacting factor of the two transcription factors in any significantly co-occurring pair. 9 TFs in the liver network share such a common interacting factor (denoted as orange edges in Figure 4.6). As an example, there are known PPIs between HNF1A and CEBPA and SOX10 and CEBPA, such that a possible interaction (and co-occurrence) of HNF1A and SOX10 is likely to exist too.

Next, we searched with Ingenuity Pathway Analysis (Ingenuity ® Systems (IPA)) for molecular functions and pathways in which the TFs from our liver network are involved. As expected, we found as the most enriched function *transcriptional regulation* and *DNA-binding*. Other interesting molecular functions were found as significantly enriched, here shown with the corresponding p-value and involved transcription factors: *development of liver* ($p = 1.37 \cdot 10^{-6}$; CEBPA, HNF1A, HNF1B, PDX1,RELA), *proliferation of hepatocytes* ($p = 5.71 \cdot 10^{-4}$; CEBPA, HNF1A, NFE2L2, NFKB1) and *liver* *hepatitis* ($p = 1.31 \cdot 10^{-2}$; ESR2, NFE2L2, PDX1, RELA). Further, transcription factor NFE2L2 is a known regulator in lipid metabolism and hepatic system development. Transcription factor RELA has a known function in liver, it regulates the degeneration of liver. In our liver network, we predict significant TF pairs between NFE2L2 and both central hubs HNF1A and HNF1B and as well significant TF pairs between RELA and these both central hubs. Known regulatory functions in liver of NFE2L and RELA indicate that there is a possible functional co-occurrence of these two factors with the central regulators HNF1A and HNF1B.

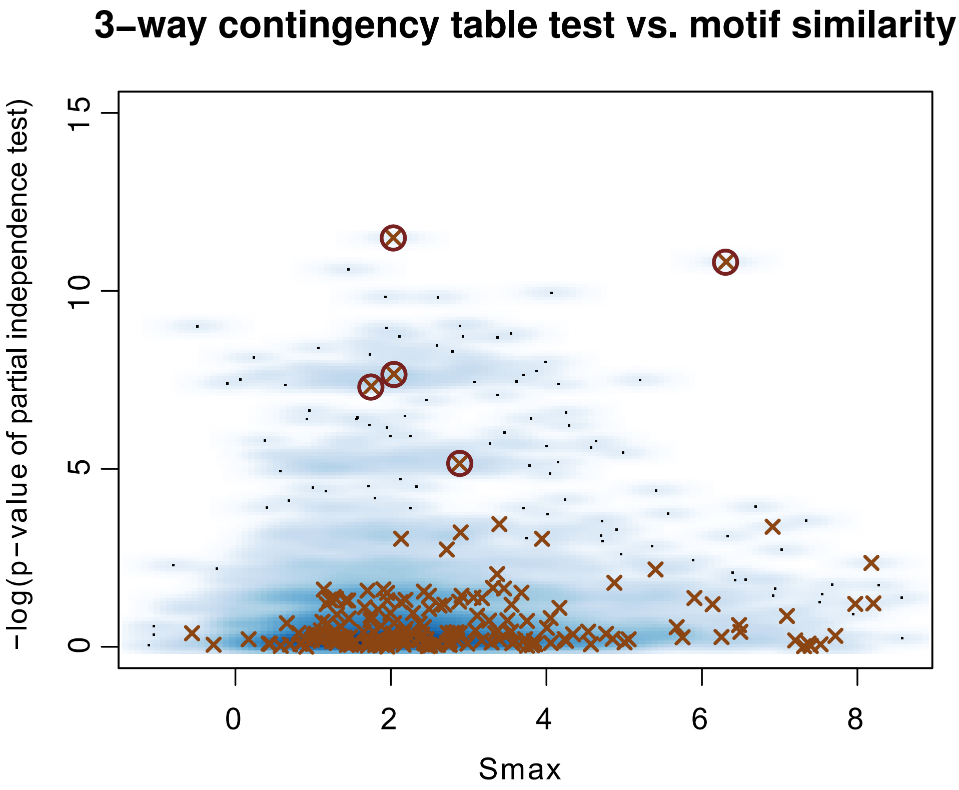


Figure 4.5: Relation of the significance value in the 3-way contingency table and the motif similarity in liver. Pairs of known interacting transcription factors are highlighted with red crosses.

4.5.2 Predicted TF pairs in skeletal muscle

The significant co-occurring TF pairs in skeletal muscle with a threshold p-value $\leq 10^{-6}$ result in a network with 41 TF pairs (edges) involving 40 transcription factors (nodes), see Figure 4.7. 38 of the predicted pairs are between TFs with nonsimilar motifs (solid edges). The remaining 3 TF pairs are between TFs with similar motifs (dashed edges).

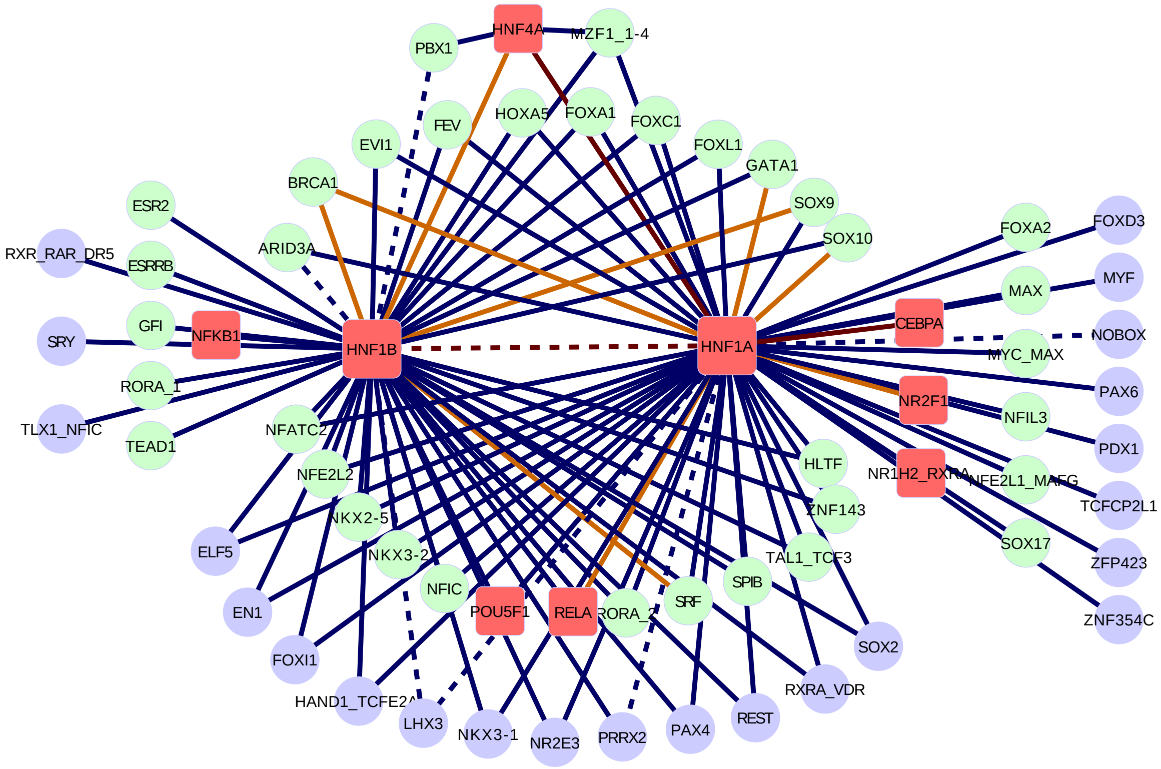


Figure 4.6: Network of significant TF pairs in liver. Red nodes are known regulators in liver, green nodes denote genes expressed in liver. Factors with similar binding motifs are connected with dashed edge, factors with non similar binding motifs with a solid edge. Red edges are known protein-protein interactions, orange edges indicate a common interacting partner of both nodes.

9 of 40 (23%) TFs in the skeletal muscle network (GATA2, MEF2A, MYF, NFIL3, PAX2, PAX6, SP1, SRF and TBP) are known regulators of gene expression in muscle (Bentzinger et al., 2012; Sartorelli and Caretti, 2005; Smith et al., 2007). In our predicted network, there are two central regulators: MEF2A and TBP. MEF2A is the central hub with the highest number of predicted co-occurring partners (36), whereas TBP with co-occurring general regulators TFAP2A and SP1 is a center of a smaller network, which is related to general tissue development. Both of the central factors (MEF2A, TBP) were classified as *facilitator* hubs by Ravasi et al. (2010), meaning that they are central regulators with widespread expression over many different tissues. We found evidence of expression in skeletal muscle for additional 22 (55%) factors (green nodes) from our network (Matys et al., 2006; Parkinson et al., 2011; Ravasi et al., 2010). Two co-occurring TF pairs (MEF2A:TEAD1 and TBP:SP1) in the muscle network are already known direct interactions (see red edges in Figure 4.7). Four of the significant TF pairs share a common known interacting TF partner, we call such connection a *trio* and denote them with an orange edge as in the liver network. The knowledge of a known shared interacting transcription factor increases confidence in the validity of our predicted TF pairs. Factors TBP and TFAP2A have two known shared co-factors: MYC and TP53 (these connections are shown as grey edges in the network, because they are not part of our predicted network). Three predicted TF pairs SRF:TBP, SRF:MEF2A and TBP:MEF2A all share a common co-factor TEAD1 (see grey edges in Figure 4.7). As found in literature, SRF, TBP and MEF2A are all known regulators in skeletal muscle, thus there is a high probability that TEAD1 can build a complex with these regulators and have a regulatory function in muscle, too. For the TF pair SRF:MEF2A, we found an experimental evidence of physical interaction between SRF and MEF2A in mouse (West et al., 1997).

We searched for the overrepresented molecular functions in which the transcription factors from the muscle network are involved. The analysis was performed with Ingenuity ® Systems (IPA). 8 TFs in the network control the differentiation of muscle cells ($p = 9.4 \times 10^{-09}$; MIZF, MEF2A, MYF5, NFIC, REST, SRF, STAT1, TP53); 6 TFs in the network are involved in the differentiation of muscle cell lines ($p = 8.1 \times 10^{-08}$; EWSR1, FLI1, MYF5, NFKB1, STAT1, ZNF423). We found two more general functional categories, that are related to molecular processes in muscle: apoptosis of fibroblast cell lines ($p = 1.31 \times 10^{-09}$; AHR, EGR1, EVI1, EWSR1, FLI1, NFE2L2, NFKB1, RELA, STAT1, TP53) and development of organs ($p = 8.5 \times 10^{-20}$; AHR, ARNT, EGR1, EVI1, FLI1, FOXD3, FOXQ1, GATA2, NFKB1, NOBOX, NR2F1, PAX2, PAX6, PLAG1, RELA, RORA, SOX2, SP1, SRF, TEAD1, TFAP2A, TP53, YY1, ZFX, ZNF423).

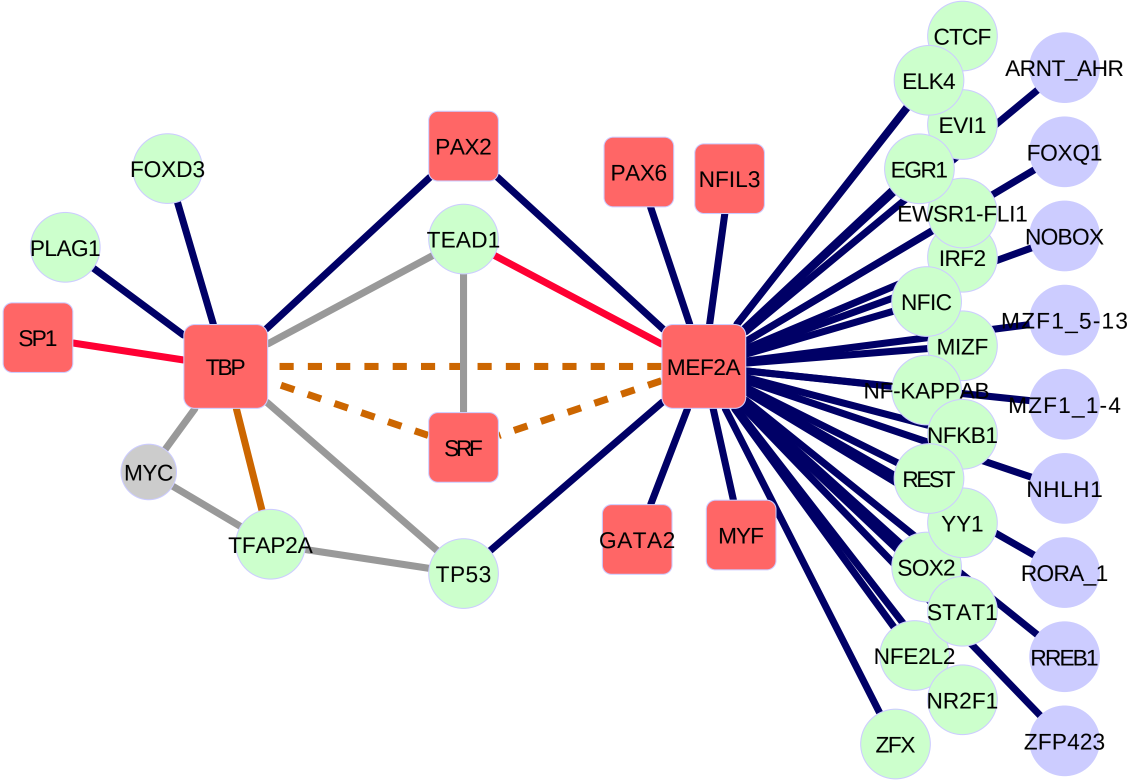


Figure 4.7: Network of significant TF pairs in skeletal muscle. Red nodes are known regulators in muscle, green nodes denote genes expressed in muscle. Factors with similar binding motifs are connected with dashed edge, factors with non similar binding motifs with a solid edge. Red edges are known protein-protein interactions, orange edges indicate a common interacting partner of both nodes.

4.5.3 Predicted TF pairs in hematopoietic stem cells

To generate a network with significant TF pairs in hematopoietic stem cells (HSCs), a threshold of p-value $\leq 10^{-11}$ was used. More strict threshold for this cell line was selected because of the large number of cell-line-specific genes (678) in HSCs which induce higher number of significant TF pairs. The predicted network of significant TF pairs consists of 50 TF-pair connections among 36 TFs, shown in Figure 4.8. 41 (82%) TF pairs are between TFs with nonsimilar motifs. The remaining 9 TF pairs are between TF with similar motifs. Similarly to the other two studied tissues, there are two subnetworks with two central hubs: ELK1 and NFYA. These two factors were identified as *facilitator*

hubs (i.e. central regulators with widespread expression) by Ravasi et al. (2010). Both ELK1 and NFYA, together with 13 other factors (ARNT/AHR, ELK4, ELF5, EGR1, GABPA, IRF1, IRF2, MYF, POU5F1, SPI1, USF1, TBP and ZFP423) are known regulators in hematopoiesis. These are 41.7% of factors in the predicted network, see red nodes in Figure 4.8. We found that further 13 (36.1%) TFs are expressed directly in HSCs or in bone marrow, where the HSCs originate from (Matys et al., 2006; Parkinson et al., 2011; Ravasi et al., 2010).
There are 4 TF pairs which are known as directly interacting proteins (ELK1:KLF4, NFYA:ELK4, NFYA:SPI1, NFYA:CREB1). Moreover, 12 TF pairs in the network share one or more of the interacting factors: BRCA1, SP1, SRF, TP53 as common co-factors (for a better representation, these interactions are not shown in the network). Half of the factors (CREB1, CTCF, E2F1, EBF1, EGR1, ELK1, ELK4, GABPA, HIF1A, HNF1A, IRF1, IRF2, KLF4, MYB, NFYA, PBX1,RXRA-VDR, SPI1) in the predicted network play a role in the hematopoiesis, as shown with Ingenuity ® Systems (IPA) analysis ($p = 7.19 \times 10^{-17}$). Further, 13 factors (CREB1, E2F1, EBF1, EGR1, ELK1, ELK4, GABPA, HIF1A, HNF1A, IRF1, IRF2, MYB, SPI1) from the network have function in the development of lymphocytes and leukocytes ($p = 1.77 \times 10^{-11}$), one of the processes that take place in the HSCs.

4.6 Comparison of predicted TF pairs obtained with different computational methods

The results of our study suggest that the gene expression in various tissues or cell types is regulated by a large number of tissue-specific TF pairs which are dominated by only a few central regulators. The experimental findings of Ravasi et al. (2010) confirmed the central hubs in various tissues which were detected with our methodology. Ravasi et al. further separate the hubs by their expression specificity into *specifier* (with tissue-specific expression) and *facilitator* (with wide expression over tissues). In our networks of significant TF pairs, we could find both of these groups of central hubs. In the following, we compare our predicted TF pairs in liver, muscle and hematopoietic stem cells with predictions obtained with two different computational methods predicting tissue-specific interactions of TFs.
The first method from Yu et al. (2006) first searches for highly significant hits of transcription factor binding sites in human promoters. Then, Yu et al. (2006) evaluate the relations between all TF pairs with the co-occurrence of their highly significant binding sites and their relative positions in promoters of tissue-specific genes. In contrast to our

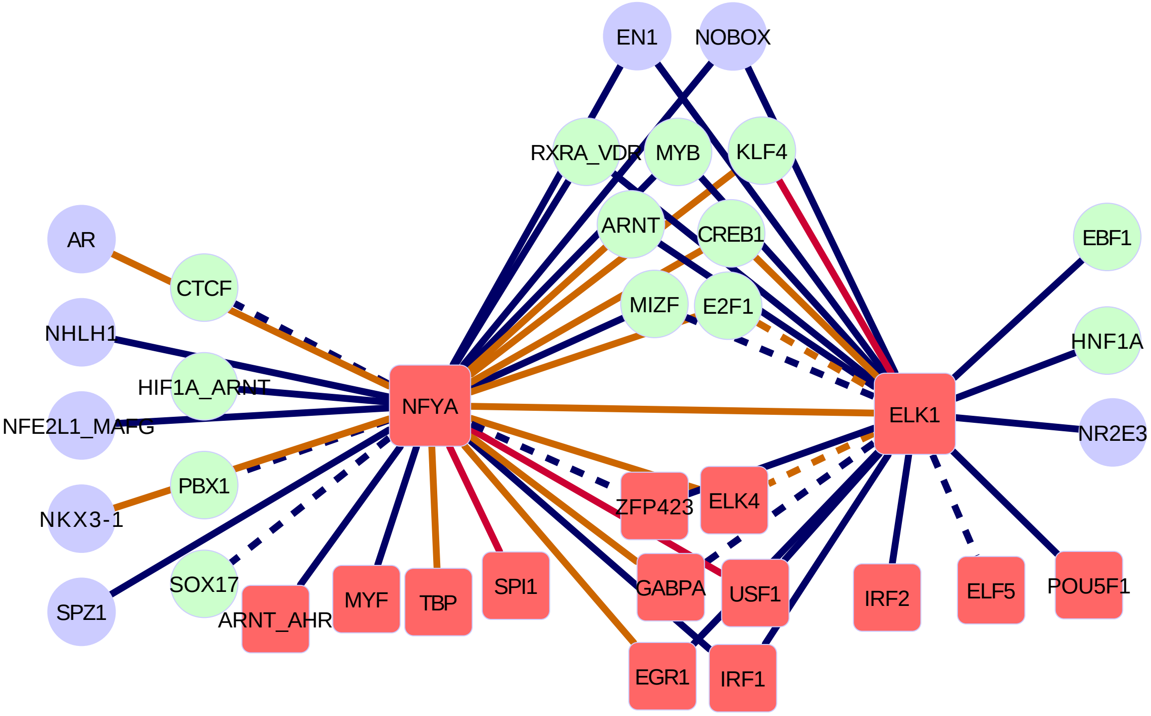


Figure 4.8: Network of significant TF pairs in hematopoietic stem cells. Red nodes are known regulators in hematopoiesis, green nodes denote genes expressed in white blood cells. Factors with similar binding motifs are connected with dashed edge, factors with non similar binding motifs with a solid edge. Red edges are known protein-protein interactions, orange edges indicate a common interacting partner of both nodes.

analysis, another database of PWMs from TRANSFAC (Matys et al., 2006) was used. In liver, we identified 11 TF pairs predicted by Yu et al. that are in agreement with our significant TF pairs (HNF1:NFIL3, PBX1:HNF1, HNF4:HNF1, HNF4A:HNF1, HNF1:FOXC1, CEBPA:HNF1, FOXD3:HNF1, HNF1:NKX2-2, HNF1:FOXL1, HNF1:NKX3-A, RORA1:HNF1). The pair HNF1:NFIL3 belongs to the top three liver interactions defined by Yu et al. In concordance with our predicted network in liver (with two main central regulators HNF1A and HNF1B), HNF1 is the central regulator in liver in Yu et al. too.

The comparison of the muscle-specific TF regulatory networks gives very similar results. 8 of our predicted co-occurring TF pairs in muscle (MYF:MEF2, TBP:MEF2, SRF:MEF2, SRF:TBP, RREB1:MEF2, PAX2:MEF2, NFkB:MEF2, TBP:TFAP2A) could be found in the muscle-specific network identified by Yu et al. Among them, the TF pair MYF:MEF2 is one of the top three interactions identified by Yu et al. The central regulator in muscle predicted by Yu et al. (2006) is MEF2 which corresponds to our main central hub MEF2A.

A direct comparison of the results in hematopoietic stem cells is not possible, because Yu et al. (2006) do not provide an analysis of this cell line. Therefore, we examined bone marrow tissue provided from Yu et al., which is the most related tissue to HSCs, as HSCs originate from bone marrow. 5 of our predicted TF pairs in HSCs (ELK1:GABPA, ELK1:CREB1, ELK1:NFY, ELK1:MYB1, NFY:VDR) could be found in the bone-marrow-specific regulatory network from Yu et al.

The second method from Hu and Gallo (2010) makes use of the evolutionary conservation of biological function and high expression level of genes in human tissues to predict TF pairs which control tissue-specific gene expression. In general, the predicted networks in skeletal muscle and in liver by our methodology and by Hu and Gallo differ a lot. We could identify only two of our predicted TF pairs in liver (HNF1A:PAX4, HNF1:SRY) and one TF pair in skeletal muscle (PAX:TBP) among predicted TF pairs by Hu and Gallo One reason for the small agreement may be the different predicted central regulators in studied tissues. The hubs identified by Hu and Gallo in liver are CEBP, HNF3, and HNF4 whereas the hubs found with our methodology in liver are HNF1A, HNF1B and HNF4A. The central hub in our predicted muscle network (MEF2A) does not occur in the muscle-specific network of interacting TF pairs from Hu and Gallo The agreement of predictions between Hu and Gallo and Yu et al. is very low too although they use a very similar set of PWMs.

We see several reasons why the agreement of our tissue-specific predictions and those from Yu et al. is much larger than the agreement in comparison with Hu and Gallo

First, we use the sets of tissue-specific genes derived by Yu et al. for our tissue-specific predictions too. Second, the predictions of Yu et al. are much more numerous (e.g. 1052 significant TF pairs in muscle compared to 121 significant TF pairs predicted by Hu and Gallo) such that the chance to find some common TF pairs is much higher. Third, whereas our method and the method of Yu et al. focus on the overrepresentation of pairs of transcription factor binding sites, Hu and Gallo uses the functional co-evolution and co-expression of the common target genes of TF pairs.

4.7 Conclusion

Tissue-specific gene expression is regulated by an interplay of multiple transcription factors (Remenyi et al., 2004). The identification of transcription factors which co-occur in the promoter regions and regulate together the expression of their target genes is very important to better understand how cells in different tissues and developmental states achieve their specificity. Previous computational studies were usually based on common sequence features of promoters corresponding to some tissue function or function related groups of genes (Klein and Vingron, 2007; Smith et al., 2007; Yu et al., 2006). Another approach was based on evolutionary conservation and co-expression of genes which are co-regulated by interacting TFs (Hu and Gallo, 2010). Although these studies make plausible predictions, the mechanisms which control gene expression are still not fully understood, nor are the exact relationships between various transcription factors in different stages.

In this chapter, we presented a novel method for predicting **co-occurring transcription factors in tissue-specific manner**. We represented each TF as ranked list of promoters ordered by the predicted binding affinity of the TF to the promoter sequence. To identify co-occurring TF pairs in a tissue-specific aspect, tissue-specificity information of the target genes was included. Then, we applied statistical testing in **3-way contingency tables** to detect significant TF pairs co-occurring in the studied tissue. Since four different null models for testing in 3-way contingency tables are available, we conducted goodness-of-fit tests for all null models using all TF pairs. The suitable model was selected based on the distribution of resulting p-values, which was closest to a uniform distribution with moderate enrichment of significant p-values. With this approach, we identified the **partial independence model** as the best underlying independence model fitting the data.

Then, we identified highly significant co-occurring TF pairs in 34 human tissues fulfilling the p-value threshold. In total, we found 1061 **significant TF pairs in 22 human**

tissues, corresponding altogether to 767 unique TF pairs. In our analysis, we focused on TF pairs between transcription factors with nonsimilar binding motifs. The majority (86.6%) of our significant TF pairs had nonsimilar motifs which reduces the possible source of false positive predictions only due to high similarity of binding motifs. The biological relevance of our discovered tissue-specific TF pairs was demonstrated by both known protein-protein interactions in validated databases and by the expression of TFs in the corresponding tissue. We have shown that known protein-protein interactions are enriched (1.8-fold) in the set of predicted TF pairs with tissue specification. However, the proportion of known PPIs in the groups of significant TF pairs varies in different tissues from 10-fold enrichment to no occurrence of known PPIs.

A large majority ($60-70\%$) of the predicted tissue-specific factors have experimental evidence to be **expressed in the corresponding tissue**. All factors in the tissue-specific TF networks were found just by the selection criterion from the statistical test in the contingency table, without any knowledge about their functions in the tissue of interest. Thus, these results indicate that our predicted tissue-specific TF pairs and thereby the tissue-specific regulatory networks are very likely functional in the corresponding tissues. Furthermore, we investigated significantly enriched gene functions related to the examined tissue which support the hypothesis of the regulatory function of these predicted factors in the tissue.

Our predicted networks consisting of significant tissue-specific TF pairs are characterized by one or two **central regulators** with a high number of respective partners. These central hubs are HNF1A, HNF1B and HNF4A in liver; MEF2A and TBP in skeletal muscle and NFYA and ELK1 in hematopoietic stem cells. All these factors have known regulatory function in the corresponding tissue and were experimental validated as *specifier* (e.g. tissue-specific central regulator) or *facilitator* (e.g. widely expressed central regulator) hubs by Ravasi et al. (2010). These findings demonstrate that both, non-specific and tissue-specific TFs play a large role in regulation of tissue-specific genes. Furthermore, individual TFs can contribute to tissue specificity in different tissues by interacting with distinct TF partners.

Despite the fact that we were able to successfully predict novel pairs of co-occurring TFs in various tissues, our method could be improved. Since our method is not able to distinguish between a cooperative binding of two TFs with highly similar motifs and their competing for one binding site, such TF pairs have to be excluded from our analysis. However in general, factors with very similar motifs can in reality jointly bind to the DNA sequence and regulate the transcription of the target gene (e.g. FOS and JUN, Gerstein et al. (2012); Glover and Harrison (1995)).

In this chapter, we use a simple definition of promoter regions as a fixed size region upstream from the TSS of a gene. We could achieve much higher accuracy of transcription factor binding affinity prediction by using the open chromatin regions in various tissues or cell lines (Boyle et al., 2008a). In that case, we would have an exact knowledge of the accessibility of the genomic sequences. Moreover, these regions are generally much smaller than our definition of promoter regions thus the rankings of these regions according to binding affinity might differ.

Further, for our predictions, we have used the groups of genes which are specifically expressed in the tissue of interest. This information was derived from the ESTs measurements. But many mammalian tissues are highly heterogeneous and consist of large number of different types of cells which might be regulated by different combinations of transcription factors. So, rather focusing on cell-type-specific genes or regulatory regions than on the whole tissue would improve the accuracy of predicted co-occurring TF pairs. One problem of the usage of the cell-type-specific groups of genes is that they might include smaller numbers of genes, and then the probability of having common cell-type specific genes at the top of the ranked lists (e.g. the n_{111} entry in the 3-way contingency table) will be even smaller than for tissues.

The two challenges mentioned above (e.g. using only accessible DNA-regions and usage of cell-type-specific information) are addressed and further studied in the following chapter. In addition, we hope that there will be more experimental data available (such as a positive set of co-occurring TFs and a negative set of not co-occurring TFs) in the near future which could provide a measure of the specificity and sensitivity of our predictions. Finally, our findings showed that comparing the usage of rank based statistics for transcription factor targets results in plausible predictions of co-occurring transcription factors in various human tissues.

5 Cell-type-specific transcription factor co-occurrence in genomic regulatory regions

5.1 Motivation

The prediction of tissue-specific cooperative binding of transcription factors on promoters was discussed in the previous chapter using the ranked list representation of TFs and applying the 3-way contingency table test. For this prediction, we assumed that the complete promoter regions are open and accessible for transcription factor binding and applied the affinity prediction method on the complete promoter. However, the *cis*-regulatory sequences, such as promoters and enhancers, are embedded in chromatin and are accessible in dependance of the temporal and spatial development of the cell. The presence or absence of nucleosomes, the basic repeating unit of the chromatin determines whether *cis*-regulatory elements are accessible for binding of transcription factors or not. Thus, **chromatin accessibility** is necessary for *cis*-regulatory elements to exert their regulatory effects.

The accessible regulatory regions can be identified using various experiments such as chromatin immunoprecipitation (ChIP) experiments or DNase I hypersensitivity experiments, see Section 1.3.2. Previous studies (Boyle et al., 2008a; John et al., 2011; Pique-Regi et al., 2011) showed that the integration of the TF binding prediction models and the DNase I hypersensitivity score considerably improve the prediction of putative TF binding to the DNA by decreasing the false positive rate.

In this chapter, we make use of a recent large study (The ENCODE Project consortium, 2012) where the open chromatin was assessed in more than 100 human cell types using the DNase I experiments combined with sequencing (DNase-seq). With this large scale data set, we can identify hundreds of thousands of **cell-type-specific open reg-**

ulatory regions and **ubiquitously open regulatory regions**. Then, we study the co-occurrence of various TFs in these cell-type-specific regions. Moreover, we are able to detect pairs of TFs which preferentially co-occur in the cell-type-specific manner when compared to their behavior in the ubiquitous regulatory regions. The large number of distinct genomic regulatory regions in all cell types motivated the development of a new method for detecting **cell-type-specific co-occurring TFs**. Namely, we compare the similarity of two ranked lists (which represent two TFs) within the cell type of interest against their similarity within the ubiquitously open regulatory regions. The significant cell-type-specific TF pairs are pairs with high similarity within the cell-type-specific regulatory regions and small similarity within the ubiquitously open regions.

In the past years, approaches predicting cooperation between TFs in the open regulatory regions were developed. One class combines the experimental data of TF binding such as ChIP-seq or ChIP-chip, for different factors to detect significantly co-binding TFs. Usually, the main idea is to compare number of peak occurrences of two TFs on shared locations to the number of single peak occurrences (Gerstein et al., 2012; Wang et al., 2012) or to integrate overrepresentation analysis of secondary motifs in peak regions bound by the primary TF (Oh et al., 2012; Wang et al., 2012; Whitington et al., 2011). These approaches usually give highly precise predictions but they are restricted by the availability of the experimental data. The largest available human study of the The ENCODE Project consortium (2012) has generated ChIP-seq data sets for 119 distinct transcription factors in five cell lines, and out of them only 87 have a DNA-binding domain with a sequence-specific binding motif (Wang et al., 2012). The number of ChIP-seq experiments in other cell lines is much smaller. It generally includes only few TFs.

The second group of methods for predicting TF co-occurrence uses the experimental evidence of open chromatin derived from the DNase I hypersensitive experiments to find significantly enriched pairs of TFs. Jankowski et al. (2013, 2014); Kazemian et al. (2013) focus on the prediction of direct TF-TF dimerization with fixed spacing and orientation. Neph et al. (2012) investigated the occupancy of binding motifs in DNase I footprints which provides precise information of DNA-protein binding due to a nonuniform DNase I cleavage. Then, they focus on the tethered binding of an indirect DNA interaction of one TF through an interaction with another TF.

The third type of methods predicting TF cooperation (Park et al., 2014; Vandenbon et al., 2012) is based on integration of gene expression measurements where the regulatory regions of co-expressed genes are investigated for TF motif overrepresentation. These approaches have the advantage of the evidence of the functional effect on the differentially expressed genes by the combination of TFs on the promoter regions. However, they are limited by analysis of promoter sequences only or by a small number of known enhancer-target pairs, since the general targets of distant regulatory regions are not known. Further, they are also limited by the availability of experimental data.

Thus, to our knowledge, our approach is a novel method to detect cell-type-specific co-occurrence of TFs with fuzzy spacing using arbitrary sets of binding motifs of interest and experimental evidence of open chromatin region in the studied cell type. The advantage of our method lies in the **TF representation** as a **ranked list of the regulatory regions** which does not require setting of thresholds for binding motif hits and their spacing (see Section 3.2), or for the identification of the accessible genomic regions. Our rank based method requires only one parameter to define the top-ranked regions in the ranked lists, which could be chosen based on a biological motivation or the results can be aggregated over different choices of these parameters.

In this chapter, we first present the method for the determination of cell-type-specific chromatin accessible regions with the DNase-seq data and discuss shortly the properties of these cell-type-specific regions (Section 5.2). In Section 5.3, the representation of TFs as a ranked list of genomic regions in a cell-type-specific manner is introduced. Section 5.4 presents single TF motifs which are overrepresented on the cell-type-specific regulatory regions and examines in depth the TF motifs in immune-related cells, lung cells, embryonic stem cells and in muscle cells. In Section 5.5, the method for predicting co-occurring TFs on cell-type-specific regulatory regions is introduced. Further, the results in immune-related cells, embryonic stem cells and in muscle cells are discussed. The predicted TF pairs are validated with other computational and experimental based approaches in Section 5.6. The last section provides a short summary of this chapter.

5.2 Cell-type-specific chromatin accessibility

The usual approach to measure chromatin accessibility genome-wide is to digest chromatin with the endonuclease DNase I followed by sequencing (DNase-seq). The accessible chromatin regions are preferentially cleaved by the endonucleases DNase I, therefore they are referred to as **DNase hypesensitive sites (DHS)**. The DNase-seq experiment generates a genome-wide map of the accessible chromatin (Boyle et al., 2008a); the more sequenced reads map to a certain region, the more the region is hypersensitive to DNase I digestion and thus more accessible.

DNase I hypersensitivity, as measured by DNase-seq, has been used previously to characterize cell-type-specific promoters and enhancers (Ernst et al., 2011; Song et al., 2011;

Xi et al., 2007). However, such analyses focused either on immortalized cell lines or cell lines derived from cancer cells. For our purpose, we want to use the DNase-seq experiments in various healthy cell types to determine genomic regions with high degrees of chromatin accessibility which are most specific for certain cell types. Moreover, we use a statistical test to create ranked lists of accessible genomic regions ordered by their specificity for each cell type. To do so, we used data from The ENCODE Project consortium across 88 healthy and 2 cancer cell lines. This large number of experimental data allowed us to derive a whole set of cell-type-specific regulatory regions for well-studied human cell types such as white blood cells, fibroblasts, myoblasts, epithelial cells, endothelial cells and others. The detailed description of the experimental data used, with cell line and corresponding tissue information is listed in Appendix, Table A.1. Some of the cell lines which are biologically highly similar were grouped into one cell type, resulting in a total number of 64 cell types (see Section 5.2.1 for detailed information).

In the selection of the experiments we focused mainly on healthy cell lines since we want to study the natural epigenetic landscape in different cell types rather than to study particular cell lines (e.g. immortalized or cancer cell lines) where the differences in chromatin accessibility might be a consequence of being a cancer or being immortalized. For this reason we considered all DNase-seq experiments in healthy cell lines available from the ENCODE consortium which were conducted in the same center (University of Washington) to avoid high technical variability. In addition, we included two cancer cell lines (K562 and HeLa-S3) because of a large number of experimental studies analyzing these two cell lines for later comparison of our results.

5.2.1 Clustering of DNase hypersensitive sites

To investigate the reproducibility of the DNase-seq experiments in biologically related cell lines first, a large matrix of read counts over genomic regions across the whole human genome and over all cell types was created.

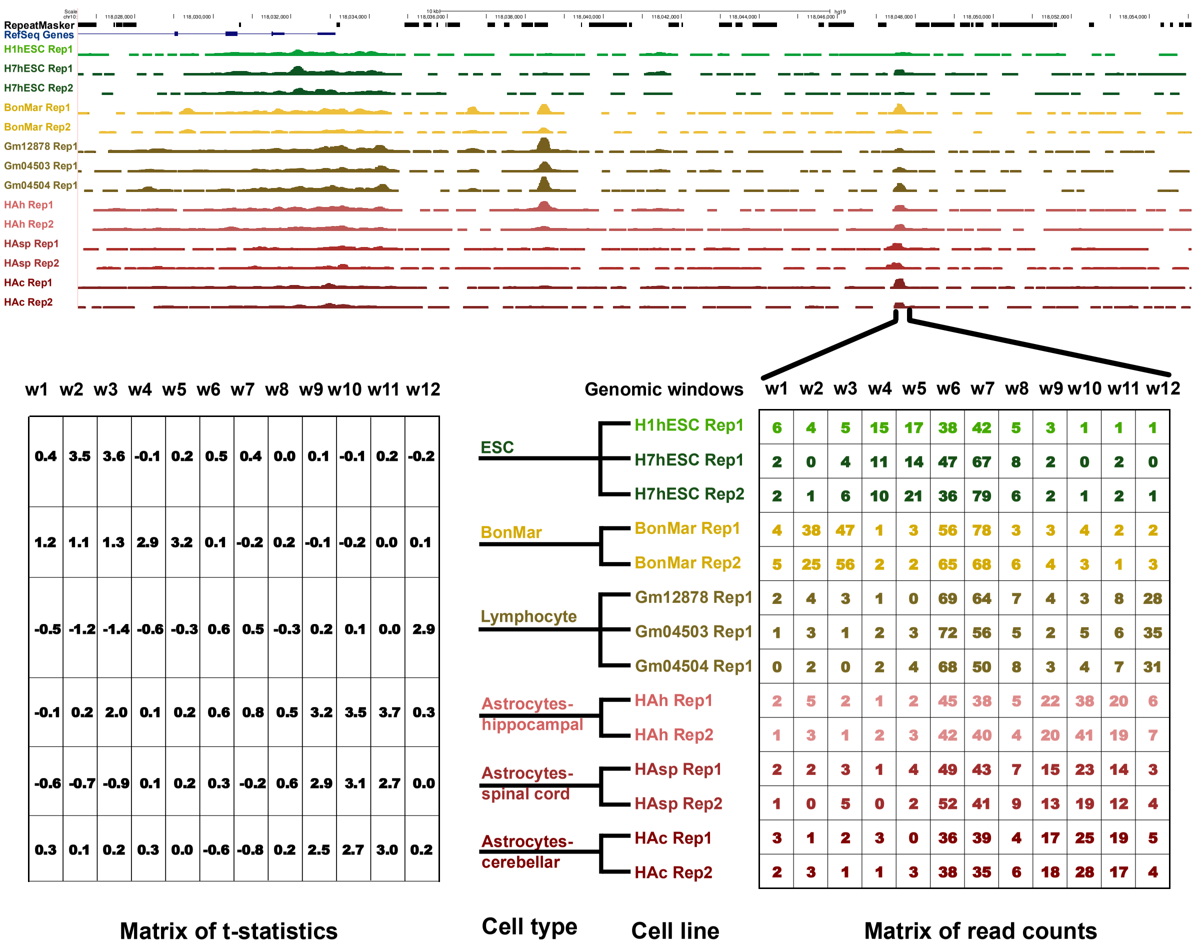
Namely, the human genome (hg19 Ensembl assembly from genome.ucsc.edu) was divided into 200 bp long, non-overlapping windows. Windows which overlay repetitive elements in RepeatMasker (Smit et al., 1996-2010) with scores higher than 1000 were eliminated resulting in total of 9.7 million windows. The DNase-seq reads from a total of 164 experiments in 90 cell lines were counted, and counts were normalized for sequencing depth by multiplying each sample by the average read count over all samples divided by the sample's average read count. Then the decadic logarithm of the normalized counts with a pseudocount of one read was taken. An illustrative example for 14 data files from 7 different cell lines (highlighted with different colors) is shown in Figure

5.1. The upper figure shows the DNase-seq raw data files, the bottom right part shows the corresponding matrix with normalized read counts of 12 genomic windows.

To investigate the similarity of the read counts of biologically related cell types, we calculated Pearson's correlation of the read counts in all genomic windows over all samples. Afterwards, we used complete linkage clustering to find clusters. The correlation matrix together with the dendrogram is visualized in Figure 5.2, the rows and columns of the symmetric matrix correspond to 164 samples used for our analysis. The color of each heatmap cell indicates the strength of the correlation, the darker the color, the higher the correlation between samples. Further, we depicted the corresponding tissues and types of cells where the studied samples came from. The tissue is displayed as a horizontal color-barcode above the heatmap; the type of cell is shown as a vertical color-barcode on the left side of the heatmap.

We could observe that the majority of the biological replicates from the same cell line have a high correlation and cluster together. Further, functionally related cell lines, such as *renal cortical epithelial cells* and *renal epithelial cells*, usually build small, highly correlated clusters (see highlighted samples in green in Figure 5.2). Such highly correlated cell lines (which are biologically similar) were manually grouped into cell types, reducing the number of groups from 90 to 64. The exact grouping of all cell lines into cell types can be taken from the Appendix, Table A.1.

In addition, a very distinct clustering of some cell types or tissues can be identified in the heatmap and in the corresponding dendrogram. The most striking groups are highlighted in blue on the left dendrogram in Figure 5.2. For example, white blood cell samples, T-cells and monocytes create a large distinct cluster as well as the skin and gum fibroblasts and microvascular endothelial cells. Further, renal epithelial cells, B-cells, embryonic stem cells (ESCs), skeletal myoblasts and others form small distinct clusters. The large highly correlated cluster in the left lower corner consists of various cell types such as: brain astrocytes, fibroblasts, myoblasts and endothelial cells originated from different tissues. However, some biologically unrelated cell lines build relatively distinct clusters too. For example leukemia samples cluster together with retinal epithelial cell and lymphatic microvascular endothelial cell samples; and cervical carcinoma samples cluster with myoblasts and epithelial cell samples, see highlighted groups in violet on the left dendrogram in Figure 5.2. One possible reason for this clustering of unrelated cell lines could be general high correlation level of the read counts over all samples.



w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12
0.4	3.5	3.6	-0.1	0.2	0.5	0.4	0.0	0.1	-0.1	0.2	-0.2
1.2	1.1	1.3	2.9	3.2	0.1	-0.2	0.2	-0.1	-0.2	0.0	0.1
-0.5	-1.2	-1.4	-0.6	-0.3	0.6	0.5	-0.3	0.2	0.1	0.0	2.9
-0.1	0.2	2.0	0.1	0.2	0.6	0.8	0.5	3.2	3.5	3.7	0.3
-0.6	-0.7	-0.9	0.1	0.2	0.3	-0.2	0.6	2.9	3.1	2.7	0.0
0.3	0.1	0.2	0.3	0.0	-0.6	-0.8	0.2	2.5	2.7	3.0	0.2

Cell type	Cell line	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12
ESC	H1hESC Rep1	6	4	5	15	17	38	42	5	3	1	1	1
	H7hESC Rep1	2	0	4	11	14	47	67	8	2	0	2	0
	H7hESC Rep2	2	1	6	10	21	36	79	6	2	1	2	1
BonMar	BonMar Rep1	4	38	47	1	3	56	78	3	3	4	2	2
	BonMar Rep2	5	25	56	2	2	65	68	6	4	3	1	3
Lymphocyte	Gm12878 Rep1	2	4	3	1	0	69	64	7	4	3	8	28
	Gm04503 Rep1	1	3	1	2	3	72	56	5	2	5	6	35
	Gm04504 Rep1	0	2	0	2	4	68	50	8	3	4	7	31
Astrocytes-hippocampal	HAh Rep1	2	5	2	1	2	45	38	5	22	38	20	6
	HAh Rep2	1	3	1	2	3	42	40	4	20	41	19	7
Astrocytes-spinal cord	HAsp Rep1	2	2	3	1	4	49	43	7	15	23	14	3
	HAsp Rep2	1	0	5	0	2	52	41	9	13	19	12	4
Astrocytes-cerebellar	HAc Rep1	3	1	2	3	0	36	39	4	17	25	19	5
	HAc Rep2	2	3	1	1	3	38	35	6	18	28	17	4

Figure 5.1: Overview of the method for the determining of cell-type-specific DNase hypersensitive sites (CTS-DHSs). The top figure shows the raw DNase-seq data for 14 samples of 7 different cell lines (highlighted with different colors). Then, for each genomic window, the normalized read counts for each sample are calculated (matrix of read counts in the right bottom part). The t-statistics is then calculated for each genomic window over all cell types corresponding to the cell lines (matrix of t-statistics in the left bottom part).

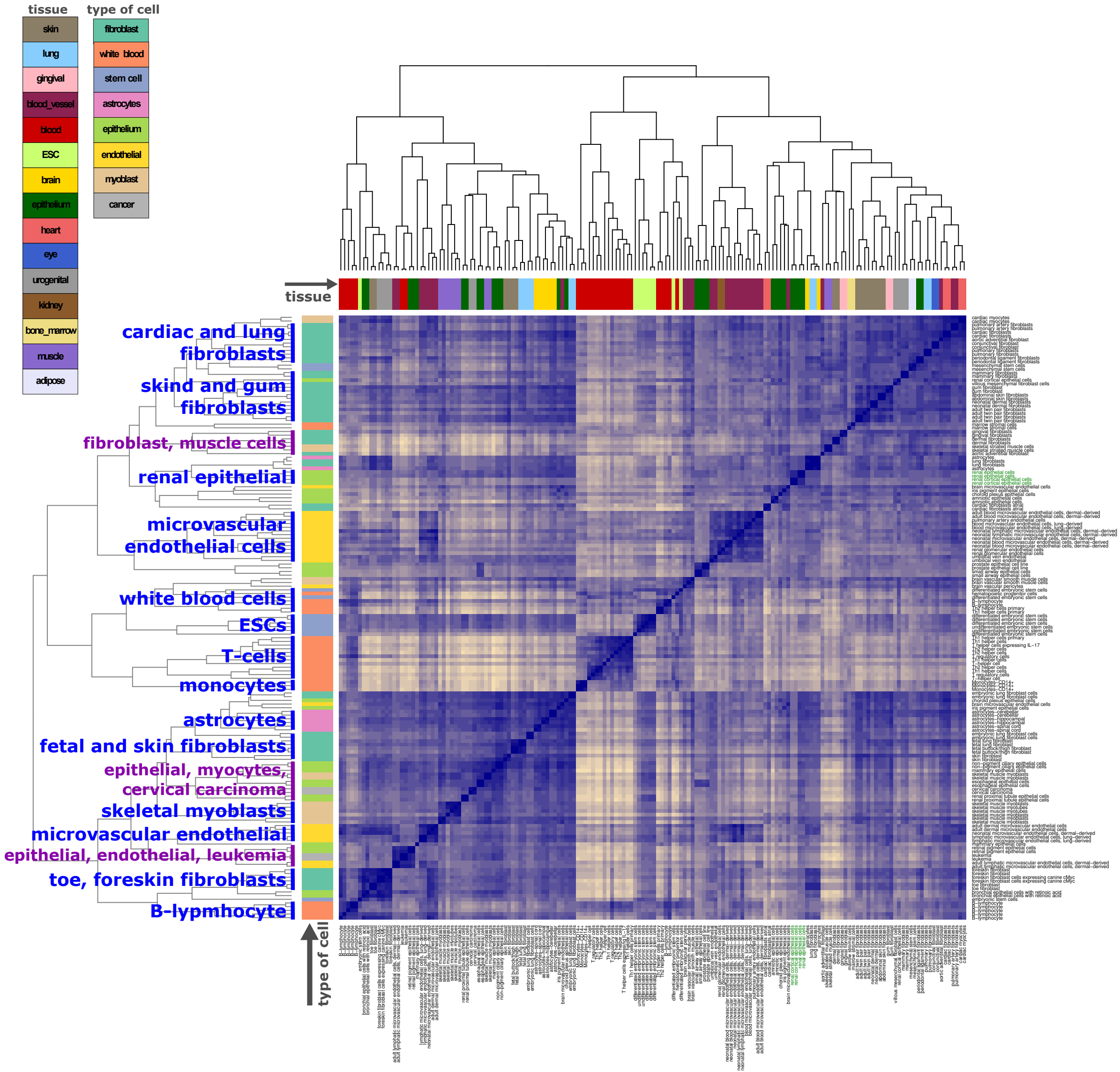


Figure 5.2: Correlation matrix among all DNase-seq samples calculated from the read coverage in 9.7 million genomic windows. The horizontal barcode shows the corresponding tissue of the sample, the vertical barcode the corresponding cell type of the sample. The complete linkage method was used for clustering.

5.2.2 Ranking the DNase hypersensitive sites by cell-type specificity

The cell-type specificity of the genomic windows is quantified with a t-statistic taking into account within-tissue variation of the DNase hypersensitivity. For the calculation, the matrix of normalized read counts described in Section 5.2.1 and the grouping of cell lines into cell types as listed in Appendix Table A.1 was used. This method for quantification of cell-type specificity was adapted from Love and Chung (2012).

Formally, the t-statistics is derived for each genomic window separately, as described in the following. For a given window w, $w = 1, \dots, W$, let C_{wi} denote the normalized log read count for sample i which belongs to cell type $t(i) = l$. Let us denote the total number of cell types with m and each cell type l, $l = 1 \dots, m$ with n_l number of samples (e.g. replicates). Then the cell type average count for each window w over all samples belonging to cell type l is:

$$\overline{C}_{wl} = \frac{1}{n_l} \sum_{i:t(i)=l} C_{wi}. \tag{5.1}$$

Assuming equal variance among cell types, the pooled within-cell-type standard deviation for each window w can be calculated as:

$$s_G^w = \sqrt{\frac{\sum_{l=1}^{m} \sum_{i:t(i)=l} (C_{wi} - \overline{C}_{wl})^2}{\sum_{l=1}^{m} (n_l - 1)}}. \tag{5.2}$$

We can also define the global average count for each window w over all cell types as:

$$\overline{C}_{wG} = \frac{1}{m} \sum_{l=1}^{m} \overline{C}_{wl}. \tag{5.3}$$

Then, the Student's t-statistic for cell type l and each window w can be calculated to quantify the divergence of the cell type average from the global mean:

$$t_{wl} = \frac{\overline{C}_{wl} - \overline{C}_{wG}}{\sqrt{\frac{1}{m} + \frac{1}{n_l}} (s_{wG} + s_0)}, \tag{5.4}$$

where s_0 is the mean over all windows: $s_0 = \frac{1}{W} \sum_{w=1}^{W} s_{wG}$. s_0 can be understood as value to moderate the t-statistic, by preventing division by very small within-cell-type estimates of standard deviation. Thus the t-statistic provides a measure of the cell-type specificity of DHS for the corresponding cell type. An illustrative example for 14 data files grouped in 6 different cell types is shown in Figure 5.1. The matrix of t-statistics (bottom left figure) for 12 genomic windows over all 6 cell types is calculated from the matrix of read counts (bottom right figure).

Genomic windows with the largest positive t-statistic are cell-type-specific DNase hypersensitive regions, we call the top-ranked windows as *cell-type-specific DNase hypersensitive sites (CTS-DHSs)*. In contrast, globally open genomic windows have cell type average counts $\overline{C}_{wl}$ close to the global means $\overline{C}_{wG}$ in all cell types, thus the global t-statistic over all cell types $t_G^w := \frac{1}{m}\sum_{l=1}^{m} t_l^w$ is close to zero. Then, we define sites with the smallest global t-statistic t_G^w as *ubiquitous* DHSs, e.g. sites which are globally open in all cell types. We will use these sites as control sequences for the comparison of TF co-occurrence in the cell-type-specific regions to the TF co-occurrence in the ubiquitous regions.

5.2.3 Genomic location of CTS-DHSs

Next, we investigate the genomic location of the identified CTS-DHS and the identified ubiquitous DHSs. We selected 5000 most cell-type-specific sites in each cell type and 5000 most ubiquitous sites, its distribution along the genome is shown in Figure 5.3. The large majority (88%) of the top-ranked CTS-DHSs are located in intronic and intergenic regions. Only 8% of the top-ranked CTS-DHSs are situated in promoters, and a very small part ($< 4\%$) overlay with annotated exons (hg19 Ensembl assembly from genome.ucsc.edu). The only exception is the primary T-cell which has 19% of sites in exons and 22% of sites in promoters. The genomic distribution of the top CTS-DHSs is in strong contrast to the genomic distribution of the top-ranked ubiquitous DHSs, of which 43% overlap promoter regions (see the top bar in Figure 5.3). These findings suggest that the CTS-DHS are mainly cell-type-specific enhancers, a conclusion that has also been drawn in earlier studies about specific cell lines (Ernst et al., 2011; Song et al., 2011; Xi et al., 2007).

5.3 Transcription factor as a ranked list of DNase hypersensitive sites

For our goal of predicting overrepresented TFs and co-occurring TFs on the DHSs we require a representation of a TF as a ranked list of (cell-type-specific) DNase hypersensitive sites. In contrast to our definition of TF as a ranked list of its target genes in Section 3.2, there is no fixed (predefined) number of DHSs which should be taken to create a ranked list. When analyzing TF co-occurrence on human promoters, the complete list of all human promoters (genes) from the Ensembl database was used. This was a biologically justified number of items, although different gene databases like RefSeq or

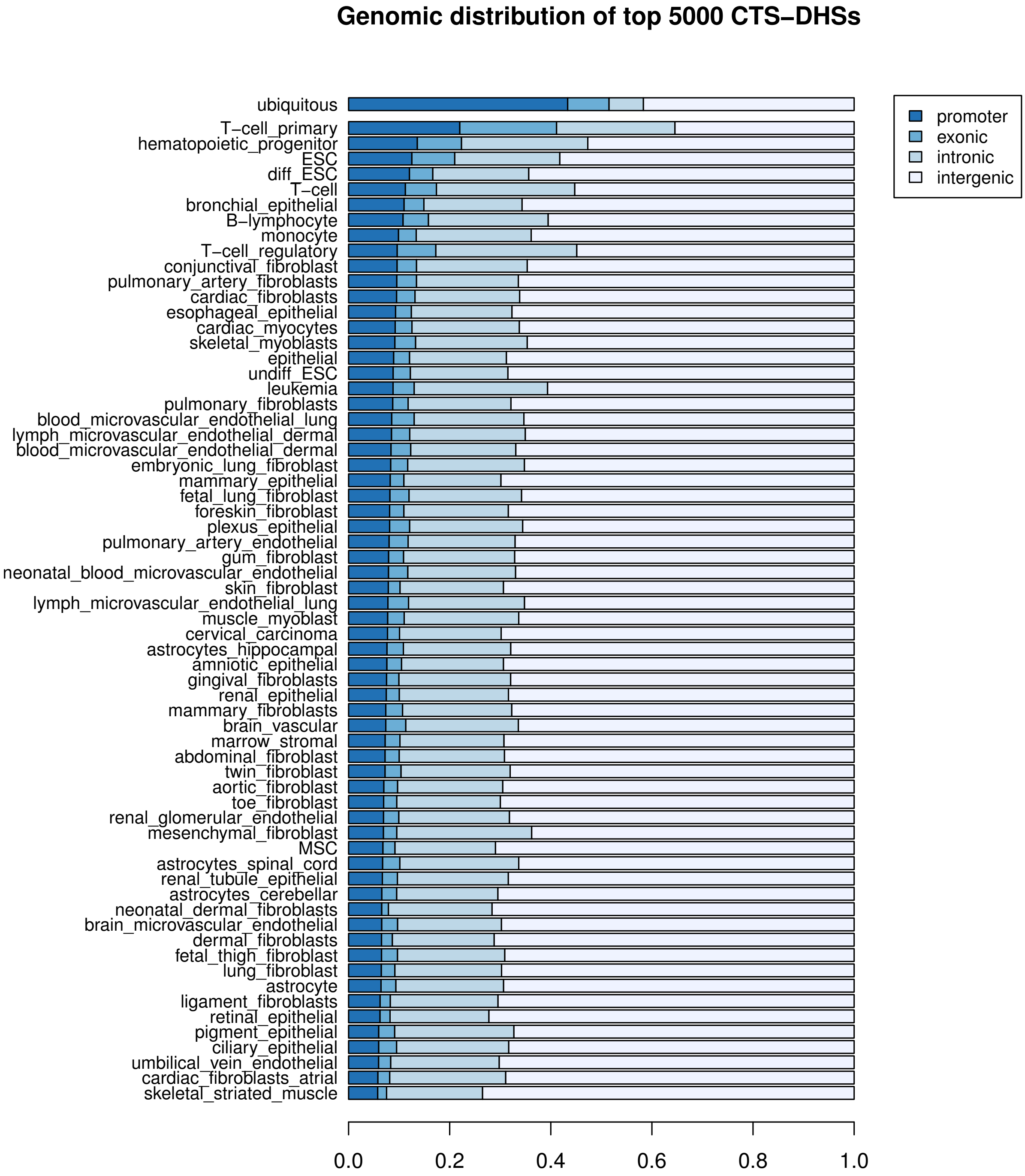


Figure 5.3: Genomic distribution of the 5000 most cell-type-specific DNase hypersensitive sites in 64 cell types and of the top 5000 ubiquitous DNase hypersensitive sites sorted by the overlap with promoter regions.

Entrez might include slightly different number of genes.

However, there is no such natural information about how many sites of open chromatin are present in the genome. This number might vary from cell to cell or cell line to cell line. For example, Song et al. (2011) report 100 000 – 140 000 DHSs in 7 different cell lines. Many of them are open over all cell lines, some of them are cell-line-specific. Thus, with the DNase-seq experiments, we are able to find hundreds of thousands of regions with enriched read counts without any knowledge whether these regions are functional or not. For this reason, representation of a TF as a ranked list of *all* cell-type-specific DHSs is not feasible.

With the method described in the previous Section 5.2, we can detect up to several thousands of cell-type specific DHSs which are unique for the corresponding cell type. A full list of all CTS-DHSs from all available cell types would result in a ranked list with more than 600 000 items for each TF. When looking for co-occurring TFs, the focus is on the very top of the list, usually up to several thousands of items (e.g. sites). In this case, a comparison of three ranked lists would lead to a large number of highly significant results even for a very small number of shared top-ranked DHSs.

In general, increasing the universe of the contingency table (e.g. the number of all ranked items in the list) can dramatically change the significance of the results. Let us consider an example of two different 3-way contingency tables for top-ranked cell-type-specific DHSs with different length of the lists as shown in Table 5.1. The threshold for top-ranked sites is set in both lists to $k_1 = k_2 = 2000$ and there are total of 5000 cell-type-specific DHSs in both tables. Further, both TFs share the same number of cell-type specific and non-specific sites, but the total length of both lists (universe of the contingency table) in the first table is 200 000 sites and the total length of both lists in the second table is 10 000 sites. Hence, p-value of the partial independent test (see Section 4.2.2) corresponding to the first table equals 0, whereas the p-value corresponding to the second table is 0.88. This simple example demonstrates how drastically the significance of p-values can be influenced by the length of the lists.

Therefore, to avoid an artificially large contingency table, we construct the ranked lists of the DHSs for each TF in a cell-type specific manner including the ubiquitous DHSs as a sort of contrast. First, in analogy to Section 3.2 we estimate the binding preferences with TRAP (Roider et al., 2007) of all TFs of interest to the t most cell-type-specific DHSs and to the t most ubiquitous DHSs. Then we construct for each TF a ranked list R of length $2t$ of the t most cell-type-specific DHSs and of t most ubiquitous DHSs, separately for each cell type. Whereas for given cell type, the CTS-DHSs change due to their specificity, the set of ubiquitous DHSs remains the same for all studied cell types.

Then we order the cell-type-specific DHSs and ubiquitous DHSs jointly by the binding affinity, separately for each TF. An example of TFs represented by a ranked list of DHSs is shown in Figure 5.4. Here, the $t = 10$ most *heart*-specific DHSs (in red) and $t = 10$ most *ubiquitous* DHSs (in grey) are ordered for a particular TF by its binding affinity. Then, the same procedure is repeated for *lung*, *brain* and all other cell types.
For our analysis a list of 477 known TF motifs obtained from TRANSFAC 2012 database from BIOBASE Corporation (Matys et al., 2006, www.biobase-international.com) was used. We chose the TRANSFAC database for the analysis on DHSs because it includes motifs for more TFs than the JAPSAR database, which was used in the previous Chapters 3 and 4. Further, TRANSFAC database includes motifs for the majority of experimentally studied TFs in the ENCODE project (Wang et al., 2012) which can be used for later comparison of our results. However, TRANSFAC database contains redundant entries, i.e. multiple motifs corresponding to a single TF, since transcription factors are known to recognize more than one consensus sequence (Badis et al., 2009). In turn, similar DNA sequences can be recognized by different TFs (Ehret et al., 2001), thus different TFs might have same motifs in the database. Therefore, we manually annotated the 477 TF motifs to 262 single TFs or TF groups/families, using the information provided by the TRANSFAC database and by Oh et al. (2012), see Appendix, Table A.2. In the following, we refer to TF motifs when analyzing the 477 binding motifs and to simple TFs or factors when discussing the matched results to the set of 262 TFs or TF groups.

5.4 Overrepresented TF motifs in cell-type specific DHSs

Next, cell-type-specific DNase hypersensitive sites are studied for overrepresented transcription factor motifs. We want to find the transcription factors which are responsible for the cell-type specific gene regulation in the CTS-DHSs. To do so, Fisher's exact test described in Section 2.3.3 is applied to determine the significance of the overrepresented TFs.
First, for each TF motif and for given cell type, we create ranked lists according to TF binding affinity of the cell-type-specific DHSs and the ubiquitous DHSs, as described in previous Section 5.3. Then, for each TF motif a 2-way contingency table is constructed in such way that the row variable partitions the ranked list into the top-k-ranked DHSs according to TF affinity. The column variable identifies the cell-type-specific and ubiquitous DHSs. Formally, we define a binary variable X_k indicating DHSs ranked according

Table 5.1: Two 3-way contingency tables for shared cell-type specific sites among the top-k_1 and top-k_2 ranked target sites of two different TFs universe size (a) $n_1 = 200\ 000$ and (b) $n_2 = 10\ 000$ (b).

(a)

	cell-type-specific		not specific		
	$\lvert i\rvert : r_y(i) \leq 2000$	$\lvert i\rvert : r_y(i) > 2000$	$\lvert i\rvert : r_y(i) \leq 2000$	$\lvert i\rvert : r_y(i) > 2000$	$\sum$
$\lvert i\rvert : r_x(i) \leq 2000$	209	791	202	798	2 000
$\lvert i\rvert : r_x(i) > 2000$	791	3 209	798	193 202	198 000
$\sum$	1 000	4 000	1 000	194 000	**200 000**

Partial independence test p-value $= 0$.

(b)

$\lvert i\rvert : r_x(i) \leq 2000$	209	791	202	798	2 000
$\lvert i\rvert : r_x(i) > 2000$	791	3 209	798	3 202	8 000
$\sum$	1 000	4 000	1 000	4 000	**10 000**

Partial independence test p-value $= 0.88$.

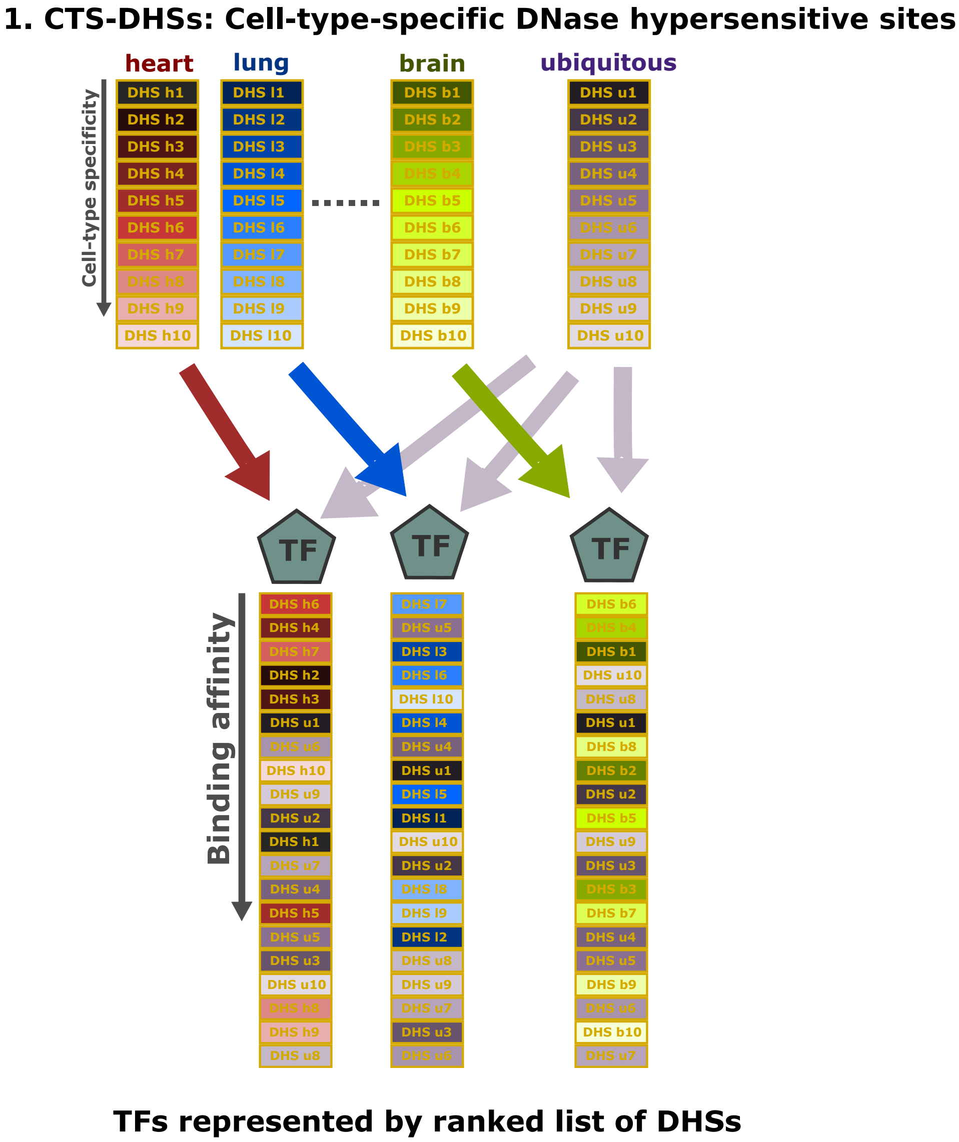


Figure 5.4: Transcription factor represented as a ranked list of CTS-DHSs and ubiquitous DHSs ordered by the binding affinity. For each cell type separately, the CTS-DHSs and ubiquitous DHSs are jointly order by the binding affinity of each transcription factor.

to TF affinity among the top-k as:

$$X_k(i) = \begin{cases} 1 & \text{if } r_x(i) \leq k, \;\; k \in \{1, \ldots, 2t\} \\ 0 & \text{otherwise,} \end{cases} \tag{5.5}$$

for each DHSs $i, \;\; i = 1, \ldots, 2t$. Further, the binary variable Z_l indicates specific DHSs of cell type l and is defined for each DHS $i, \;\; i = 1, \ldots, 2t$ as:

$$Z_l(i) = \begin{cases} 1 & \text{DHS } i \text{ is cell-type-specific for cell type } l \\ 0 & \text{DHS } i \text{ is ubiquitous .} \end{cases} \tag{5.6}$$

The significance of the top-ranked cell-type-specific DHSs ($n_{11} := \sum_i \mathbb{1}(X_k(i) = 1, Z_l(i) = 1)$) is assessed with the Fisher's exact test.

A toy example for 10 most heart-specific and 10 most ubiquitous DHSs with $k = 7$ top-ranked DHSs and the corresponding contingency table is shown in Figure 5.7 in module 3.

Analogous to the prediction of co-occurring TFs on general promoters discussed in Section 2.3.3 a choice of the cutoff k defining the top-ranked DHSs has to be made. Moreover, since we are able to derive hundreds of thousands of cell-type-specific DHSs for each cell type, another cutoff t defining the most cell-type specific and most ubiquitous DHSs is necessary.

To test the consistency of the enriched transcription factors in the CTS-DHSs we calculated Fisher's exact test for all possible combinations of thresholds $k \in \{500, 1000, 2000\}$ and $t \in \{1000, 2000, 5000, 7000\}$. The dependency of the log_{10} p-value on the choice of different thresholds for all 477 TF motifs in 3 different cell types is shown in Figure 5.5. In general, we can observe a decline of the significance with decreasing number of selected CTS-DHSs (t) and with decreasing number of selected top-scored DHSs (k). However, this trend is relatively weak in the majority of studied cell types as shown for leukemia in Figure 5.5a and B-lymphocyte in Figure 5.5c.

In addition, the top-20 enriched motifs selected with $k = 500$ and $t = 5000$ were highlighted in red and the selection of the top-20 enriched motifs with $k = 1000$ and $t = 1000$ was highlighted in blue. The first combination of thresholds $k = 500$; $t = 5000$ represents one extreme case when k is relatively small and t large. The second combination of the parameters with $k = 1000$ and $t = 1000$ corresponds to the case when only the most cell-type-specific DHSs are selected and the top-ranked DHSs account for half of the list (since the length of the list is $2t$).

For many cell types, the agreement of the top-20 enriched motifs between these two extremes is very large and is consistent over the majority of the threshold combinations, as depicted in the case of leukemia (see Figure 5.5a). However, there are several cell types (mainly the embryonic stem cells, see Figure 5.5b) where the agreement between the two selections of thresholds is very small. The top-ranked motifs based on $k = 1000$ and $t = 1000$ have lower significance for other combinations of thresholds with $t > 1000$ and the top-ranked motifs based on $k = 500$ and $t = 5000$ have smaller significance for $t \leq 2000$. For some of the cell types one can observe a sort of mixture of the two above cases; relative consistency over all threshold combinations with few outliers which show low significance in one of the extreme combination of thresholds, as shown in the case of B-lymphocyte in Figure 5.5c.

After evaluation of all cell types, the combination $k = 500$ and $t = 5000$ was selected for the further analysis of the overrepresented TF motifs. First, the biological interpretation of this choice is reasonable, 5000 of the most cell-type-specific accessible regions in the genome are analyzed and our focus is on the very top of the ranked lists. Second, this combination of thresholds does not overestimate the significance as for other combinations (e.g. $k \in \{1000, 2000\}$ and $t \in \{5000, 7000\}$, respectively) where many of the TF motifs reach the minimal possible p-value of 10^{-150}. With these selected cutoffs of $k = 500$ and $t = 5000$ the significance for all contingency tables corresponding to all possible combinations of TFs and cell types was calculated.

The regulators of interest are those TFs with the greatest significance (smallest p-values) in each cell type. First of all we identified several general enriched factors which were ranked among the top 50 significant TFs in at least 40 out of 64 cell types. These general factors are: ARNT (Aryl hydrocarbon receptor nuclear translocator) with HIF1A (Hypoxia Inducible Factor 1), ETFA (Electron-Transfer-Flavoprotein, Alpha Polypeptide), TEAD2 (TEA Domain Family Member 2), GABP (GA Binding Protein Transcription Factor), KLF4 (Kruppel-Like Factor 4), HIC1 (Hypermethylated In Cancer 1), MYC (V-Myc Avian Myelocytomatosis Viral Oncogene Homolog) with MAX (MYC Associated Factor X), NFY (Nuclear Transcription Factor Y), NKX6-2 (NK6 Homeobox 2), NRF1 (Nuclear Respiratory Factor 1), POU2F1 (OCT1, POU Class 2 Homeobox 1), SP1 (Sp1 Transcription Factor), TFAP2A (Activating Enhancer-Binding Protein 2-Alpha), ZFP161 (Zinc Finger And BTB Domain) and E2F(Retinoblastoma-Associated Proteins) family, EGR (Early Growth Response) family and ETS Oncogene Family.

Most of these factors are known regulators of many genes and are involved in general cellular function such as: apoptosis, energy metabolism or cellular growth (HIF1, NRF1, SP1), cell cycle (E2F, MYC, MAX) or in general development of organs (TFAP2A, EGR family, KLF4, HIC1, TEAD2). Some of these factors control mitochondrial functions

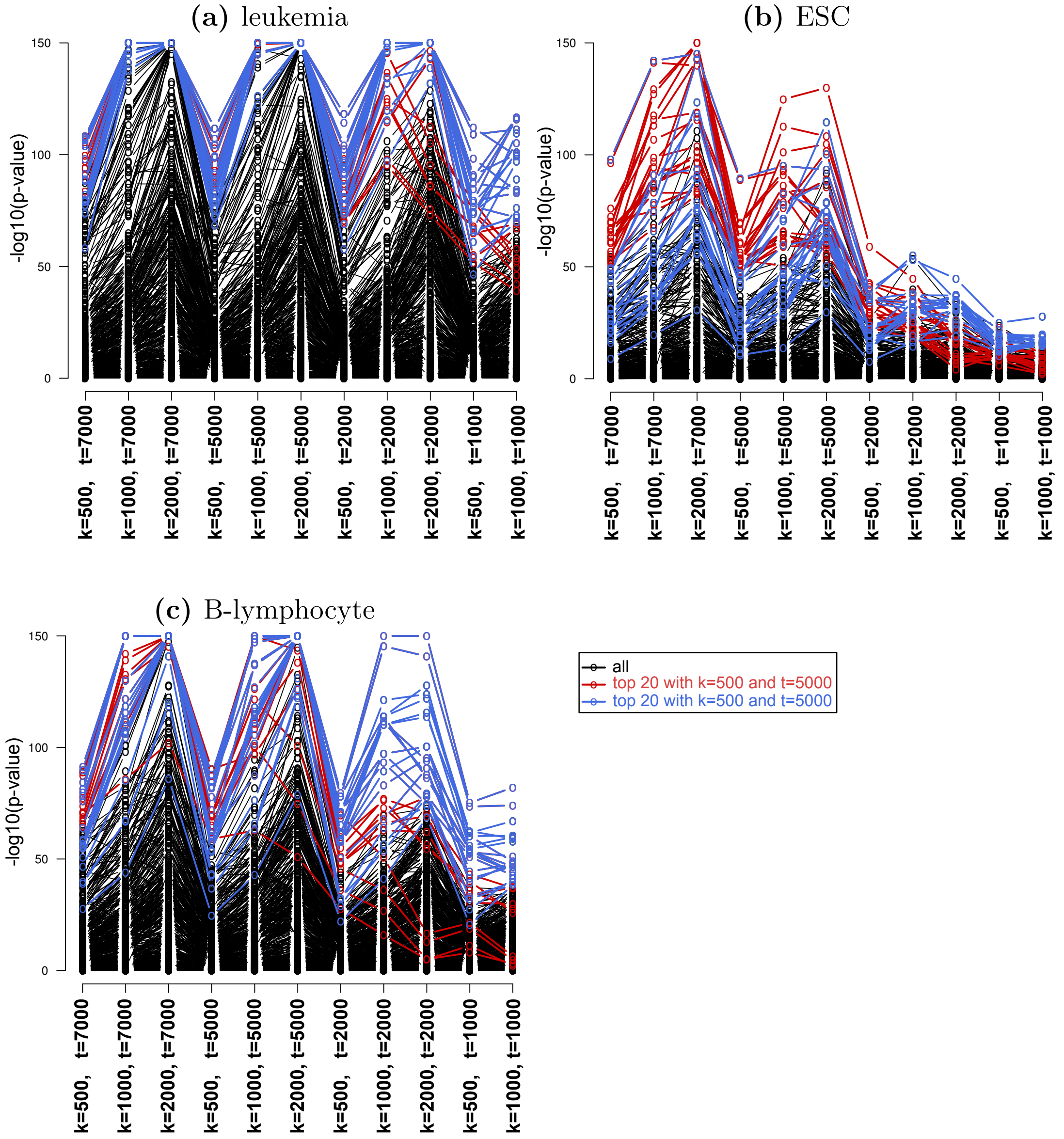


Figure 5.5: Dependency of the significance in Fisher's exact test on threshold selection for 477 TF motifs in (a) leukemia, (b) embryonic stem cells (ESCs) and (c) B-lymphocytes. The significance is represented as $-\log_{10}$ p-value on the vertical axis. 11 combinations of thresholds k (defining the top-ranked DHSs) and t (defining the number of cell-type-specific and ubiquitous DHSs) is depicted on the horizontal axis. Top-20 enriched TF motifs selected with two extreme values of k and t are highlighted in red and blue, respectively.

(ETFA, GABPA) or the biological rhythm (EGR). Three of these factors are tumor supressors (E2F,TEAD2, HIC1). The search of gene and protein functions was carried out with the Entrez Gene database (Maglott et al., 2011, www.ncbi.nlm.nih.gov/gene) and UniProtKnowledgebase (UniProt Consortium, 2011, www.uniprot.org). Although the general TFs are not cell-type-specific regulators, the overrepresentation of their motifs on the CTS-DHSs is reasonable. Most of these factors are very important transcriptional regulators so the high frequency of their motifs on distal regulatory elements is plausible.

To easily visualize the overrepresented TFs in a cell-type-specific manner, we created a color heatmap with significance for each TF in each cell type. For simplification, the general TFs mentioned above were removed from the matrix as well as TFs which did not show high significance ($-\log_{10}(p\text{-value}) < 20$) in any of the cell types. In case of multiple motifs corresponding to one TF the maximum significance (the minimal p-value) over the multiple motifs was taken. The heatmap of 192 resulting TFs (in rows) in all 64 studied cell types (in columns) is shown in Figure 5.6, the darkness of the cells indicates the significance of the association between the corresponding cell type and the occurrence of the corresponding TF.

First, we can see a large block of TFs enriched in a group of cell types in the left upper corner, namely there are TFs of the FOX (forkhead) family, POU family, NK homeoboxes, CEBP (CCAAT/Enhancer Binding-) proteins, EGR family, CREB (CAMP Responsive Element Binding-) proteins, GATA (GATA Binding-) proteins and AHR (Aryl Hydrocarbon Receptor) - ARNT complex (AHR nuclear translocator), all enriched in various fibroblasts (mesenchymal, neonatal dermal, pulmonary, gingival, gum, cardiac), some endothelial cell lines (renal glomerular and umbilical vein), in astrocytes of spinal cord, in pigment epithelial cells and in cervical carcinoma. Some of these factors such as EGR family members 1 and 4, NKX3-1, FOXN1, CREB, AHR, together with EVI1, HIF1A, LHX3, OCT4 and ETS1 (highlighted in red in Figure 5.6) are enriched in the majority of the cell types.

In the following sections, we focus on the overrepresented TFs in some selected groups of cell types such as immune cells, muscle cells, lung cells and embryonic stem cells.

5.4.1 Overrepresented factors in immune-specific DHSs

As first, let us investigate the most overrepresented factors in immune-related cell types in the data set: B-lymphocyte, T-cell, primary T-cell, regulatory T-cell, monocyte, hematopoietic progenitor cell, marrow stromal cells and leukemia.

The cell-type-specific TFs with the strongest association to these cell types are listed in

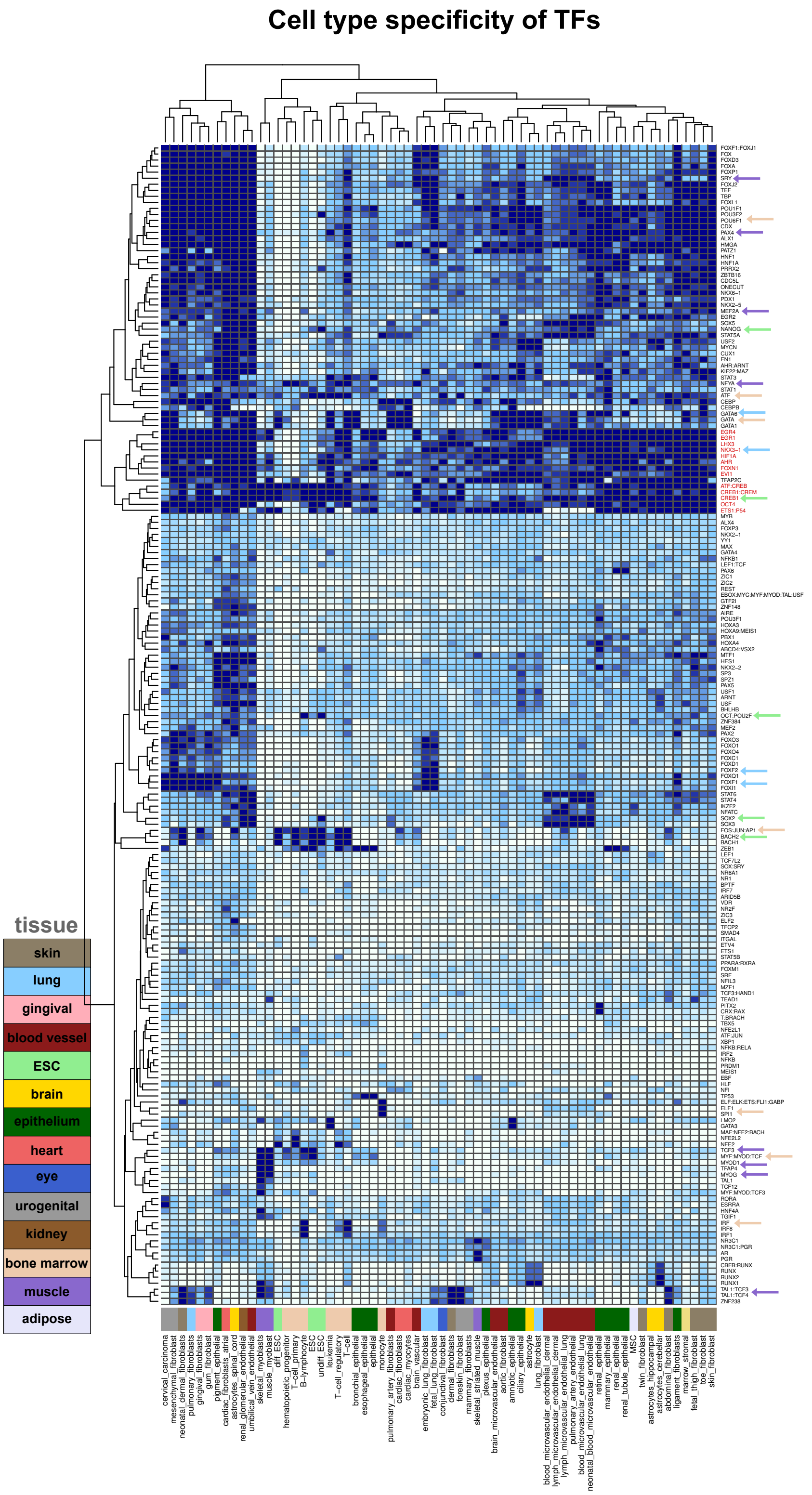


Figure 5.6: Overrepresented transcription factors over 64 cell types. Each cell in the matrix indicates the significance ($-\log_{10} p$-value) of the association between the cell type and the corresponding TF, the darker the matrix cell the higher the significance. TFs overrepresented in the majority of cell types are highlighted in red. Cell-type-specific TFs are marked with arrows of corresponding color.

Table 5.2. There is a high agreement among the most associated factors in the different T-cells with a strong distinction of the hematopoietic progenitor cells from all other immune cell lines. Among the top enriched TFs, large number of known regulators in immune cells such as: ATF family members (with ATF, CREB1, JUN, FOS), ETS family members (with ELF, ELK, ETS, FLI1, GABPA, SPI1), EGR proteins, IRF proteins, GATA proteins, POU-domains proteins (with OCT:POU2F, POU3F2, POU6F1), STAT proteins, CEBP proteins, BACH2, MYOD and TCF3 were found (Matys et al., 2006; Nutt and Kee, 2007). These immune regulators are highlighted in Table 5.2 in bold and marked with a light brown arrow in Figure 5.6.

The top-scoring TFs were enriched in the following functional categories: *transcription regulatory region DNA binding* (the majority of the TFs by definition), *activation of innate immune response* (ATF, CREB1, ELK1, FOS, IRF1, IRF7, JUN, MEF2A, NFKB1, RELA), *cellular response to type I interferon* (IRF1, IRF2, IRF7, IRF8, EGR1) and *erythrocyte differentiation* (ARNT, ETS1, HIF1A, SPI1). The functional analysis was conducted with Ingenuity ® Systems (IPA).

Three of the top-100 factors in B-lymphocyte (E2F, PRDM1, ZNF384) were found as associated TFs with so-called high plasticity regions in B-lymphocytes by Pinello et al. (2014). The high plasticity regions (HPRs) characterize the chromatin-state plasticity in cell-type-specific manner and are enriched in promoters, enhancers and DHSs (Pinello et al., 2014), suggesting that TFs associated with HPRs might be overrepresented on the CTS-DHSs from the corresponding cell line too.

Our findings of enriched TFs in monocytes are in agreement with previous studies in monocytes (Huber et al., 2014; Martens et al., 2014). The enriched transcription factor families in monocyte-specific DHSs, such as: ETS, ATF, IRF, SPI/KLF and CEBP, were found to be expressed in monocytes, to have overrepresented motifs in DNase footprints (Martens et al., 2014) and to have an important regulator function in monocyte differentiation (Huber et al., 2014).

Concerning the enriched factors in leukemia, many of them such as GATA factors and EVI1 were identified in previous studies (Nucifora et al., 2006; Tenen et al., 1997) to have an important function in myeloid leukemia. Further, ETS1, HIF1A and GATA1 are regulators of myeloid cell differentiation (Maglott et al., 2011). In comparison of other immune cell lines, the GATA factors, ETS1 and particularly EVI1 with HIF1A are enriched only in the cancer cell line suggesting that these factors are essential for the regulation of the myeloid leukemia.

Table 5.2: Most significant cell-type-specific TFs in various cell types. TFs in bold are known transcription regulators in the corresponding tissue.

Cell type	cell group	Associated TFs
B-lymphocyte	immune	**CREB1, FOS:JUN, IRF, OCT:POU2F, TCF3, ZEB1**
T-cell	immune	BACH1, **BACH2, CREB1, EGR1, EGR4, FOS:JUN, NKX3-1, SRY, STAT3, TFAP2C**
T-cell primary	immune	BACH1, **BACH2, CREB1, FOS:JUN, NFE2, NFYA, TFAP2C**
T-cell regulatory	immune	**ATF, BACH1, BACH2, CREB1, FOS:JUN, LHX3, NFE2, NFYA, ZEB1**
hematopoietic progenitor	immune	**CREB1, FOS:JUN, GATA, MYOD, NFYA, TCF3, ZEB1**
marrow stromal	immune	AHR, **CREB1, EGR1, EGR4, LHX3, NKX3-1, POU1F1, POU3F2, POU6F1**
monocyte	immune	**CEBP, EGR1, EGR4, ELF:ELK:ETS:FLI1:GABP, HIF1A, IRF, SPI1, STAT5A**
leukemia	immune cancer	AHR, **ATF, CREB1, EVI1, ETS1:P54, GATA1, GATA3, GATA6, HIF1A, LMO2**
brain vascular smooth muscle	muscle	CREB1, EGR1, EGR4, ETS1, FOX, **NFYA, SRY**
cardiac myocytes	muscle	AHR, CEBP, EVI1, GATA, HMGA, **NFYA, PAX4, STAT5A**
muscle myoblast	muscle	AHR, CREB1, EGR1, ETS1, **MEF2A, MYOD, MYOG, TCF3, TFAP2C, TFAP4**
skeletal myoblasts	muscle	CREB1, **MYOD, MYOG , TAL1:TCF, TFAP4**
skeletal striated muscle	muscle	AHR:ARNT, AR, CEBP, ETS1, HIF1A, NKX3-1, NR3C1, PATZ1, PGR, POU1F1, TFAP2C
fetal lung fibroblast	lung	**ETS1, FOXF1, FOXI1, FOXJ2, FOXL1, FOXQ1, NKX3-1, POU1F1, TBP, TEF**
embryonic lung fibroblast	lung	**FOXA, FOXF1, FOXI1, FOXJ2, FOXL1, FOXQ1, NKX3-1, PATZ1, TBP, TFAP2C**
lung fibroblast	lung	**ETS1, EVI1, GATA, GATA6, HIF1A, LHX3, MYCN, NKX3-1, TFAP2C, USF2**
pulmonary fibroblast	lung	CEBP, EGR4, **ETS1, FOXF1, FOXJ2, FOXL1, FOXQ1, NKX3-1, POU6F1, TBP**
ESC	stem cell	ATF, BACH1, BACH2, CREB1, CREM, ETS1, FOS, JUN, MYF:MYOD:TCF, **OCT4, ZEB1**
undifferentiated ESCs	stem cell	AHR, ATF, BACH2, CREB1, CREM, FOS, JUN, **NANOG, OCT4, SOX2**
differentiated ESCs	stem cell	ATF, BACH1, BACH2, CREB1, CREM, ETS1, FOS, JUN, GATA1, NFE2

5.4.2 Overrepresented factors in muscle-specific DHSs

The muscle-related cell types in our data set are: skeletal muscle myoblasts, skeletal striated muscle cells, cardiac myocytes and brain vascular smooth muscle cells. There are several known muscle regulators among the most overrepresented factors, such as MEF2A (myocyte specific enhancer factor 2A), MYOG (myogenic factor inducing myogenesis), NFYA (nuclear transcription factor Y) and TCF3 (Immunoglobulin Transcription Factor 1), marked with a purple arrow in Figure 5.6. Remarkably, none of these known muscle regulators is overrepresented in skeletal striated muscle (see Table 5.2 with the most enriched factors in selected cell lines).

The functional annotation of factors overrepresented in skeletal striated muscle showed one enriched functional category which might take place in muscles: *cellular response to oxygen levels* (for ARNT, HIF1A and NKX3-1). Further, the overrepresented factors in other muscle cells were enriched in the following functional categories (except all functional categories related to DNA-binding and transcriptional regulation): *cell fate commitment* (MYOG, MYOD1, TAL1, TCF3) and *muscle cell differentiation* (GATA6, MEF2A, MYOD1, MYOG, TCF3). The functional analysis was conducted with Ingenuity ® Systems (IPA).

5.4.3 Overrepresented factors in lung-specific DHSs

There are four cell types in the studied data set which originated from lung tissue: fetal lung fibroblast, embryonic lung fibroblast, lung fibroblast and pulmonary fibroblast. Several of the most enriched factors are known regulators in lung development and lung morphogenesis, such as: ETS family members (with ETS1), FOX family members (FOXA, FOXF1, FOXL1) and GATA6 (Maeda et al., 2007; Matys et al., 2006), see blue arrows in Figure 5.6. Many FOX family members are enriched in the embryonic lung fibroblast and the fetal lung fibroblast which is in agreement with previous studies which found FOX genes involved in embryonic development (Maglott et al., 2011).

Interestingly, factor NKX3-1 (NK3 Homeobox 1) belongs to the most enriched factors in all lung-related cell types although it plays an important role in normal prostate development but without any known function in lung. However, another homeobox factor NKX2-5 (NK2 Homeobox 5) is a known regulator in lung morphogenesis (Maeda et al., 2007), which binds to the consensus sequence `5’-[CT]AAGTG-3’`. This motif is actually very similar to the the consensus sequence of the enriched factor NKX3-1, namely `5’-TAAGT[AG]-3’`, suggesting that the motif of NKX2-5 might be enriched in the lung-specific DHSs as well but ranked lower than NKX3-1. The most enriched cell-type specific TFs are listed in Table 5.2.

5.4.4 Overrepresented factors in ESC-specific DHSs

Among the studied cell types, there are three cell lines originating from the embryonic stem cells (ESCs), namely: undifferentiated ESCs (H7-hESC cell line), differentiated ESCs (H7-hESC cell line differentiated after 2-9 days) and ESCs (H1-hESC cell line).
The most enriched factors in all of the three cell lines are: BACH2, CREB1, CREM, FOS and JUN. All these 5 factors have molecular function in tumor apoptosis or proliferation (Ingenuity ® Systems, IPA), FOS-JUN complex is known to be involved in cell differentiation (UniProt Consortium, 2011), CREB1 is implicated in synchronization of circadian rhythmicity and differentiation of adipose cells (UniProt Consortium, 2011), BACH2 is a transcriptional activator or repressor inducing apoptosis (Yoshida et al., 2007) and CREM plays a role in spermatogenesis (Maglott et al., 2011).
There is a large agreement in the top-scoring factors in the differentiated ESCs and in the general ESCs. The main factors responsible for the pluripotency of the ESCs, such as OCT4, NANOG and SOX2 (Chen et al., 2008a; Yeo and Ng, 2013), are enriched only in the undifferentiated ESCs but not in the differentiated ESCs (see green arrows in Figure 5.6). This fact suggests that the undifferentiated-specific DHSs are very distinct from the differentiated-specific DHSs (after few days only) and thus contain different enriched TF motifs. This phenomenon is confirmed in the clustered correlation heat map of the DNase-seq read counts in Figure 5.2. Here, three samples of the differentiated ESCs (after 9 or 14 days of differentiation) cluster together with other differentiated cell lines such as B-lymphocytes, monocytes, and T-cells (see lower left corner of the matrix in Figure 5.2). In contrast, the undifferentiated ESCs and few samples of the early differentiated ESCs (after 2 or 5 days of differentiation) build a separate cluster (see the center of the matrix in Figure 5.2).

5.5 Co-occurrence of TFs in cell-type specific DHSs

After identifying the main enriched factors in cell-type-specific DHSs in previous Section 5.4 our interest now is in predicting pairs of co-occurring TFs which regulate jointly the cell-type-specific gene expression.
To do so, we use the representation of TFs described in Section 5.3 as a ranked list of DHSs (both, cell-type-specific and ubiquitous) ordered by the binding affinity. When studying the co-occurrence of two TFs, we can make use of the information about the behavior in the cell-type specific regions and in the ubiquitous regions. In contrast to the

procedure in Chapter 4, where a 3-way contingency table of shared tissue-specific regions was constructed, we can now derive two 2-way contingency tables of TF co-occurrence in cell-type-specific and in ubiquitous manner, respectively. This new strategy is enabled due to the DHS properties, since we constructed the ranked lists with the *same* number of cell-type specific sites and ubiquitous sites. Such information was not available for the gene promoters analyzed in Chapter 4.

Thus, with two contingency tables describing the TF co-occurrence in the cell-type-specific and in the ubiquitous sites, we can construct a score comparing the significances of the two tables. The information about the association of the TF pair in the ubiquitous sites is used as a sort of background. Then, we select those TF pairs which show a highly significant co-occurrence in the cell-type specific DHSs and not in the ubiquitous DHSs. Thus the selected TF pairs co-occur specifically in the given cell-type and not in a general way. Further, with this method we avoid the problem of artificially large contingency tables when the ranked list would be constructed out of cell-type-specific DHSs of all cell types, as discussed in Section 5.3.

5.5.1 Methods

To predict co-occurring TFs on the cell-type-specific DNase hypersensitive sites (CTS-DHSs), log ratio of p-values from two contingency tables is calculated to compare the significance of a TF pair on the CTS-DHSs to the significance of the TF pair on ubiquitous DHSs. The ranked list representation of TFs as described in Section 5.3 is used to construct the contingency tables.

Similarly to prediction of TF co-occurrence on gene promoters, let us define two binary variables X and Y identifying the top-ranked DHSs for the first TF and for the second TF, respectively. For given thresholds k_1, k_2 defining the top-ranked DHSs for the first and the second TF, respectively and for each DHS i, $i = 1, \dots, 2t$:

$$\begin{aligned} X_{k_1}(i) &= \begin{cases} 1 & \text{if } r_x(i) \leq k_1,\ \ k_1 \in \{1, \dots, 2t\} \\ 0 & \text{otherwise} \end{cases} \\ Y_{k_2}(i) &= \begin{cases} 1 & \text{if } r_y(i) \leq k_2,\ \ k_2 \in \{1, \dots, 2t\} \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{5.7}$$

where $r_x(i)$ and $r_y(i)$ are the ranks of item i in the ranked list R_x and R_y, respectively. In analogy to the previous Section 5.4, binary variable Z_l indicates cell-type specific

DHSs for a particular cell type and is defined as follows:

$$Z_l(i) = \begin{cases} 1 & \text{DHS } i \text{ is cell-type-specific for cell type } l \\ 0 & \text{DHS } i \text{ is ubiquitous .} \end{cases} \tag{5.8}$$

Then two individual tables with fixed values of Z_l can be constructed as shown in Table 5.3, the same notation for cell counts as in Section 4.2 is used.
Due to the construction of the underlying ranked lists with t most cell-type specific and t most ubiquitous DHSs, both of the tables have the same universe t and therefore are simply comparable.

Table 5.3: Two partial contingency table for shared cell-type-specific DNase-hypersensitive sites (top) and for shared ubiquitous DNase-hypersensitive sites (bottom).

cell-type specific: $Z_l = 1$

	$\|i\| : r_y(i) \leq k_2$	$\|i\| : r_y(i) > k_2$	$\sum$
$\|i\| : r_x(i) \leq k_1$	n_{111}	n_{121}	n_{1+1}
$\|i\| : r_x(i) > k_1$	n_{211}	n_{221}	n_{2+1}
$\sum$	n_{+11}	n_{+21}	$n_{++1} = t$

ubiquitous: $Z_l = 0$

	$\|i\| : r_y(i) \leq k_2$	$\|i\| : r_y(i) > k_2$	$\sum$
$\|i\| : r_x(i) \leq k_1$	n_{112}	n_{122}	n_{1+2}
$\|i\| : r_x(i) > k_1$	n_{212}	n_{222}	n_{2+2}
$\sum$	n_{+12}	n_{+22}	$n_{++2} = t$

We aim to predict TF pairs co-occurring in a cell-type-specific manner. Thus we define a log ratio score as the log ratio of the p-values obtained from Fisher's exact test in the cell-type-specific table and of the p-value obtained from the ubiquitous table. Using the same calculation for the p-value as in Eq. 2.7, we define the L_l score for cell type l as:

$$L_l = -\log\left[\frac{P(n_{111} \geq m_l))}{P(n_{112} \geq m_u)}\right] = -\log\left[\frac{\sum_{m=m_l}^{\min(k1,k2)} \frac{\binom{n_{1+1}}{m}\binom{n_{2+1}}{n_{211}}}{\binom{n_{++1}}{n_{+11}}}}{\sum_{m=m_u}^{\min(k1,k2)} \frac{\binom{n_{1+2}}{m}\binom{n_{2+2}}{n_{212}}}{\binom{n_{++2}}{n_{+12}}}}\right] , \tag{5.9}$$

where m_l stands for the observed value of shared top-ranked cell-type-specific DHSs and m_u for the observed value of shared top-ranked ubiquitous DHSs in particular tables. With this definition, the larger the L_l score, the greater the association between the two TFs on the CTS-DHSs in comparison with the ubiquitous DHSs. Thus TF pairs with the highest score in each cell type are predicted as co-occurring TFs in a cell-type-specific manner. Moreover, TF pairs with the largest negative L_l score are TF pairs which co-occur generally on the ubiquitous DHSs and not in the cell-type-specific way. We call them the ubiquitous-specific co-occurring TF pairs. For easier understanding, the method described above is summarized in four steps in Figure 5.7.

Our comparison method with the L_l score can be extended for comparisons between two different cell types of interest. For example, one could investigate TF pairs which co-occur *specifically* in the differentiated cell line but not in the primary cell line. Then, the L_d score will be defined as $L_d = -\log \frac{P(n_{111} \geq m_{\text{differentiated}})}{P(n_{112} \geq m_{\text{primary}})}$.

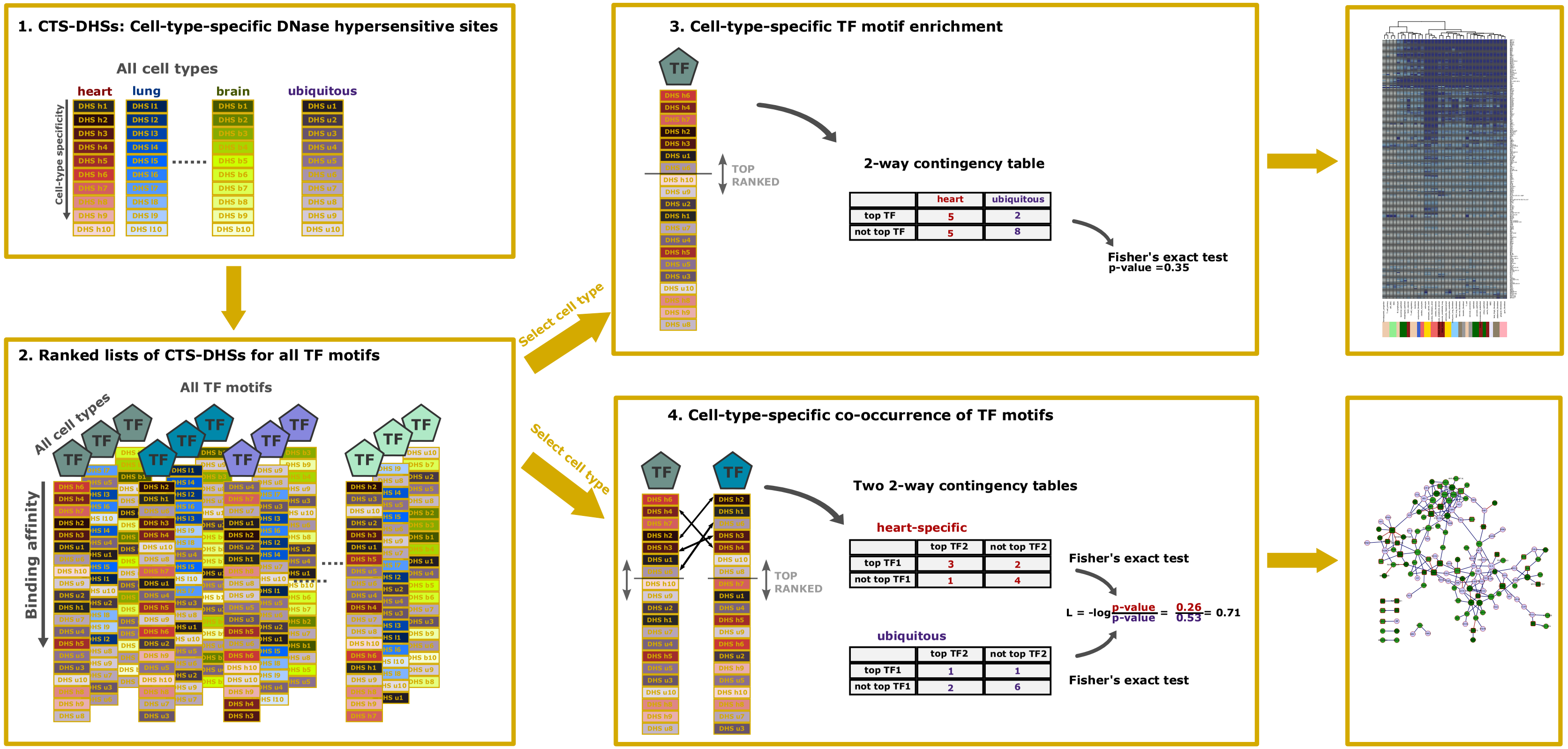


Figure 5.7: Overview of the method for the detection of TF overrepresentation and co-occurrence in the cell-type-specific DNase hypersensitive sites (CTS-DHSs). First, the most cell-type-specific DHSs and most ubiquitous DHSs are determined (1). Then, for each cell type separately, and for each TF of interest, CTS-DHSs and ubiquitous DHSs are jointly ranked by the binding affinity (2). Enriched TF motifs in CTS-DHSs are identified by constructing a 2-way contingency table and calculating Fisher's exact test (3). Co-occurring TF pairs on CTS-DHSs are predicted from the log score of p-values derived from two contingency tables with Fisher's exact test (4).

5.5.2 Parameter choice

The L_l score depends on two threshold parameters k_1 and k_2 defining the top-ranked sites for the first and second TF in the pair, respectively. In addition, the construction of the partial contingency tables depends on the choice of parameter t which selects the most cell-type-specific sites and most ubiquitous sites.

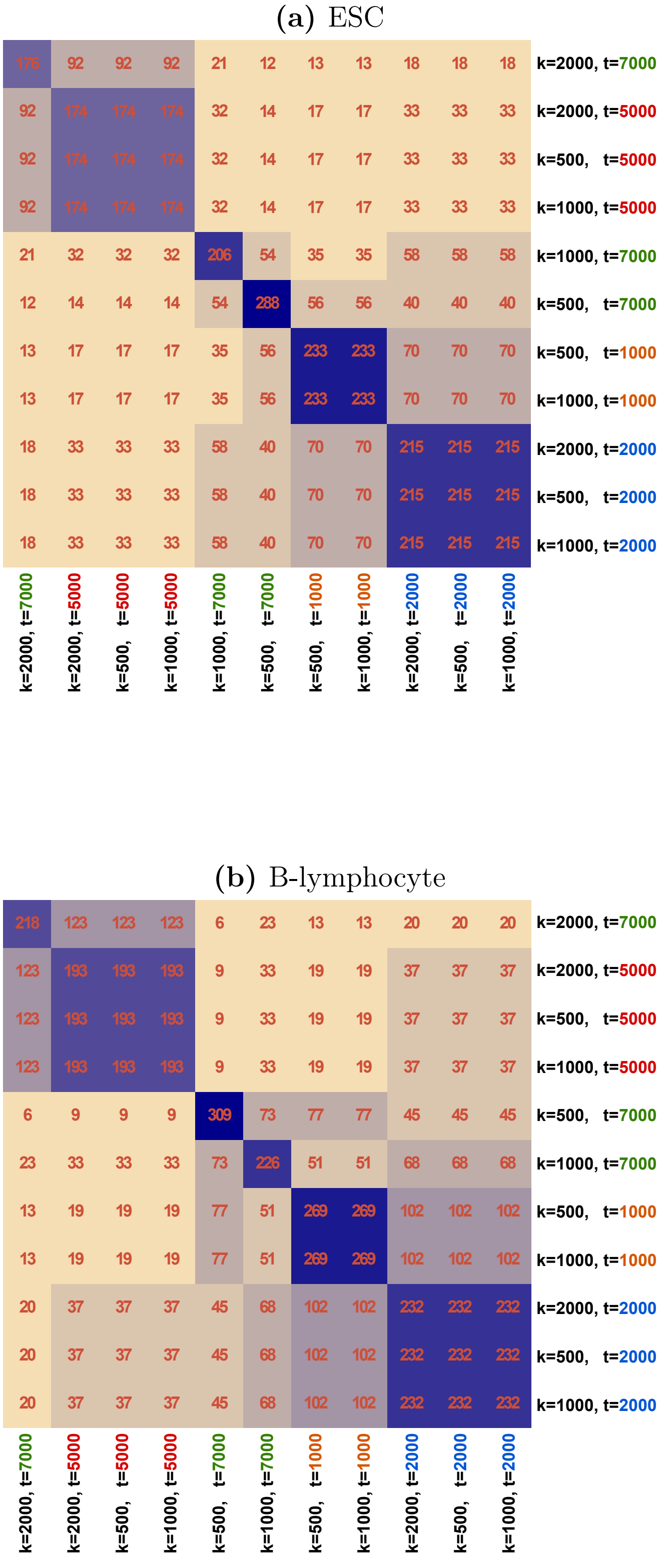
To investigate the stability of the obtained results, we calculated the L_l scores for all TF pairs for different combinations of parameters. Then TF pairs with the largest L_l scores ($L >$ 99.5%-quantile) were selected and studied for the consistency of obtained results. For easier comparison, thresholds k_1 and k_2 were chosen to be equal in all combinations; all pairs of TF motifs were mapped to the corresponding TF names or groups.

Two matrices with numbers of identical TF pairs predicted with various combinations of parameters $k_1 = k_2$ and t in embryonic stem cells (ESCs) and in B-lymphocytes are shown in Figure 5.8. Both matrices show very similar trend: with the choice of $t \in \{1000, 2000, 5000\}$ the choice of k_1 and k_2 is not important for the group of highly significant TF pairs; the predicted TF pairs remain exactly the same for different values of $k_1 = k_2$ (see dark blue boxes along the diagonal in Figure 5.8). Setting the threshold $t = 7000$ causes a high variability between the significant TF pairs defined with different thresholds $k_1 = k_2 \in \{500, 1000, 2000\}$. Further, there is a high similarity of the predictions derived with $t = 2000$ and $t = 1000$ with different values of $k_1 = k_2$. The agreement between the two sets of significant TF pairs is 102 out of 232 and 269 TF pairs (in B-lymphocytes, Figure 5.8b) and 70 out of 215 and 233 TF pairs (in ESCs, Figure 5.8a). The combination of thresholds $t = 7000$ and $k_1 = k_2 = 2000$ gives very similar results to threshold $t = 5000$ with various values for $k_1 = k_2$. This behavior can be observed for the majority of studied cell types.

These results suggest that the strongest cell-type-specific signal is among the top 5000 CTS-DHSs, since for this threshold $t \leq 5000$ highly consistent results were obtain. For further analysis, we use the longest possible list of CTS-DHSs and ubiquitous DHSs with $t = 5000$ and choose as appropriate threshold $k_1 = k_2 = 1000$.

5.5.3 Predicted co-occurring TF pairs on CTS-DHSs

Before applying calculation of the L_l score to all possible TF pairs, the TF motifs were clustered into 138 distinct groups based on their similarity according to Oh et al. (2012). Using a group (e.g. cluster) of TF motifs rather than a single motif representing the whole TF family to identify their binding regions is more effective and practical as suggested by Oh et al. (2012).



(a) ESC

k=2000, t=7000	k=2000, t=5000	k=500, t=5000	k=1000, t=5000	k=1000, t=7000	k=500, t=7000	k=500, t=1000	k=1000, t=1000	k=2000, t=2000	k=500, t=2000	k=1000, t=2000	
176	92	92	92	21	12	13	13	18	18	18	k=2000, t=7000
92	174	174	174	32	14	17	17	33	33	33	k=2000, t=5000
92	174	174	174	32	14	17	17	33	33	33	k=500, t=5000
92	174	174	174	32	14	17	17	33	33	33	k=1000, t=5000
21	32	32	32	206	54	35	35	58	58	58	k=1000, t=7000
12	14	14	14	54	288	56	56	40	40	40	k=500, t=7000
13	17	17	17	35	56	233	233	70	70	70	k=500, t=1000
13	17	17	17	35	56	233	233	70	70	70	k=1000, t=1000
18	33	33	33	58	40	70	70	215	215	215	k=2000, t=2000
18	33	33	33	58	40	70	70	215	215	215	k=500, t=2000
18	33	33	33	58	40	70	70	215	215	215	k=1000, t=2000

(b) B-lymphocyte

k=2000, t=7000	k=2000, t=5000	k=500, t=5000	k=1000, t=5000	k=500, t=7000	k=1000, t=7000	k=500, t=1000	k=1000, t=1000	k=2000, t=2000	k=500, t=2000	k=1000, t=2000	
218	123	123	123	6	23	13	13	20	20	20	k=2000, t=7000
123	193	193	193	9	33	19	19	37	37	37	k=2000, t=5000
123	193	193	193	9	33	19	19	37	37	37	k=500, t=5000
123	193	193	193	9	33	19	19	37	37	37	k=1000, t=5000
6	9	9	9	309	73	77	77	45	45	45	k=500, t=7000
23	33	33	33	73	226	51	51	68	68	68	k=1000, t=7000
13	19	19	19	77	51	269	269	102	102	102	k=500, t=1000
13	19	19	19	77	51	269	269	102	102	102	k=1000, t=1000
20	37	37	37	45	68	102	102	232	232	232	k=2000, t=2000
20	37	37	37	45	68	102	102	232	232	232	k=500, t=2000
20	37	37	37	45	68	102	102	232	232	232	k=1000, t=2000

Figure 5.8: Consistency of the most significant TF pairs on CTS-DHSs for different combinations of parameters in a) embryonic stem cell (ESC) and b) B-lymphocyte. The matrix entries denote the number of identical TF pairs with the highest L_l score for 11 combinations of thresholds $k_1 = k_2$ (the first number) and of threshold t (second number).

In addition, TF pairs solely from different clusters are considered for testing to avoid unnecessary comparison of pairs with highly similar motifs and to reduce the total number of tests, resulting in total of 111 241 pairs of TF motifs. Moreover, with the nature of the L_l score method, the significance of TF pairs with very similar TF motifs as discussed in Section 3.3 should not confound the analysis. Namely, we focus on those TF pairs which show high association in the ranked lists only in the CTS-DHSs and not in the ubiquitous DHSs. Thus TF pairs with highly similar TF motifs should not obtain high L_l scores because both p-values - in the CTS-DHSs and in the ubiquitous DHSs - should be very small.
With the selected parameters $k_1 = k_2 = 1000$ and using $t = 5000$ most cell-type-specific and ubiquitous DHSs, the corresponding tables for all 111 241 TF pairs in all 64 cell types were built and the corresponding p-values were calculated. This results in total number of 14 238 848 tests. Before calculating the L_l score, the p-values were corrected for multiple testing with Benjamini-Hochberg method (Benjamini and Hochberg, 1995), considering each cell type separately. To avoid comparisons in the extreme tails of the hypergeometric distribution, the minimal value of all corrected p-values were set to 10^{-10}.
Then, separately for each cell type we identified significant TF-motif pairs as pairs with the L_l score larger than the 99.5%-quantile of the empirical distribution of L_l scores in the corresponding cell type, achieving total number of 5 257 significant TF-motif pairs. The significant TF-motif pairs were then aggregated to their corresponding pairs of transcription factors (or transcription factor groups), resulting in total number of 2 359 significant TF pairs.

First, the agreement of identical significant TF pairs among the different cell types was investigated. Our aim is to find co-occurring TFs in a *cell-type-specific* way, thus we expect that the significant co-occurring TF pairs would be unique for the particular cell type. This insight could be compared when looking at the matrix of overlapping significant TF pairs over all 64 cell types as shown in Figure 5.9. The majority of the significant TF pairs is unique for the corresponding cell type (shown with the dark cells on the diagonal). The pluripotent cell groups on the top of the matrix (primary T-cell, hematopoietic progenitor cells, embryonic stem cells) are very distinct from all other differentiated cell types, the agreement of the significant TF pairs predicted in the pluripotent cells and in the differentiated cells is very low (see light yellow cells in Figure 5.9). Further, many functionally related cell types such as renal glomerular endothelial cells and umbilical vein endothelial cells or microvascular endothelial cells originated from different tissues share high number of significant TF pairs (see dark blue cells in

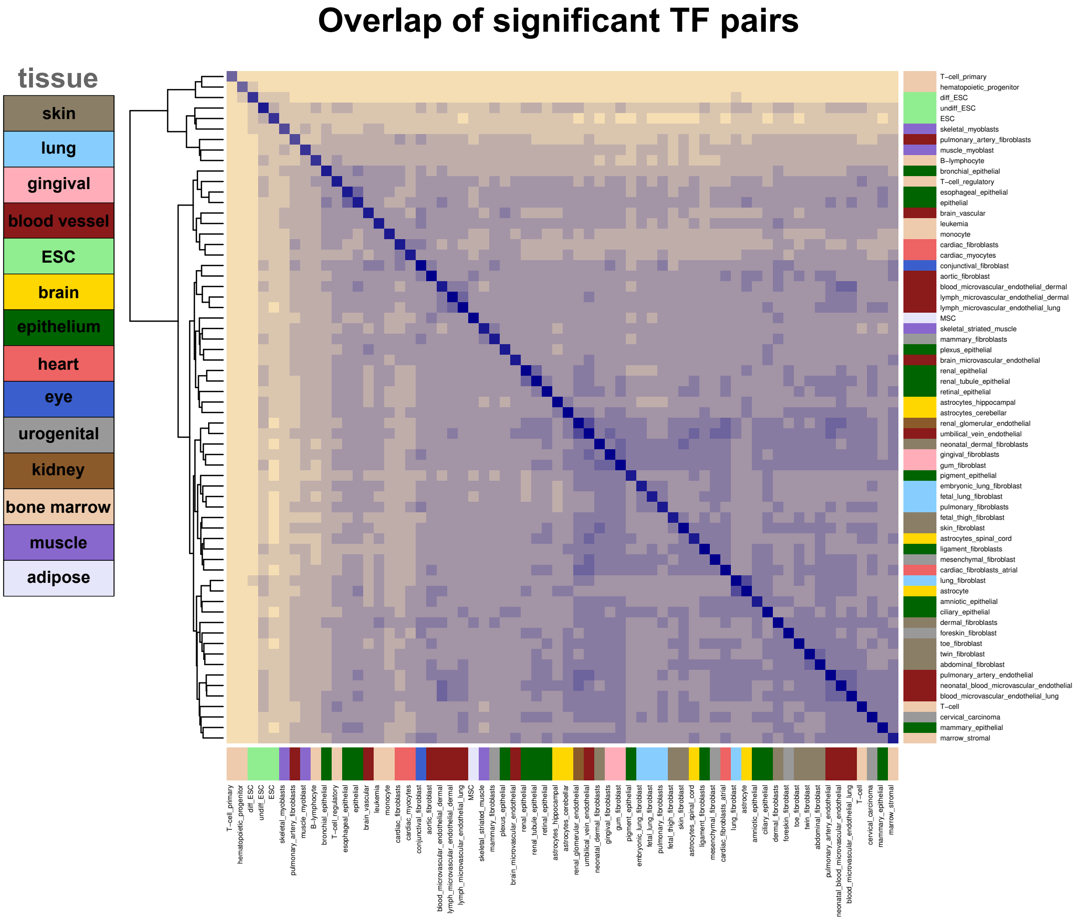


Figure 5.9: Heatmap of overlapping significant TF pairs over 64 cell types. Each cell corresponds to the number of identical TF pairs significant in the corresponding cell types. Dark blue color denotes large numbers, light yellow color denotes small numbers close to zero.

Figure 5.9).

In general, we identified 158 highly frequent TF pairs which are significant in at least 30 out of 64 cell types. The main factors among these frequent TF pairs are mainly homeoboxes (ALX1, POU2F1, ONECUT, HNF1, homeodomain NKX factors) and members of the forkhead-box (FOX) family. This finding suggests that the cell-type-specific DHSs are enriched for homeobox and forkhead-box binding motifs. The functions of the most frequent TFs are very broad: POU2F1 have a general function in embryonic development, HNF1 and ONECUT are involved in the development of endoderm, TBP and ALX1 are more general factors involved in various processes, NKX factors are involved in cell differentiation, proliferation and death and FOX genes are involved in various processes such as embryonic development, cell cycle regulation, tissue-specific gene expression, cell signaling, and tumorgenesis (Bieller et al., 2001; Ingenuity ® Systems, IPA; Maglott et al., 2011; UniProt Consortium, 2011).

The most enriched functions of these TFs found with Ingenuity ® Systems (IPA) analysis were general functions such as *cellular and organismal development* (52 and 39 out of 68 TFs, respectively) and *embryonic development* (39 out of 68 TFs). The network derived from the most frequent TF pairs is shown in Figure 5.10. It is dominated by a large interconnected subnetwork with main nodes (TFs): POU2F1, ALX1, TBP, ONECUT, HNF1, NKX6-2 and NKX3-1. Among the very frequent significant TF pairs, there were several already known protein-protein interactions between: POU2F1:TBP, POU3F2:TBP, FOXA:ONECUT and STAT3:NFKB1 (highlighted in red in Figure 5.10). Over all cell types, most significant TF pairs were found in dermal fibroblast with 320 pairs, the smallest number of significant TF pairs was found in primary T-cells with 161 TF pairs.

In the following, the regulatory networks derived from significant co-occurring TF pairs in immune cells, embryonic stem cells (ESCs) and in muscle cells are discussed.

5.5.4 Co-occurring TF pairs in immune-specific DHSs

The significant TF pairs in immune cells, such as hematopoietic progenitor cells, B-lymphocyte, T-cell (primary and regulatory), monocyte and leukemia were investigated. The most significant co-occurring TF pairs were found in T-cells (312) and leukemia (288), the least significant TF pairs were found in primary T-cells (161) and hematopoietic progenitor cells (184).

The single TFs in the co-occurring TF pairs were first analyzed for their expression in the corresponding cell type and second for a known regulatory function in hematopoiesis. Roughly one quarter of the factors (from 21% in B-lymphocytes to 38% in monocytes)

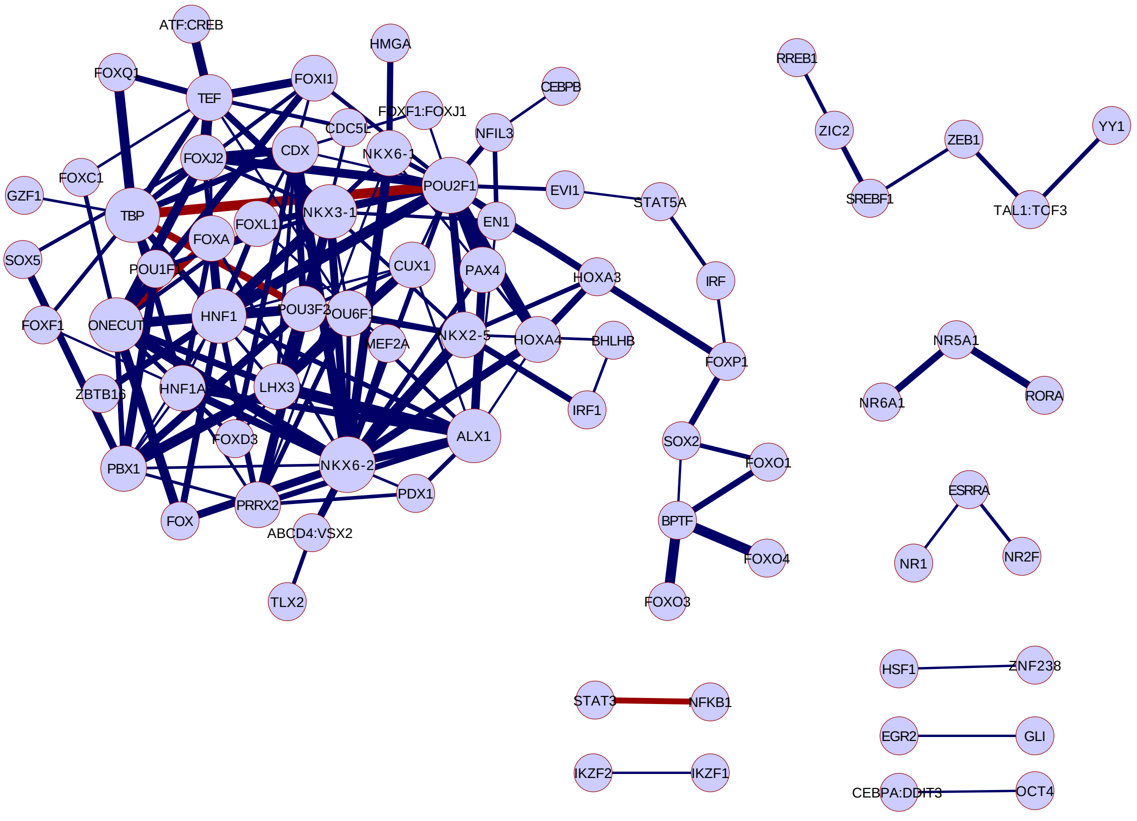


Figure 5.10: Network of highly frequent significant TF pairs in the majority of cell lines. Nodes in the network represent transcription factors, edges are drawn between the significant TF pairs. Red edges are known protein-protein interactions. The width of the edges corresponds to the frequency over various cell lines.

are known regulators in hematopoietic differentiation or in the corresponding cell line (Matys et al., 2006; Nutt and Kee, 2007; Ramirez et al., 2010; Wilson et al., 2010). Further, more than three quarters of all factors (from 75% in T-cells to 80% in hematopoietic progenitor cells) are expressed in immune cells, when comparing with the Ensembl database (Flicek et al., 2014, release 75 based on RNA-seq experiments). One has to point out that all these factors were selected *only* by the high L_l score without any knowledge of their possible function or expression in the corresponding cell type.

To focus on the cell-type-specific TF co-occurrence, we removed all general significant TF pairs which appear in 30 or more cell lines and construct regulatory networks from all significant TF pairs in the particular cell type.

The regulatory network in the hematopoietic progenitor cells consists then of 178 edges (TF pairs) among 120 nodes (TFs) and is dominated by a highly connected large subnetwork with 150 edges among 78 nodes and a smaller subnetwork with 9 nodes and 9 edges, followed by several triplets and pairs of co-occurring TFs. Most of the key regulators of white blood progenitor cells such as: EVI1, TCF3, GATA, LEF1, IKZF1, IRF1 and bHLH-binding proteins SREBF and USF (Matys et al., 2006; Nutt and Kee, 2007; Ramirez et al., 2010; Wilson et al., 2010) are present in the network (see Figure 5.11, known regulators are highlighted as rectangles with red borders). Factors with most co-occurring partners in the network are EVI1, GATA, POU6F1, ONECUT and TEF. Whereas EVI1 and GATA are known regulators in hematopoietic stem cells, the homeodomain proteins POU6F1 and ONECUT are known regulators of pluripotency and differentiation (Maglott et al., 2011; UniProt Consortium, 2011). Thyrotrophic Embryonic Factor (TEF) is involved in the embryonic development of pituitary gland and is expressed in various adult tissues, however, it shows functional homology with other basic region/leucine zipper (bZIP) family members such as the hepatic leukemia factor, which is involved in transcriptional regulation in lymphoblastoids (Maglott et al., 2011). Further, the predicted network includes several experimental validated protein-protein interactions (PPIs) listed in Chatraryamontri et al. (2013) or validated by Ravasi et al. (2010). The known complex of NFKB/RELA and STAT3 builds one of the triplets, the known interactions GATA:MEF2A and GATA:POU1F1 as well as the known interactions CEBP:ATF/CREB and CEBP:STAT5A are part of the large subnetwork (red edges in Figure 5.11).

For comparison, let us investigate the transcriptional network derived for a differentiated immune cell such as monocytes, see Figure 5.12. The network consists of 209 edges (TF pairs) among 120 nodes (TFs) and is dominated by a large subnetwork of 179 TF pairs among 86 TFs. The most connected hubs are distinct from the hubs in hematopoietic progenitor cells, namely: POU2F1, CDX, LHX3, PBX1, ALX1, NKX3-1, NKX6-1 and TEF. POU2F1 is a known regulator in the hematopoiesis (Matys et al., 2006); among the caudal type homeoboxes (CDXs), CDX1 can inhibit T-cell factor transcriptional activity and CDX4 is involved in hematopoiesis. It is known, that LHX3 with POU1F1 synergistically enhance the transcription of prolactin, which is a growth regulator of immune cells (Maglott et al., 2011). This TF pair LHX3:POU1F1 is significant in our transcriptional network in monocytes. Further, factor PBX1 can be associated with lymphoblastoid leukemia (Maglott et al., 2011). Factors ALX1, NKX3-1, NKX6-1 and TEF do not have any known specific function in monocytes or immune cells.

Further, the large network in monocytes includes two small subnetworks of 15 and 8 TFs, one quintet, one triplet and two single TF pairs. The subnetwork with 15 TFs includes 4 known regulators in immune cells (ATF, TCF3, retinoid receptor RAR/RXR/THR and

VDR-nuclear receptor NR1/RXRA) and some general receptors like TFAP2 (Activating Enhancer Binding Protein 2) and ESR (Estrogen Receptors). The subnetwork with 8 TFs involves 4 immune cell regulators NFKB1, ZIC3 and zinc finger proteins EGR1 and ZNF148. The quintet describes a co-occurrence of 4 regulating factors in immune cells: SREBF1, KLF12, MZF1 and ZIC1. In agreement with other studies (Huber et al., 2014), most of the known regulators of monocyte differentiation such as SPI1, CEBP proteins, IRF proteins, VDR nuclear receptors (RXRA, NR1), STAT1 and STAT3 proteins are present in the predicted regulatory network. Further, known direct PPIs were detected as significantly co-occuring TF pairs: CEBP:ATF/CREB, CEBP:POU2F1, ATF/CREB: STAT5A, MAF and ETS-binding proteins, ESR:NR2F1 and the complex STAT3:NFKB1:RELA.

Further, we compare the two regulatory networks in immune cell such that we calculate the number of shared TF partners in both networks and the number of distinct significant TF partners for all nodes in the networks. These values for all nodes are summarized in Figure 5.13. The bar plots show the number (or proportion) of co-occurring partners in hematopoietic progenitor cells (red), in monocytes (blue) and in both networks (black). TFs with large number of significant partners in hematopoietic progenitor cells that are not present in monocytes are: GATA factors, MYC:MAX, ETFA/TEAD2, ARID5B and E2F family factors.

GATA proteins are of special interest in erythropoiesis as they play a crucial role in the maintenance and proliferation of immature hematopoietic progenitors (Doré et al., 2012; Ohneda and Yamamoto, 2002). With our analysis we showed that GATA factors co-occur with their TF partners in regulatory regions of hematopoietic progenitor cells but do not occur in the regulatory regions of monocytes. This is in agreement with the known functionality of GATA factors in erythropoiesis and is reflected in the cell-type-specific regulatory network in hematopoietic progenitor cells. Furthermore, c-Myc (with MYC:MAX binding motif) is one of the transcriptional regulators of pluripotency, thus it very likely co-occur with other TFs in the undifferentiated cells (hematopoietic progenitor cells) rather than in the differentiated cells (e.g. monocytes).

On the other hand, factors which do not occur in the hematopoietic progenitor network but have many significant pairs in monocyte network are: SPI1 and other ETS transcription factor family members, ZIC3, PRDM1, FOXD3 and FOXJ2. Among them, SPI1 is the most interesting as it was shown to be the major regulator of monocytic differentiation (Gangenahalli et al., 2005; Huber et al., 2014). Factors, which share the most partners in both networks are: POU6F1, CDC5L, LHX3, TEF, PBX1, STAT3 and EVI1. Among them, EVI1, STAT3, PBX1, and LHX3 have known functions in white

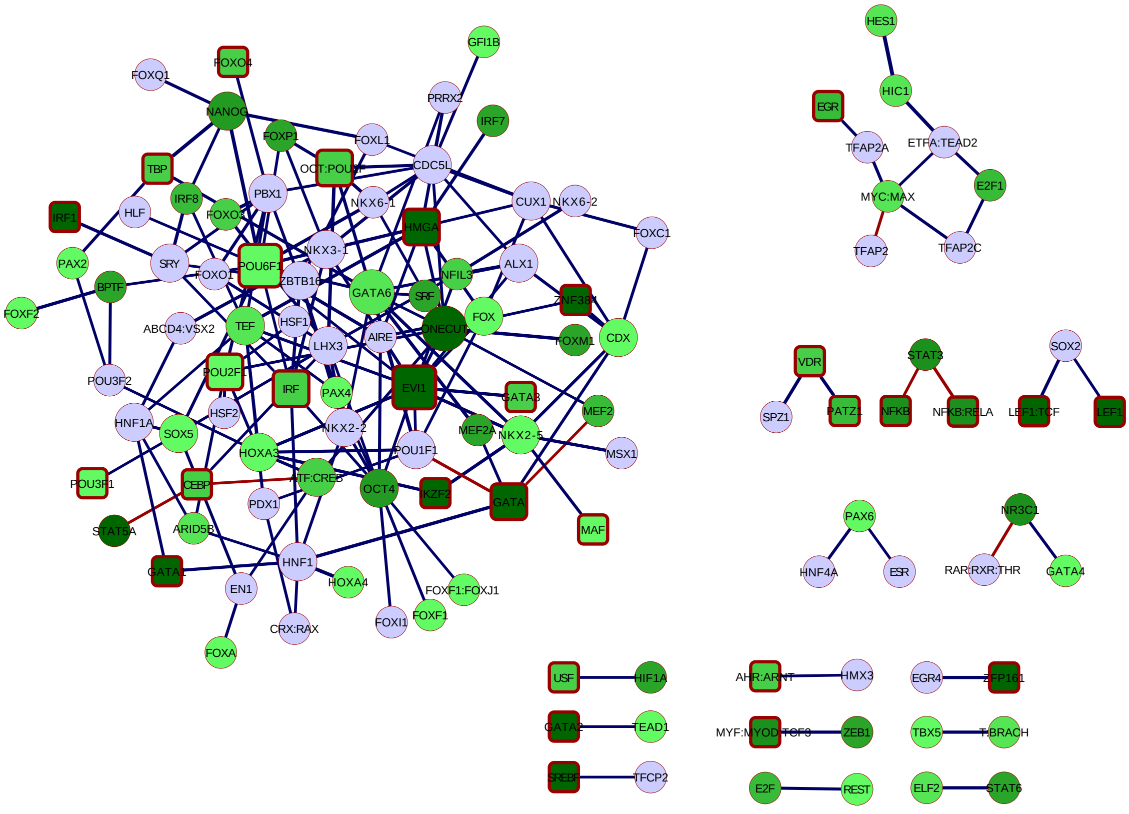


Figure 5.11: Network of significant TF pairs in hematopoietic progenitor cells. Nodes in the network represent transcription factors, edges are drawn between the significant TF pairs. Red edges are known protein-protein interactions. TFs expressed in the cell line are highlighted in green with darker tone indicating higher evidence; known regulators in the corresponding cell type are highlighted as rectangles with red border.
]

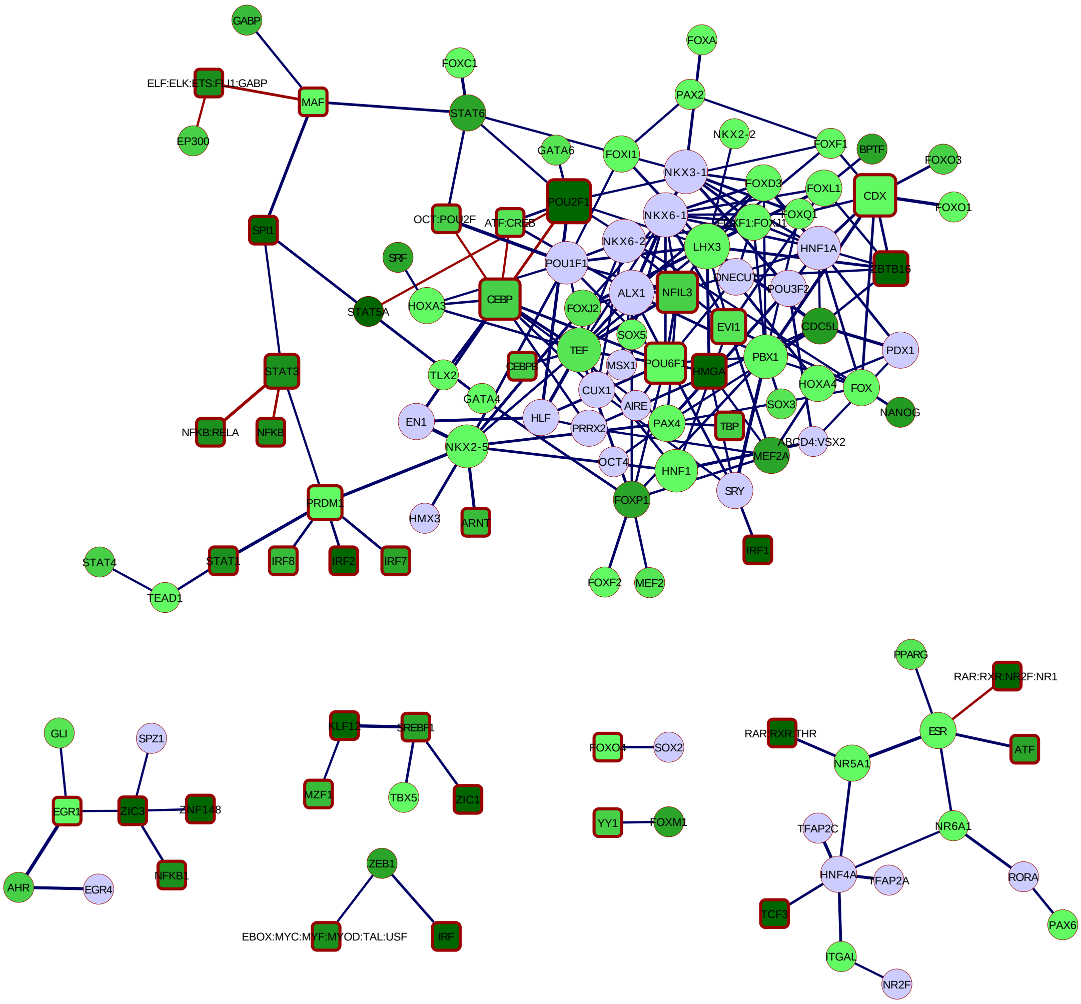


Figure 5.12: Network of significant TF pairs in monocytes. Nodes in the network represent transcription factors, edges are drawn between the significant TF pairs. Red edges are known protein-protein interactions. TFs expressed in the cell line are highlighted in green with darker tone indicating higher evidence; known regulators in the corresponding cell type are highlighted as rectangles with red border.

blood cells.

In addition, we compared the results which we obtained from the prediction of co-occurring TF pairs on promoters specific for hematopoietic stem cells (HSCs) in Section 4.5.3 with the results obtained from the prediction on white-blood-specific DHSs. The network derived on promoters is dominated by two central regulators NFYA and ELK1. However, NFYA do not occur in any of the white-blood-specific regulatory network on DHSs, thus we could not find any agreement between the two approaches. For the ELK1 co-occurring partners, only two pairs ELK1:IRF1 and ELK1:KLF4 could be found in the white-blood-specific regulatory networks too.

5.5.5 Co-occurring TF pairs in ESC-specific DHSs

Next, we studied the transcriptional network in embryonic stem cells (ESCs). There are three different sorts of cell type in our data which are related to ESCs: undifferentiated ESCs (H7-hESC cell line), differentiated ESCs (H7-hESC cell line differentiated after 2-9 days) and ESCs (H1-hESC cell line).

To focus on the ESC-specific significant TF pairs only, the frequently occurring TF pairs in many cell types were removed from the networks, as described in Section 5.5.4. Then, the transcriptional networks in embryonic stem cells have between 216 (differentiated ESCs) and 234 significant TF pairs (undifferentiated ESCs) involving 127-147 distinct TFs. The majority ($67 - 76\%$) of the regulators in the networks is expressed in the ESCs (Flicek et al., 2014) and approximately 20% of the regulators have known function in pluripotency or early development.

The main regulators in the undifferentiated ESCs with the highest number of significant TF partners are: OCT4, NANOG, POU2F1, LHX3, ZBTB16 and PAX4. Among them, OCT4 and NANOG are two of three most important pluripotent factors in human embryonic stem cells (Chen et al., 2008a). Besides, most of the known pluripotent factors (SOX2, general motif SOX:SRY and OCT:POU2F motif) and most of known early developmental regulators (STAT3, FOXD3, ESRR, KLF4, MYC, TCF3, zinc finger proteins and YY1 as a part of polycomb complex, Chen et al., 2008a,b) are present in our transcriptional network (see rectangles with red borders in Figure 5.14).

The transcriptional network consist of a main subnetwork with 210 edges and 119 nodes, a small subnetwork of 17 edges and few single TF pairs or triplets. The main subnetwork is composed of two parts which are connected with factors PAX4 and STAT6. Both of these factors have no known function in ESCs. PAX4 is a transcriptional regulator which is able to differentiate the stem cells into primary pancreatic cells (Blyszczuk et al., 2003; Liew et al., 2008), thus the presence of its binding motif in the ESC regulatory regions

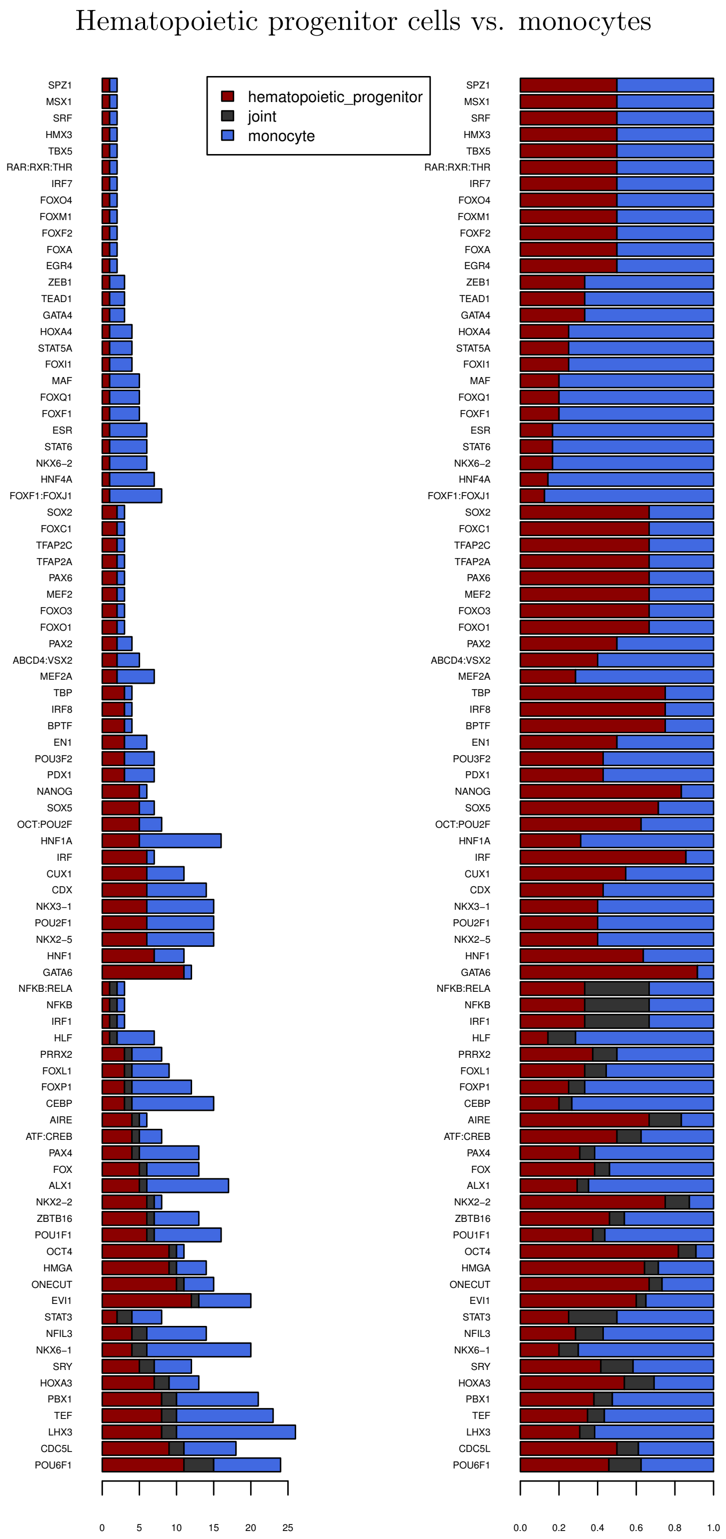


Figure 5.13: Comparison of significant TF pairs in hematopoietic progenitor cells and in monocytes. For each TF, the barplot shows the number of distinct co-occurring partners in hematopoietic progenitor cells (red) and in monocytes (blue) and the number of shared co-occurring partners on both cell lines (black). The left column shows the absolute numbers, the right column shows the proportions.

is possible. The two parts of the main subnetwork suggest different functionality in ESC regulation. One part includes the main pluripotent factors (OCT4, NANOG, SOX) with CDX and FOXD3 regulators and with other FOX genes and some NK-Homeoboxes. The function of this cluster is mainly the maintenance of the pluripotency and of the self-renewal (Chen et al., 2008a). The other part of the subnetwork includes KLF4, STAT3, ZIC3 and ZNF148 transcription factors which are all regulated by the three pluripotent factors OCT4, NANOG and SOX2 and which regulate the early development (Neph et al., 2012). The small subnetwork of only 17 TFs could be defined as a MYC-cluster with TCF3, TBX5 and YY1.

It was shown in various studies that the main pluripotent factors OCT4, NANOG, SOX2 and SMAD1 bind in clusters in the regulatory regions in ESCs (Chen et al., 2008b). In our predicted network, the TF pairs OCT4:NANOG, OCT4:SOX2, NANOG:SOX/SRY have large L_l scores too and fully agree with the experimental findings. Further, the network contains some experimentally proven direct protein interactions: NANOG:TCF/LIF1, POU2F1:TLX2, CEBP:ATF/CREB, CEBP:OCT/POU2F, MAF:ETS, STAT3:NFKB, TAL1:LMO2, LMO2:MYC (see red edges in Figure 5.14).

The predicted transcriptional network in the differentiated ESCs consists of 216 significant TF pairs with 147 TFs. It includes a main subnetwork of 209 edges and 134 nodes, one triplet and 5 single TF pairs, see Figure 5.15. The main regulators with the most significant partners are ALX1, EGR, E2F1, ESRRA, LEF1/TCF, NANOG, GATA4 and TFAP2 proteins. Among them, NANOG, ESRRA and E2F1 (Chen et al., 2008a; Yeo and Ng, 2013) are pluripotent factors in stem cells, LEF1/TCF is a regulator in the hematopoietic primary cells (Nutt and Kee, 2007). GATA4 together with GATA6 is necessary for differentiation into primitive endoderm (Aronson et al., 2014; Murakami et al., 2005) and EGR and TFAP2 proteins are general transcription factors involved in development of many tissues and organs (Maglott et al., 2011). The specific function of ALX1 is still unknown.

The large network in differentiated ESCs can be separated in four smaller subnetworks. The first part with the pluripotent factors SOX2, NANOG, FOXC1 is connected to the second part, that is dominated by the GATA proteins with co-occurring TFs important for the development of blood (EVI1, IRF) and for the development of muscles (SRF). The third part of the network, as well connected with the pluripotent factors, is dominated by the ESR (or ESRRA) factor, which is a target of NANOG and OCT4 serving for the maintenance of the cell pluripotency. Here, ESR co-occurs with HNF4 and NR2F factors which are important for the mesoderm differentiation. The fourth part of the network is a MYC-centered subnetwork with E2F1 factor, KLF4, zinc finger proteins and general regulators TFAP2 and SP1. MYC-centric clusters of co-binding with E2F1, zinc finger protein ZFX and CTCF were already shown with a ChIP-seq study by Chen et al. (2008b). The regulatory network includes some of the known protein interactions such as: MYC:SP1, ESR:NR2F/HNF4, SOX:NANOG, GATA:SRF, GATA1:ZBTB16 and TCF3:LMO2.

Undifferentiated ESCs

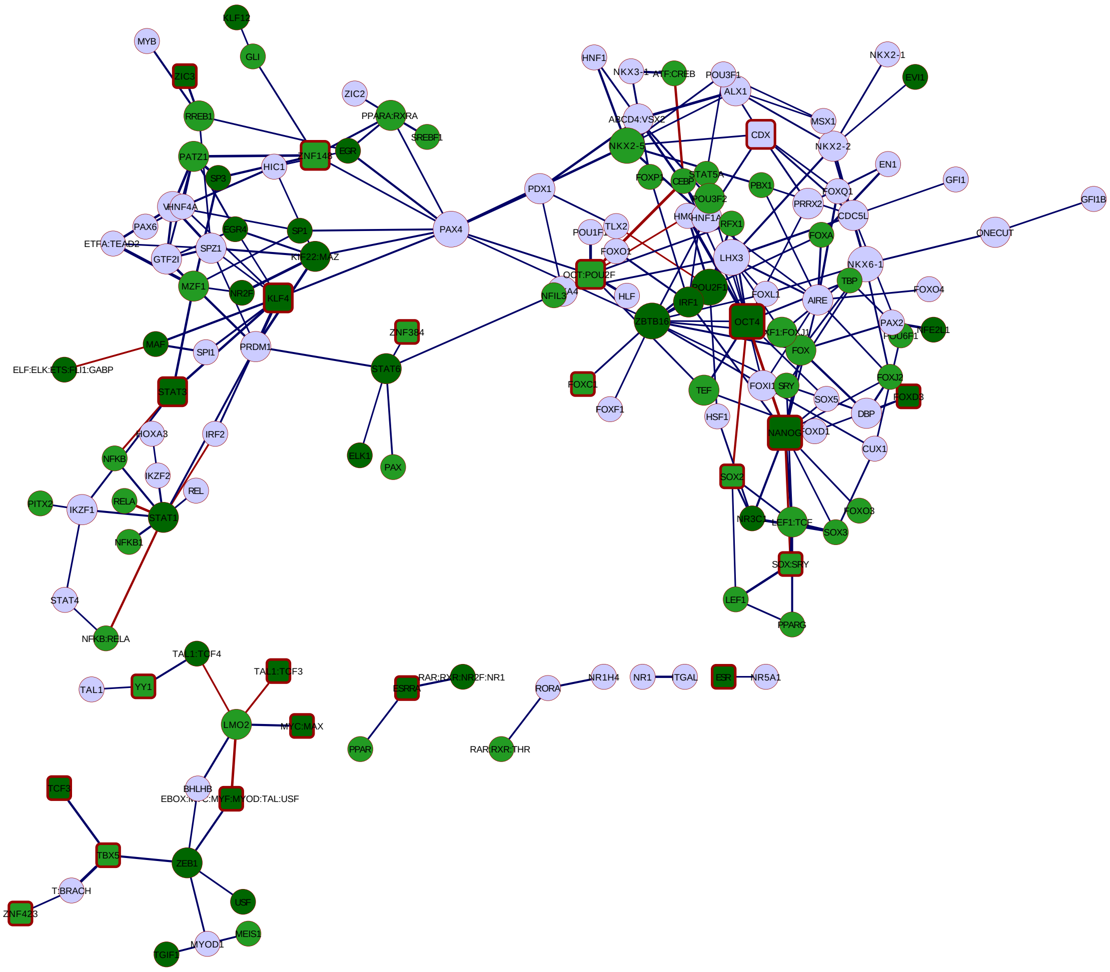


Figure 5.14: Network of significant TF pairs in undifferentiated embryonic stem cells. Nodes in the network represent transcription factors, edges are drawn between the significant TF pairs. Red edges are known protein-protein interactions. TFs expressed in the cell line are highlighted in green with darker tone indicating higher evidence; known regulators in the corresponding cell type are highlighted as rectangles with red border.

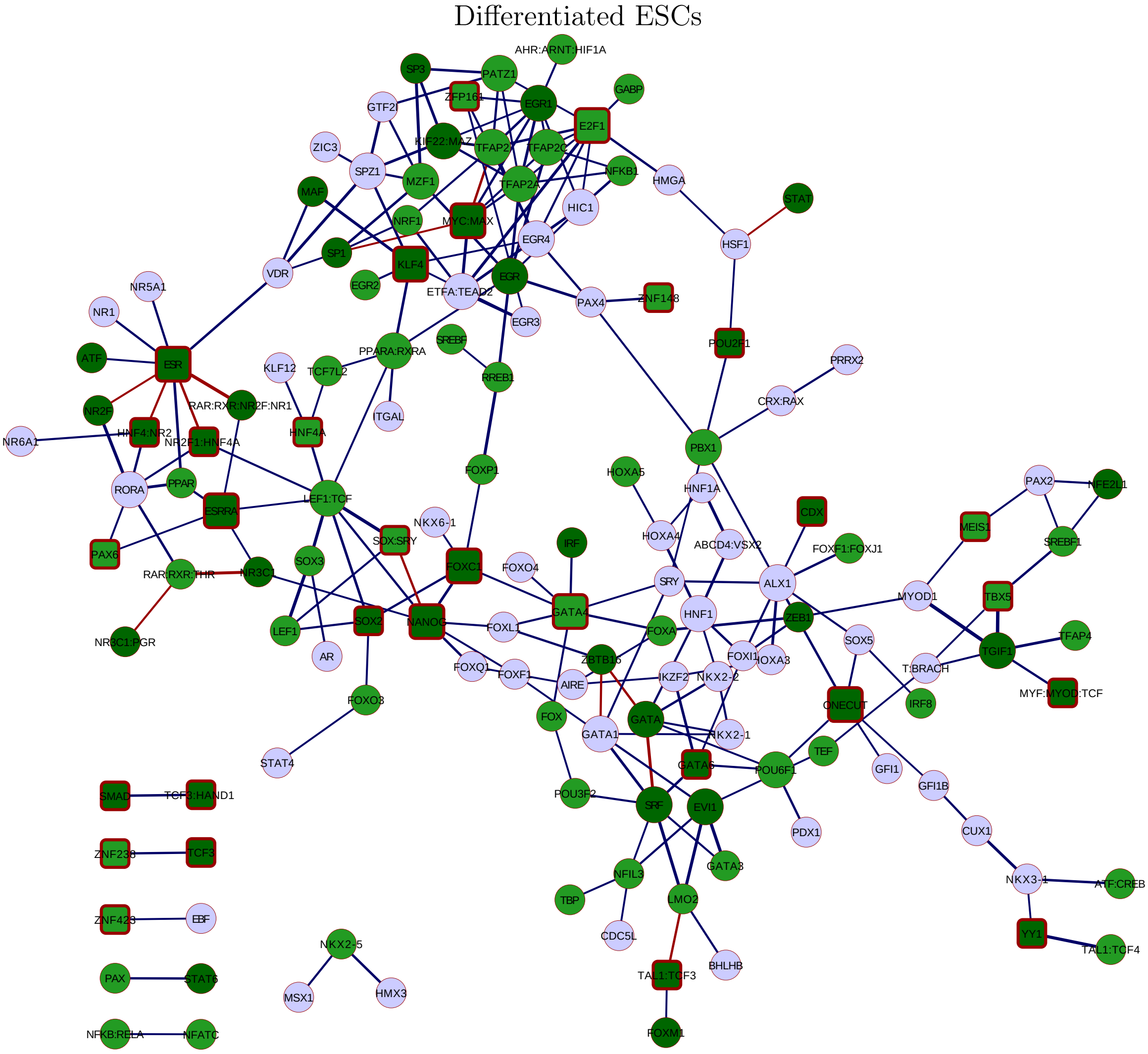


Figure 5.15: Network of significant TF pairs in differentiated embryonic stem cells. Nodes in the network represent transcription factors, edges are drawn between the significant TF pairs. Red edges are known protein-protein interactions. TFs expressed in the cell line are highlighted in green with darker tone indicating higher evidence; known regulators in the corresponding cell type are highlighted as rectangles with red border.

Further, we examined the coherence and the differences between the two predicted regulatory networks. TFs which share the most significant co-occurring partners in both networks are the pluripotency factors SOX/SRY, NANOG, KLF4 and other factors like LEF1/TCF, EGR, MZF1 and SPZ1. Factors, which have large number of co-occurring factors in undifferentiated ESCs but do not occur in the differentiated ESCs were the pluripotency factors OCT4 and FOXD3 (necessary for the maintenance of the pluripotency and self-renewal) and factors LHX3 and DBP. On the other hand, factors with many co-occurring partners in the differentiated ESCs which are not present in the undifferentiated ESCs are factors important for differentiation in endoderm (GATA factors) or mesoderm (HNF4/NR2F) or other developmental factors (TFAP2, E2F1, EGR1 and SRF). The differences between the two networks are summarized in Figure 5.16; each TF in the network is shown with the corresponding number of distinct and shared TF partners in the two networks.

5.5.6 Co-occurring TF pairs in muscle-specific DHSs

The cell-type-specific regulatory networks (without frequently occuring TF pairs in many cell types) in muscle-related cell types such as skeletal myoblast, muscle myoblast, skeletal striated muscle and cardiac myocytes were investigated. The networks in muscle-related cell types consist of 165 (muscle myoblast) - 180 (skeletal myoblast) sinificant TF pairs among roughly 130 TFs. Analog to other cell types mentioned above, at least three quarters of all TFs in the network $(75-78\%)$ are expressed in myoblasts, in myocytes or in differentiated muscle cells. Between $15-20\%$ of TFs in the derived network are known regulators in muscle development. Furthermore, the most relevant regulators of the myogenic lineage MEF2A, MYF, MYOD, MYOG and PAX are present in all muscle-specific regulatory networks.

The regulatory network in skeletal myoblast (see Figure 5.17) consists of a large interconnected network, small network with five nodes and several single TF pairs or triplets. The main network could be divided into two subnetworks connected with a single factor FOXA. The first subnetwork is dominated by two regulators of early muscle differentiation MYOD and MYOG. These factors co-occur in the network with general factors such as TFAP2, HNF4A, ZEB1 and with other myogenic regulators TAL1, TCF, RUNX1 and TBX5 (Bentzinger et al., 2012; Sartorelli and Caretti, 2005). The second subnetwork includes more general regulators of cell and organ development (homeobox proteins POU6F1, POU3F1, ONECUT, CUX1) co-occurring in many cases with MEF2A, which is a myogenic regulator usually activated by MYOD (Sartorelli and Caretti, 2005). It co-occurs with some other factors with known function in muscle development such as GATA and LEF1. The two subnetworks suggest that there might be two different processes regulated by distinct groups of TFs: one of the early development directed by MYOD/MYOG and a second one directed by homeobox proteins and MEF2 which is recruited by MYOD. We observed this phenomenon in the networks of skeletal muscle cells and of muscle myoblasts. One reason for the high frequency of homeobox factors in the muscle network might be their motif similarity to myogenic

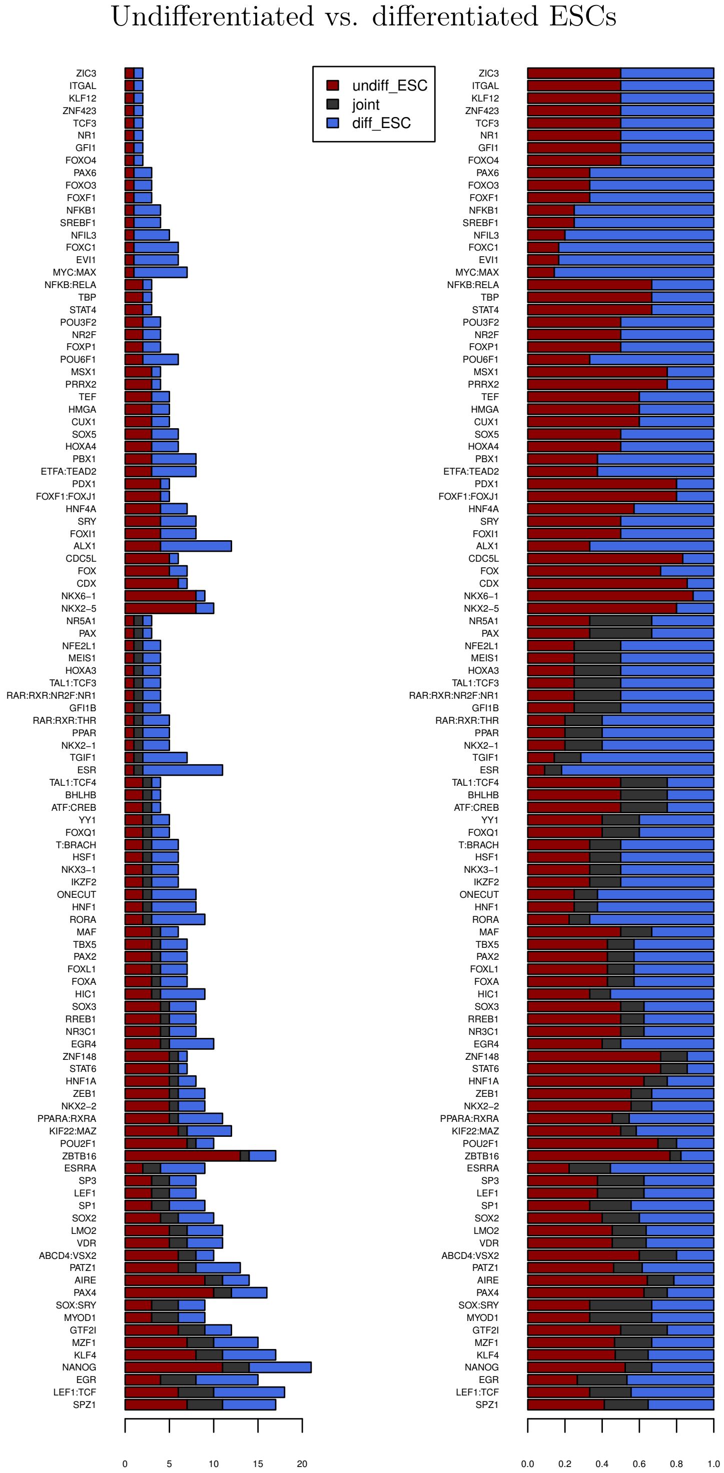


Figure 5.16: Comparison of regulators in undifferentiated and differentiated embryonic stem cells. For each TF, the barplot shows the number of distinct co-occurring partners in undifferentiated ESCs (red) and in differentiated ESCs (blue) and the number of shared co-occurring partners on both cell lines (black). The left column shows the absolute numbers, the right column shows the proportions.

regulator Six1/4 which as well belongs to the homeobox family but was not included in the list of the studied motifs. The network includes TF pairs which are known to be directly interacting proteins: TAL1:LMO2, HNF4A:TP53, ESR:NR2F, CEBP:CREB, ATF:NFE2L2, POU2F1:STAT5A, HSF1:STAT, AR:NR3C1 (Chatraryamontri et al., 2013; Ravasi et al., 2010).

Skeletal myoblasts

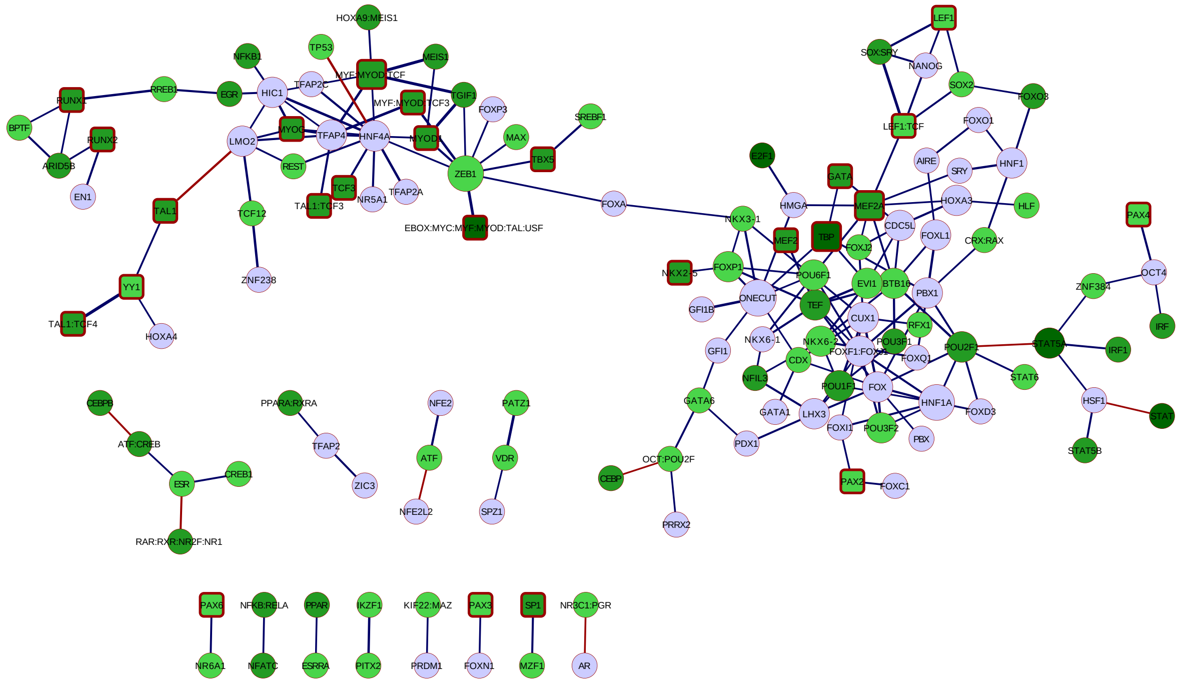


Figure 5.17: Network of significant TF pairs in skeletal myoblasts. Nodes in the network represent transcription factors, edges are drawn between the significant TF pairs. Red edges are known protein-protein interactions. TFs expressed in the cell line are highlighted in green with darker tone indicating higher evidence; known regulators in the corresponding cell type are highlighted as rectangles with red border.

The network in cardiac myocytes (see Figure 5.18) has a different structure than the the network in skeletal myoblasts. It consist of a main network, small subnetwork with 5 nodes and several single TF pairs and triplets. Out of the myogenic regulators, the main network is dominated by MEF2, SRF, NKX2-5 and GATA factors, all of them are main modulators of cardiac transcription network (Schlesinger et al., 2011). Other downstream regulators of the myogensis such as LEF1, TCF, RUNX and YY1 are

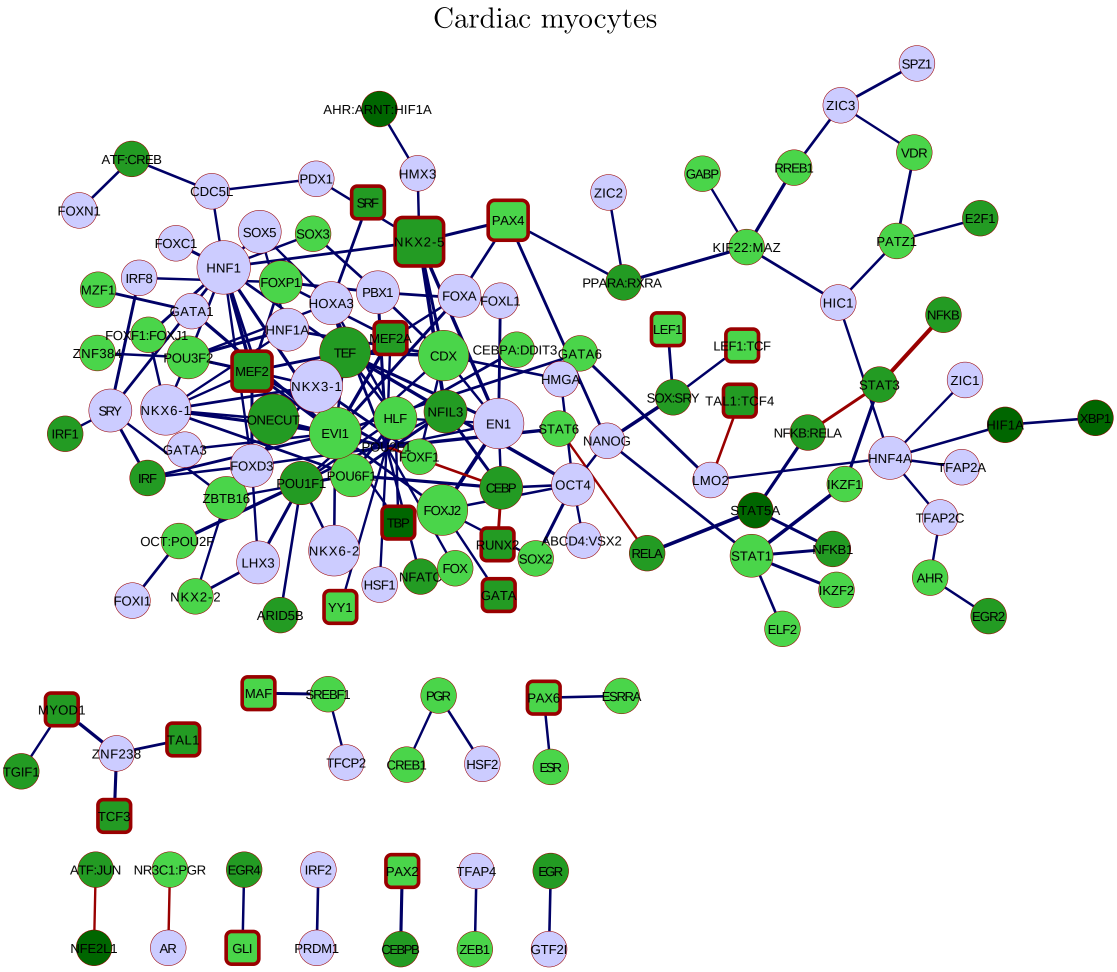


Figure 5.18: Network of significant TF pairs in cardiac myocytes. Nodes in the network represent transcription factors, edges are drawn between the significant TF pairs. Red edges are known protein-protein interactions. TFs expressed in the cell line are highlighted in green with darker tone indicating higher evidence; known regulators in the corresponding cell type are highlighted as rectangles with red border.

present in the network, too. The small subnetwork consists of quartet of co-occurring proteins MYOD1, TAL1, TCF3 and ZNF238. Previous studies showed a co-regulation of MYOD and TCF/LEF of the Wnt proteins, which are important for the formation of groups of muscles myotomes (Dubinska-Magiera et al., 2013). The cardiac myocyte regulatory network includes several known PPIs such as TAL1:LMO2, STAT6:RELA, STAT3:NFKB, CEBP:RUNX, ATF/JUN:NFE2L1 and NR3C1:AR (Chatraryamontri et al., 2013; Ravasi et al., 2010).

When comparing the two regulatory networks of the progenitor myoblasts and the differentiated cardiac myocyte, some interesting facts rise up. Regulators which have many co-occurring partners in myoblasts but do not appear in the cardiac myocyte network are MYOG, MYOD/MYF motif and CUX1. MYOG, MYOD and MYF are the main regulators of the early myogenic differentiation from embryonic progenitors into myoblasts (Sartorelli and Caretti, 2005) and thus, are expected to co-occur with other factors in myoblasts but not to be active in the further differentiated myocytes. On the other hand, factors with many co-occurring partners in myocytes which do not occur in the myoblast network are STAT1, STAT3 and NKX2-2. STAT proteins regulate the expression of genes that are important (among other functions) for differentiation, proliferation, and apoptosis (Levy and Darnell, 2002) and thereby are expected to be more active in differentiated cell lines. Factors, which share the most co-occurring partners in both networks are homeoboxes POU2F1, HNF4, SOX/SRY and EVI, factors involved in the embryonic development and development of organs. The differences between the co-occurring partners for all TFs in both networks are summarized in Figure 5.19.

Surprisingly, the agreement among the significant TF pairs predicted on the muscle-specific DHSs and on the muscle-specific promoters as presented in Chapter 4, Section 4.5.2 is very low. Only 4 significant TF pairs were found with both methods: MEF2A:EVI1, MEF2A:FOX, MEF2A:PAX and TBP:FOX, where only the FOX and PAX family (with similar binding motifs) was in agreement but not the exact members of the families.

5.5.7 Co-occurring TFs in ubiquitous DHSs

As a contrast to cell-type-specific co-occurring transcription factors, we have investigated TF pairs with largest negative L_l scores in at least 20 cell types. These TF pairs tend to co-occur on the ubiquitous DHSs more likely than on the cell-type-specific DHSs. The network of the ubiquitous co-occurring TF pairs is visualized in Figure 5.20. The network is dominated by a large, highly connected subnetwork with main regulators such as SP1, E2F and PAX factors. Furthermore, a small subnetwork of STAT fac-

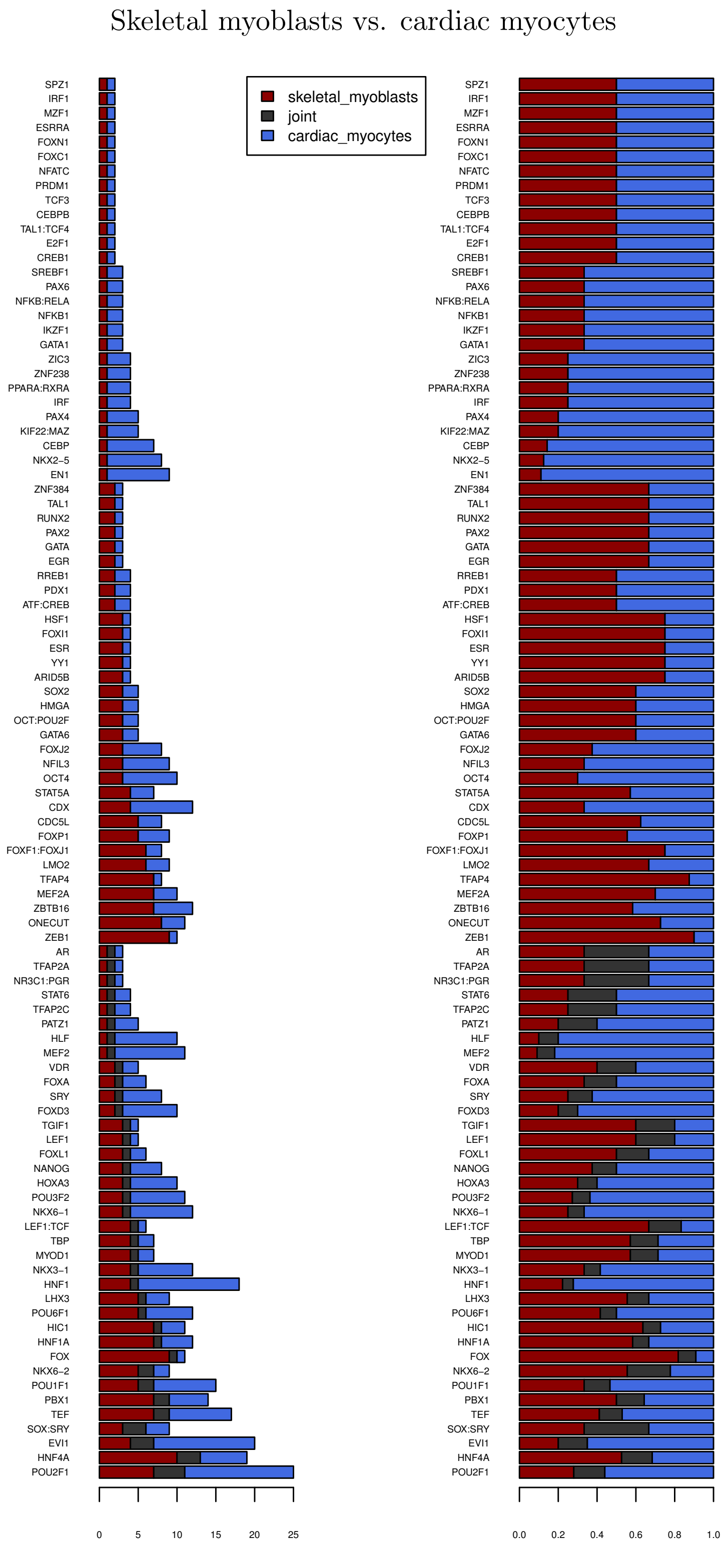


Figure 5.19: Comparison of regulators in skeletal myoblast and cardiac myocyte. For each TF, the barplot shows the number of distinct co-occurring partners in skeletal myoblast (red) and in cardiac myocyte (blue) and the number of shared co-occurring partners on both cell lines (black). The left column shows the absolute numbers, the right column shows the proportions.

tors, GABPA and NKX2-5 factors is connected to the large subnetwork via EP300. A small separated subnetwork with JUN:FOS central hub, a small subnetwork with NFY central hub and few single TF pairs are presented in the network. We could identified 14 factors (ATF, CREB, E2F1, NRF1, NFY, SP1, TBP, STAT factors) which were described as promoter-specific or promoter-centric TFs in previous studies (Neph et al., 2012; Whitfield et al., 2012). This is in agreement with our predictions since the ubiquitous DHSs overlay in large part promoter regions. Further, we investigated those TFs which were predicted as significant co-occurring pairs on general promoters based on diverse rank based association measures, see Section 3.5. There were 27 (33%) factors or factor families which appeared in both of the networks (highlighted as green nodes in Figure 5.20). In addition, both approaches predicted identically 6 co-occurring TF pairs (FOXD:PAX4, HNF1A:NKX2-5, TFAP2A:E2F1, TFAP2A:SP1, PAX4:EN1, SOX5:FOXL1). This corresponds to an odds score of 1.26 and corresponding p-value of Fisher's exact test $p = 0.37$. However, we do not expect a complete agreement between the two methods. The co-occurring TF pairs in Section 3.5 are pairs of TFs with significantly similar ranked list of all promoters. But the predictions on the ubiquitous DHSs are predicted co-occurring TF pairs with a prioritization on ubiquitously open DHSs than on cell-type-specific DHSs.

5.6 Comparison with other computational and experimental methods

To validate the accuracy of our predicted co-occurring TFs on CTS-DHSs and on ubiquitous DHSs, a systematic comparison with a large-scale experimental database of protein-protein interactions (PPIs) and with two other computational studies was conducted. The results of this comparison are discussed in the next paragraphs.

5.6.1 Comparison with a database of protein-protein interactions

One possible verification of our predicted TF pairs is a comparison with experimental validated direct protein-protein interactions (PPIs) between transcription factors. We compared our predicted co-occurring TF pairs with the atlas of TF-TF interactions inferred from mammalian two-hybrid assays (Ravasi et al., 2010) and from other forms of experimental evidence listed in PPI databases (Chatraryamontri et al., 2013).
First, the TFs in both sets have to be mapped to each other to determine the number of possible TF pairs. Thus, only 276 TFs are included in both sets (experimental atlas

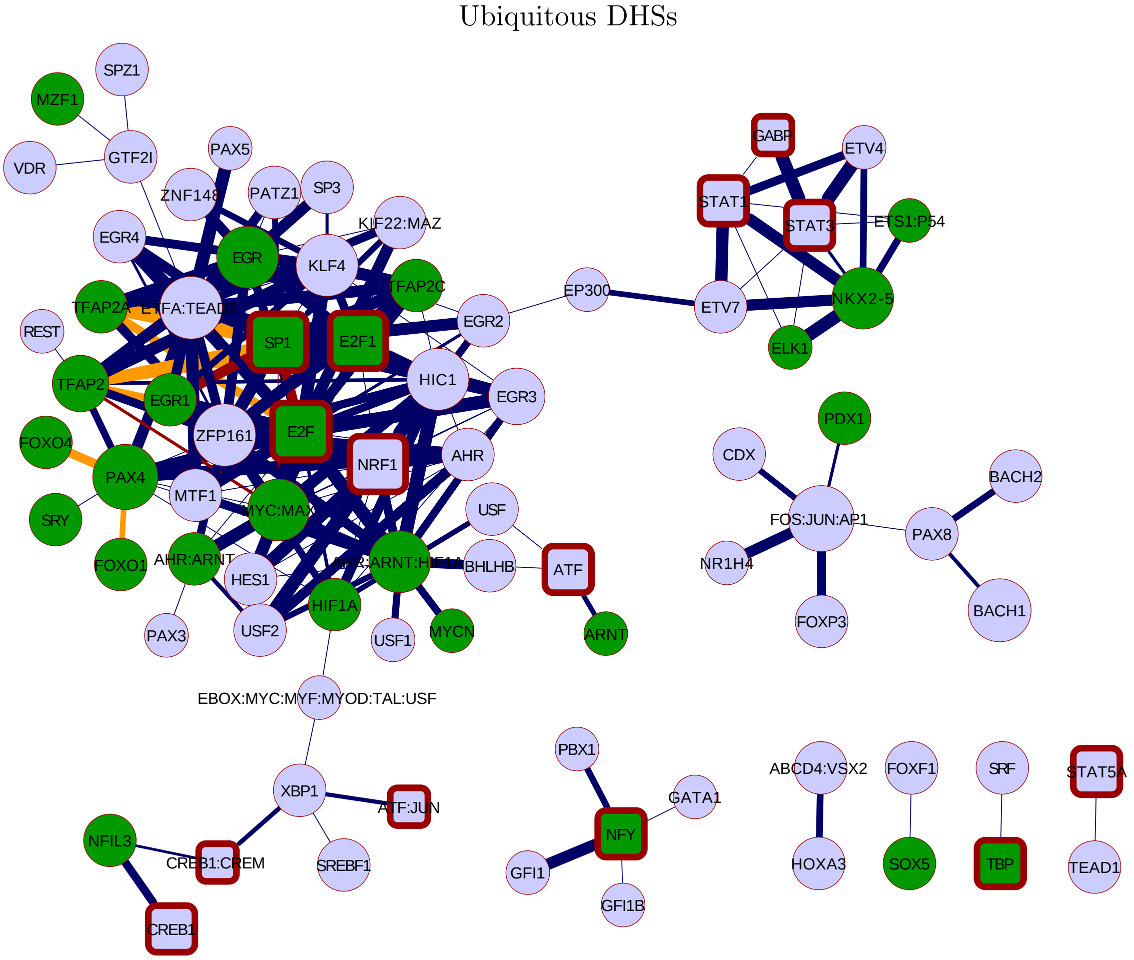


Figure 5.20: Network of significant TF pairs on ubiquitous DHSs. Nodes in the network represent transcription factors, edges are drawn between the significant TF pairs. Red edges are known protein-protein interactions. Known promoter-specific regulators are highlighted as rectangles with red border; green nodes are TFs indicated as co-occurring on promoter sequences in Section 3.5.

and our TF data), resulting in possible number of 38 226 TF pairs. Then, the total list of 5 238 TF-TF interactions in the atlas can be mapped to 1516 TF pairs with TFs in the joint universe. On the other hand, out of 2359 of our predicted co-occurring TF pairs, only 1047 TF pairs are from the joint universe. Then, the comparison results in 83 identical TF pairs (7.9%) predicted with our method and experimentally verified as a PPI. This represents a large enrichment compared to random choice of TF pairs with an odds ratio of 2.1 and a corresponding p-value of Fisher's exact test of $p = 3.03 \times 10^{-9}$. Over all studied cell types, the highest proportion of direct TF-TF interactions was found in blood microvascular endothelial cells (15.3%), mammary fibroblasts (14.9%) and cardiac fibroblasts (14.7%). The smallest overlap was found in primary T-cells, mesenchymal stem cells and brain vascular cells (3.7%, 4.4% and 5.3%, respectively).

In a similar fashion, the comparison of the predicted co-occurring TF pairs in the ubiquitous DHSs was conducted. Out of 364 significant TF pairs on the ubiquitous DHSs, 35 (9.6%) pairs were found to be interacting proteins as well. The significance of this overlap can be assessed again with Fisher's exact test and the corresponding p-value is $1.8 \times 10^{-}6$, with odds ratio of 2.6.

When comparing with the experimentally derived database of direct PPIs, it is important to consider the sensitivity (true positive rate) and the false discovery rate (FDR) of the experimental method. The mammalian two-hybrid-assay has a very low estimated sensitivity of 25% and a limited FDR of 53% (Ravasi et al., 2010). For this reason, we cannot expect, even for a perfectly predicted set of TF interactions, that all of the TF pairs would be included in the experimentally derived atlas. Similarly, even for a predicted set with all true TF interactions, we cannot expect that the whole PPI database from the atlas will be included, due to the false positives in the database. Further, there are fundamental differences between our statistical method and the experimental techniques for detecting PPIs. Whereas our method investigates predicted binding affinities of TFs on regulatory regions derived from living cells across multiple human cell types, the experimental approaches such as two-hybrid assay test the interaction between artificially expressed proteins in yeast or mammalian cell. Therewith, the two-hybrid assays measure a general ability of two proteins to form a complex - independent of binding to the DNA or not. In contrast, our method focuses only at the cooperative binding of two TFs on the DNA.

5.6.2 Comparison with a study based on ChIP-seq experiments

The largest available study of experimental mapping of TF binding regions in human cell lines with the chromatin immunoprecipitation technique coupled with high-throughput sequencing (ChIP-seq) was generated by The ENCODE Project consortium (2012). In an accompanied study, Wang et al. (2012) analyzed all 457 ChIP-seq data sets on 87 sequence-specific human TFs across 72 cell lines to determine co-binding between different TFs. Specifically, Wang et al. analyzed the canonical motifs as well as secondary motifs (possibly corresponding to a partner TF) found in the ChIP-seq peaks of each TF. In this fashion, they identified total of 155 co-binding TF pairs between 69 TF motifs. Out of these 69 motifs, only 50 map to the TF motifs used for our analysis. This restricts the number of comparable pairs from Wang et al. to only 94 co-binding TF motif pairs and the number of TF pairs with significant L_l score to only 67 TF pairs. Among them, 10 TF pairs are found with both methods, the corresponding odds ratio is 2.3 and the p-value of Fisher's exact test p=0.02. The TF pairs found with both of the approaches (with the cell type where the TF pair was significant) are: MYC:YY1 (fibroblast and epithelial cells), TBP:YY1 (esophaegal epithelial cells), STAT1:CEBPB (fibroblast), HNF4:ESRRA (HeLa cells, skeletal muscle), HNF4A:TCF12 (muscle myoblast, microvascular endothelial dermal cells), HNF4A:SP1 (fibroblast), IRF4:PAX5 (B-lymphocyte, T-cell, fibroblast, astrocytes), IRF4:MEF2A (primary T-cell), TAL1:STAT3 (fibroblast) and RXRA:TCF7L2 (differentiated ESCs). Among them, MYC:YY1, HNF4:ESRRA and STAT1:CEBPB are known PPIs (Chatraryamontri et al., 2013; Ravasi et al., 2010).

In a similar way, we compared the results based on the ChIP-seq data and TF pairs with large significant negative L_l score, i.e. TF pairs enriched on the ubiquitous sequences. We found 6 TF pairs which were in both of the lists (odds ratio 3.1, p-value = 0.02), namely E2F4:NRF1, JUN:FOXA, HNF4A:TCF12, HNF4A:ESRRA, RXRA:TCF7L2, TBP:YY1. The latter 4 are TF pairs significant on some cell-type-specific DHSs as well as on the ubiquitous DHSs which suggest a general rather than a cell-type-specific co-occurrence of these TF pairs.

The number of sequence-specific TFs which have known DNA-binding motifs and for which large scale data is available, is very limited. For our comparison, we narrow the universe of all possible TF pairs from more than 46 000 (with 306 unique TFs in our data) to only 1225 possible TF pairs (with 50 TFs for which the TF motif and ChIP-seq data is available). This restriction might be one of the reason for the relatively small but still significant overlap of both predictions, since it focuses only on a subset of analyzed TFs.

5.6.3 Comparison with statistical prediction of TF-TF dimers

Jankowski et al. (2013) presents a computational method for predicting cooperative cell-type-specific dimerization of TFs on the DNA. They have studied the occurrence of more than 450 000 TF motif pairs in cell-type-specific DHSs in 78 cell types and investigate the cell-type-specific orientation and offset (with maximal spacing of 50bp, allowing partial overlaps) of the two motifs. The significance of the enrichment of the specific offset is evaluated with the binomial distribution and corrected for multiple testing with Bonferroni correction. Thus, after conducting 1.4×10^9 statistical tests, Jankowski et al. derived 5233 significantly overrepresented TF motif complexes corresponding to 603 significant distinct TF dimer pairs (after removing multiple motifs corresponding to a single TF).

In total, the agreement of the predicted co-occurring TF motifs with our method and with the predicted TF-TF dimers by Jankowski et al. is very low, only 90 (1.7%, Fisher's exact test p-value $=1$) out of more than 5257 TF motif pairs coincide in both methods. Nevertheless, 4 of the top-10 predicted motif-complexes from Jankowski et al. (E-box dimer, OCT-SOX heterodimer, IRF homotypic dimer and EBF1 homodimer), which were additionally found in other independent studies, were included in our predictions too. Surprisingly, the cell types where these motif-complexes were significant using our method were not exactly in agreement with the cell types from Jankowski et al. Namely, the E-box dimer of YY1 and MEIS1 was found with significant spacing in retinoblastoma by Jankowski et al., but in various cell types such as lung fibroblast, muscle myoblast and amniotic epithelial cells with our method. This dimer was confirmed with an *in vitro* study by De Masi et al. (2011). The IRF homotypic dimer E2F1/IRF7 and NFAT3 was found in B-lymphocytes by Jankowski et al. and in various fibroblast and in hematopoietic progenitor cells with our method. An independent study of Tanaka et al. (1993) identified this dimer in a non-cell-type-specific experiment. The EBF1 homodimer IKZF and STAT1 had a significant offset in neuroblastoma found by Jankowski et al. and a significant L_l score in astrocytes, mesenchymal stem cells and in ligament fibroblast. Treiber et al. (2010) found this dimer in an independent experiment in B-lymphocytes. The well-known OCT-SOX heterodimer in embryonic stem cells (Chen et al., 2008b) was found with our method and by Jankowski et al. also in embryonic stem cells.

Over the different cell types, the highest agreement between the two methods was found for co-occurring TF motifs in dermal fibroblast with 22 matches, renal glomerular endothelial cells and mesenchymal stem cells, both with 20 matches. The smallest overlap was found in the primary cells such as primary T-cell, hematopoietic progenitor cells, ESCs and muscle myoblast with less than 5 matching TF motif pairs.

Similarly, when compared to the co-occurring TF motifs with a large negative L_l score, predicted on the ubiquitous DHSs, the agreement with the predictions from Jankowski et al. is 34 (1.4%, Fisher's exact test $p = 1$) out of 2379 significant TF motif pairs. Among them, there is one of the top-10 predicted motif-complexes by Jankowski et al. FOXA1:AR dimer, whose enhancer function in prostate adenocarcinoma was shown independently by Wang et al. (2011).
The reason for the relatively small concordance of predicted co-occurring TF motifs and predicted TF dimers from Jankowski et al. might be the different objectives of both prediction methods. Whereas our method focus on TFs which co-occur in cell-type-specific manner on small genomic regulatory regions, Jankowski et al. predicts directly interacting TFs which bind as a dimer on the regulatory DNA with a fixed spacing. Further, method of Jankowski et al. is sensitive to TF-dimers with widespread binding whereas our method focus on the top 1000 bound cell-type-specific regions for each TF. There might be further differences in the definition of the cell-type-specific regulatory regions defined by the DNase-seq reads. Although we used partially the same data provided by ENCODE, the definition of the CTS-DHSs by Jankowski et al. uses larger genomic regions (400bp) which were detected with the F-seq peak-calling algorithm (Boyle et al., 2008b). Then, the cell-type-specific regions are defined as regions in the corresponding cell-types which do not overlap with other cell types. In contrast, our algorithm including calculation of the t-statistic (see Section 5.2) takes into account the variability between replicates of one cell types and enables to create lists of DHSs ranked by their cell-type specificity.

5.7 Conclusion

Transcription in eukaryotic cells is regulated depending on chromatin, where the genomic DNA is wrapped around nucleosomes. The accessibility of the *cis*-regulatory DNA sequences for transcription factor binding depends on the temporal and spatial development of the cell. Early on discoveries showed that the accessible genomic regions were hypersensitive to cleavage by an enzyme DNase I and that these regions might contain regulatory sequences (Weintraub and Groudine, 1976). The DNase I technique combined with high-throughput genomic sequencing (DNase-seq) allowed systematic mapping of nucleosomes and accessible regions of the packed DNA. Recent studies (Boyle et al., 2008a; John et al., 2011; Li et al., 2011b) showed that the TF binding sites are preferentially located in those accessible regions which are hypersensitive to

DNase I and which are therefore called DNase-hypersensitive sites.

In this chapter, we used DNase-seq data in 90 human cell types from a recent large study (The ENCODE Project consortium, 2012) to derive **cell-type-specific DNase hypersensitive sites (CTS-DHSs)** and **ubiquitous DHSs** which were quantified with a t-statistic taking into account within-cell-type variation of the DNase hypersensitivity. We showed that biologically related samples have high correlations of read counts in genomic windows, suggesting a good quality and reproducibility of underlying samples. The derived CTS-DHSs localize primarily in distant regions from the TSS (mainly in introns and intergenic regions) compared to the ubiquitous DHSs which occur mainly in promoters.

Next, we derived a representation of a **TF** as a **ranked list of CTS-DHSs and ubiquitous DHSs** ordered by the binding affinity. For each cell type, we took a fixed number of CTS-DHSs and of ubiquitous DHSs and rank them by the binding affinity of the particular TF.

Next, we studied the **overrepresented TF binding motifs in the CTS-DHSs** and identified a large group of motifs present in the majority of cell types, such as homeobox proteins (NKX6-2, POU2F1), nuclear factors (NFY, NRF1), general TFs (SP1, GABP, TFAP2), ETS and E2F family members. These factors are known to regulate many important genes involved in cell and organ development, energy metabolism and cell cycle.

Notably, factors showing cell-type-specific occupancy patterns are particularly known cell-selective transcriptional regulators including (1) SPI1, IRF, GATA, STAT and ATF family members in white blood cells; (2) MEF2A, MYOD and MYOG in the myoblast cells; (3) pluripotency factors OCT4, SOX2 and NANOG in the embryonic stem cells and (4) ETS-family members and GATA6 in lung fibroblasts.

Further, we developed a novel statistical method to detect pairs of **co-occurring TFs in a cell-type-specific manner**. Our approach compares the similarity of two TFs (represented as a ranked list of DHSs) in the CTS-DHSs and in ubiquitous DHSs. Therewith, TF pairs of interest share significant number of CTS-DHSs but non-significant number of ubiquitous DHSs. Thus, the predicted TF pairs with the highest score co-occur more likely in the cell-type-specific sites than in the generally open sites.

In total, we predicted 2359 **significant unique TF pairs in 64 cell types**. Among the predicted TF pairs, 158 are significant in the majority of cell types, suggesting that these TF pairs occur together much likely in distal regulatory regions than on ubiquitously open regions. Further, we detected 739 TF pairs which share significant number of the ubiquitous DHSs compared to the CTS-DHSs over all cell types. The co-occurring

TFs in ubiquitous DHSs involve many promoter-centric and general TFs and are in a significant agreement with our predictions when comparing just two ranked lists as presented in Chapter 3.

Further, we derived **cell-type specific regulatory networks** from the predicted TF pairs in each particular cell type. In general, more than 75% of factors in the networks are **expressed in the corresponding cell type** as measured by an independent study using RNA-seq experiments (Flicek et al., 2014). Moreover, roughly one quarter of TFs in the cell-type-specific networks are **known regulators** in the particular cell type.

To gain further knowledge about the relevance of our predictions we studied in detail the regulatory networks in white blood cells, embryonic stem cells and muscle cells. The main regulators with most co-occurring partners in the **hematopoietic progenitor cells** are hematopoietic factors EVI1 and GATA and early development factors POU6F1 and ONECUT. By comparison, the main regulators in differentiated **white blood cells** such as monocytes are immune cell regulators POU2F1, CDX, LHX3 and PBX1 and other factors such as ALX1, NKX3-1, NKX6-1 and TEF. Further, majority of the known regulators of monocyte differentiation (SPI1, CEBP, IRF, VDR, STAT1 and STAT3) are present in the predicted regulatory network in monocytes. The networks includes several known direct protein interactions: NFKB/RELA:STAT3, GATA:MEF2A, GATA:POU2F1, CEBP:CREB, CEBP:STAT5A, CEBP:POU2F1, MAF:ETS, ESR:NR2F1.

Next, the transcriptional networks in (undifferentiated and differentiated) **embryonic stem cells** were investigated. The main regulators of undifferentiated ESCs are known pluripotent factors OCT4 and NANOG, but other early developmental regulators such as POU2F1, SOX/SRY, STAT3, FOXD3, KLF4, ESRR, MYC, TCF3, ZIC and YY1 are present in the network, too. The well known co-occurrence of pluripotent factors (OCT4:NANOG, OCT4:SOX2, NANOG:SOX) was found with our method too. In addition, several other known TF-TF interactions were found with our method: NANOG:TCF/LIF1, POU2F1:TLX2, CEBP:ATF/CREB, CEBP:OCT, MAF:ETS, STAT3:NFKB, TAL1:LMO2, LMO2:MYC. The regulatory network in differentiated ESCs includes more cell-selective transcriptional regulators such as LEF1/TCF, GATA4, EGR and TFAP2. Four smaller subnetworks of specific functions were identified in the network in differentiated ESCs: (1) pluripotent subnetwork with SOX, NANOG, FOXC1; (2) early blood development subnetwork dominated with GATA proteins together with EVI1, IRF and SRF; (3) ESR dominated subnetwork together with mesoderm differentiation regulators HNF4 and NR2F; (4) MYC-centric subnetwork with E2F1, KLF4, TFAP2, SP1 and zinc finger proteins. The co-occurrence of MYC with

E2F1 and ZFX (or other zinc finger proteins) was observed in other study analyzing ChIP-seq experiments in ESCs (Chen et al., 2008b).

Regulatory networks in myoblasts and differentiated **muscle cells** include the main myogenic regulators MYOG, MYOD, MYF, PAX, MEF2, NKX2-5, SRF and GATA. However, the regulators of early myogenic differentiation MYOG, MYOD and MYF were present with high number of co-occurring partners only in the progenitor myoblasts but not in the differentiated cardiac myocytes. Further, the regulatory networks in skeletal muscle cell types show 2 distinct regulatory groups of TFs: first one dominated by myogenic factors MYOD and MYOG and the second one is dominated by the myocyte factor MEF2.

The concordance of the predictions of co-occurring TFs in tissue-specific promoters presented in Chapter 4 and in the cell-type-specific (or tissue-specific) DHSs is relatively small. One reason for the diverse results might be the difference in the genomic regions which were studied. As shown in Figure 5.3, CTS-DHSs are preferentially located in intronic and intergenic regions (more than 80% of them) and therewith correspond mainly to enhancers. Less than 10% of the CTS-DHSs overlay with promoters, thus the DNA sequences on the CTS-DHSs are different from the promoter sequences analyzed in Chapter 4. Further, it is more likely that different combinations of TFs regulate the transcription of their target genes on promoters and on enhancers. In addition, predicted regulatory networks in promoters are dominated with few central hubs which co-occur with many other factors. In contrary, TFs in regulatory networks in the CTS-DHSs are usually more interconnected and include larger number of main regulators.

For a systematic validation of our results, the predicted co-occurring TF pairs were compared against a large-scale experimental databases of PPIs (Chatraryamontri et al., 2013; Ravasi et al., 2010), against predictions derived from an analysis of ChIP-seq experiments (Wang et al., 2012) and against a statistical prediction of TF-dimerization (Jankowski et al., 2013). Although the experiment-derived relationships between TFs provide measurements of an even stronger cooperation of two TFs (such as direct PPI or co-binding proteins measured by ChIP-seq experiment) than only co-occurrence on the regulatory DNA (as the aim of our prediction), these **experimental-derived TF pairs** are significantly **enriched** in our predicted set of the co-occurring TFs. The enrichment of the direct PPIs in the set of cell-type-specific significant TF pairs is 2.1 fold (corresponding Fisher's p-value= 3.3×10^{-9}) and in the set of ubiquitous-specific co-occurring TFs 2.6 ($p = 1.8 \times 10^{-6}$). The overlap of studied TFs with the ChIP-seq derived co-binding TFs is very small due to the limitation of the experimental technique. However, we found a significant agreement among the predictions on the CTS-DHSs with total of 10 TF pairs (2.3 fold enrichment, $p = 0.02$) and on the ubiquitous DHSs with 6 TF pairs (3.1 fold enrichment, $p = 0.02$). The agreement between the predicted TF-TF dimers and our set of co-occurring TFs is very low, less than 2% of our predicted co-occurring TF pairs was found as TF-TF dimer by Jankowski et al.. Nevertheless, out of the 10 top-scored predicted TF complexes by Jankowski et al. with further evidence in literature, 5 of them were found in our set of predicted co-occurring TF pairs. 3 of the 10 top-scored predicted TF dimers from Jankowski et al. are homodimers, e.g. protein complexes of two identical factors, which are impossible to be predicted with our approach, leaving only 2 of the top 10 predicted TF dimers undetected with our methods. With these findings, we could support the accuracy of our results.

Overall, the validation of our predicted TF pairs and further analysis of cell-type-specific networks show that our predictions include a significant proportion of independently validated co-occurring (or directly interacting) TFs. Moreover, the large majority of the regulators appearing in the cell-type-specific networks are actually expressed in the corresponding cell type and roughly one quarter of them have known function in the related cell type. These findings suggest that our novel **predictions of co-occurring TFs** have **functional relevance**. Further, our results indicate that the cell-type-specific enhancers contain large number of TF motifs. One part of the TFs enriched on the enhancers has more general function of cell development and metabolism. The other part of TFs is highly cell-type-specific and might influence the differentiation of the corresponding cell type.

6 Summary

One of the key questions in molecular biology is how cells with the same genetic code are able to differentiate into a large variety of cell types. The differentiation of the cell is controlled through the regulation of gene expression - a cellular mechanism which activates only specific part of the genetic information in a way that only specific gene products are generated. The main factors of the gene regulatory mechanisms are the cellular environment, accessibility of the DNA and specific proteins called transcription factors (TFs) (Coller and Kruglyak, 2008). TFs bind with sequence preferences to regulatory regions in the DNA to control the expression of their target genes. They usually do not act alone but in a combinatorial manner thus regulating cell-type-specific gene expression. This combinatorial cooperation of TFs is critical for the achievement of the cell type specificity and of the developmental level of the cell. But, the experimental techniques that are able to detect the combinatorial cooperation of TFs on the DNA are very limited and are usually able to measure only few proteins at once.

The aim of this thesis was to use estimated information about the binding affinity of **transcription factors** to the DNA to predict their **co-occurrence** in the genomic regulatory regions. Specifically, the transcription factors were represented by ranked lists of their target genes (or other regulatory regions) and then several rank based statistics were applied to detect significant associations between pairs of TFs.

The most common **rank based association measures** were introduced in Chapter 2 together with their application on a simple example of two ranked lists. In molecular biology, experimental techniques are able to measure thousands of items in a single experiment. Thus the researcher's interest focuses mainly on the top-scored measurements. Owing to this, in the statistical analysis the attention is also restricted to partial ranked lists, i.e. ranked lists of top-ranked items only. For this reason, the behavior of the rank based association measures for comparison of two partial ranked lists was studied in Chapter 2. All of the discussed rank based association measures could successfully analyze partial lists, though some of them (e.g. Spearman's and Kendall's correlation) needed to be modified to be able to deal with incomplete ranked lists.

In Chapter 3, we first defined the representation of **TFs** by **ranked lists** of their target genes as a list of promoter sequences ordered by the estimated binding affinity of the particular TF. Then we applied all five rank based association measures to detect significantly **associated pairs of TFs**, i.e. pairs of highly associated ranked lists. When corrected for similarity of binding motifs, the concordance of the predictions based on four association measures (Spearman's correlation, Kendall's τ, R2KS score and Fisher's exact test) was very large. The regulatory network derived from the predicted associated TF pairs included several known promoter-specific regulators and several known protein-protein interactions between predicted co-occurring TFs.

Of much larger interest is the **tissue-specific** or **cell type-specific** co-occurrence of TFs. The key question is how different combinations of factors co-occur in different tissues or cell types. Moreover, it is still unknown whether there are some general factors playing an important role in many tissues which specifically change only their cooperating partners. Further, it is of interest whether there are different main regulators in each cell type or tissue. We seeked to answer these questions with predicted co-occurring TFs in tissue-specific promoters in Chapter 4 and in cell type-specific DNase hypersensitive sites (CTS-DHSs) in Chapter 5.

Including additional information about the tissue (or cell type) specificity of the corresponding genomic regions led to introducing a third dimension (or third ranked lists) for the association measure. Because of an easier extension of conventional 2×2-contingency tables (and therewith of the corresponding statistical test) to three dimensions we prefered to use the contingency tables for prediction of co-occurring TFs in tissue-specific manner.

In Chapter 4, we translated the problem of the association of two TFs (represented by ranked lists of their target genes) in tissue-specific promoters into **3-way contingency table**. Then, the significance of the association of the two TFs could be assessed with the corresponding statical test for the three-dimensional contingency table. However, the choice of the correct null model in the table is essential for the obtained results. Since there is no general rule how to choose the underlying null model in the analysis of the TF co-occurrence we developed a new strategy to select the most appropriate model. We first derived 3-way contingency tables for all TF pairs, fited all possible null models to each table and calculated the corresponding test statistic and its p-value. Then, we studied the obtained distributions of p-values for various null models. We selected the model with p-value distribution closest to the uniform distribution with moderate enrichment of significant p-values that should correspond to a real biological signal. The underlying null hypothesis of the selected model was the **partial independence** of a joint variable selecting to the top-ranked promoters of both TFs and of the variable assigning the tissue-specificity to each promoter.

In Chapter 5, we make use of newly available experimental results of the DNA accessibility over many cell types using the DNase-seq technique. With this data set we were able to define tens of thousands of DNase-hypersensitive sites (DHSs) which are specifically open in the particular cell type or which are ubiquitously open in all cell types. To derive such **cell-type-specific DHSs (CTS-DHSs)** and **ubiquitous DHSs**, an approach from Love and Chung (2012) was adapted. With this new information, we could represent each TF as a ranked lists of the DHSs sorted by the binding affinity. However, if all CTS-DHSs over all cell types of interest are included the length of the ranked lists and therewith the universe of the corresponding contingency table becomes extremely large. For this reason, we developed a new method for comparing two TFs represented by a ranked lists of DHSs. Now, we defined a new representation of the TFs for each cell type of interest such that we ranked the same number of CTS-DHSs and of ubiquitous DHSs by the binding affinity. For example, in liver tissue we created the ranked list for a given TF as an ordered list (by binding affinity) of liver-specific and of ubiquitous DHSs and in lung tissue as an ordered list of lung-specific and of ubiquitous DHSs. Then, we derived two different 2×2 contingency tables for each TF pair in each cell type, namely one for the CTS-DHSs and one for the ubiquitous DHSs. The significance of these two tables was compared with a **log ratio** of its **p-values**. Thus, TF pairs with a large log ratio were strongly associated in the CTS-DHSs but not associated in the ubiquitous DHSs. With this approach we ensured that the predicted associated **TF pairs co-occur** in a **cell-type-specific manner**.

With both methods, we were able to predict a large number of **co-occurring TF pairs** in various **human tissues and cell types**. We detected several central cell-type-specific regulators in the studied cell types and a large group of TFs which are active in many cell types with some stable co-occurring partners and with some cell-type-specific partners. We could show that known protein-protein interactions are enriched in the set of predicted co-occurring TF pairs and that they are in significant agreement with other computational studies. In addition, we found that the majority of the predicted TFs is **expressed in the corresponding tissue or cell type**. Moreover, we could identify roughly one third of the predicted TFs to have a **known regulatory function** in the related tissue or cell type. Of note, all predicted TFs were selected just by the significant test statistics, without any knowledge about their functions in the cell type of interest. Thus, these results indicate that our predicted co-occurring TF pairs and therewith the related regulatory networks are very likely functional in the corresponding cell types.

Our method might be applied in future studies for prediction of cell-type-specific cooperation of other regulatory factors or other important players of gene regulation, such as microRNAs, long non-coding RNAs an others. In principle, our method requires only that these factors can be represented by a ranked list of genomic regions or other informative elements. Further, we hope, that the rapid development of experimental techniques will produce reliable data of co-occurring TFs in cell-type-specific manner which would enable the statistical validation of our predictions.

Bibliography

R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nat Biotech*, 24(5):537–544, 2006.

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Hoboken, New Jersey, third edition, 2013.

B. E. Aronson, K. A. Stapleton, and S. D. Krasinski. Role of GATA factors in development, differentiation, and homeostasis of the small intestinal epithelium. *American journal of physiology. Gastrointestinal and liver physiology*, 306(6), 2014.

G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935): 1720–1723, 2009.

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

C. F. Bentzinger, Y. X. Wang, and M. A. Rudnicki. Building muscle: Molecular regulation of myogenesis. *Cold Spring Harbor Perspectives in Biology*, 4(2):a008342+, 2012.

O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins. *Journal of Molecular Biology*, 193(4):723–743, 1987.

T. Berggård, S. Linse, and P. James. Methods for the detection and analysis of protein-protein interactions. *Proteomics*, 7(16):2833–2842, 2007.

B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson. The NIH roadmap epigenomics mapping consortium. *Nat Biotech*, 28(10):1045–1048, 2010.

A. Bieller, B. Pasche, S. Frank, B. Gläser, J. Kunz, K. Witt, and B. Zoll. Isolation and Characterization of the Human Forkhead Gene FOXQ1. *Cell Biology*, 20(9):555–561, 2001.

Y. M. Bishop. Full contingency tables, logits and split contingency tables. *Biometrics*, 25:383–399, 1969.

P. Blyszczuk, J. Czyz, G. Kania, M. Wagner, U. Roll, L. St-Onge, and A. M. Wobus. Expression of Pax4 in embryonic stem cells promotes differentiation of nestin-positive progenitor and insulin-producing cells. *Proceedings of the National Academy of Sciences*, 100(3):998–1003, 2003.

A.-L. Boulesteix and M. Slawski. Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10 (5):556–568, 2009.

A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, 2008a.

A. P. Boyle, J. Guinney, G. E. Crawford, and T. S. Furey. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24(21):2537–2538, 2008b.

J. C. Bryne, E. Valen, M.-H. E. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research*, 36(suppl 1):D102–D106, 2008.

M. L. Bulyk. DNA microarray technologies for measuring protein-DNA interactions. *Current Opinion in Biotechnology*, 17(4):422–430, 2006.

L.-W. W. Chang, R. Nagarajan, J. A. Magee, J. Milbrandt, and G. D. Stormo. A systematic model to predict transcriptional regulatory mechanisms based on overrepresentation of transcription factor binding profiles. *Genome research*, 16(3):405–413, 2006.

A. Chatraryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas, L. O'Donnell, T. Reguly, A. Breitkreutz, A. Sellam, D. Chen, C. Chang, J. Rust, M. Livstone, R. Oughtred, K. Dolinski, and M. Tyers. The biogrid interaction database: 2013 update. *Nucleic Acids Res.*, 41(D1):D816–D823, 2013.

F. E. Chen, D.-B. Huang, Y.-Q. Chen, and G. Ghosh. Crystal structure of p50/p65 heterodimer of transcription factor NF-[kappa]B bound to DNA. *Nature*, 391(6665):410–413, 1998.

X. Chen, V. B. Vega, and H. H. Ng. Transcriptional Regulatory Networks in Embryonic Stem Cells. *Cold Spring Harbor Symposia on Quantitative Biology*, 73:203–209, 2008a.

X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y.-H. H. Loh, H. C. C. Yeo, Z. X. X. Yeo, V. Narang, K. R. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W.-K. K. Sung, N. D. Clarke, C.-L. L. Wei, and H.-H. H. Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, 2008b.

G. Chua, Q. D. Morris, R. Sopko, M. D. Robinson, O. Ryan, E. T. Chan, B. J. Frey, B. J. Andrews, C. Boone, and T. R. Hughes. Identifying transcription factor functions and targets by phenotypic activation. *Proceedings of the National Academy of Sciences*, 103(32):12045–12050, 2006.

H. A. Coller and L. Kruglyak. It's the sequence, stupid! *Science*, 322(5900):380–381, 2008.

G. E. Crawford, S. Davis, P. C. Scacheri, G. Renaud, M. J. Halawi, M. R. Erdos, R. Green, P. S. Meltzer, T. G. Wolfsberg, and F. S. Collins. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Meth*, 3(7):503–509, 2006a.

G. E. Crawford, I. E. Holt, J. Whittle, B. D. Webb, D. Tai, S. Davis, E. H. Margulies, Y. Chen, J. A. Bernat, D. Ginsburg, D. Zhou, S. Luo, T. J. Vasicek, M. J. Daly, T. G. Wolfsberg, and F. S. Collins. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, 16(1):123–131, 2006b.

F. De Masi, C. A. Grove, A. Vedenko, A. Alibés, S. S. Gisselbrecht, L. Serrano, M. L. Bulyk, and A. J. M. Walhout. Using a structural and logics systems approach to infer bHLH-DNA binding specificity determinants. *Nucleic Acids Research*, 39(11):4553–4563, 2011.

R. P. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen, and R. Etzioni. Combining results of microarray experiments: a rank aggregation approach. *Statistical applications in genetics and molecular biology*, 5, 2006.

W. Deming and F. Stephan. Ontables squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11:427–444, 1940.

L. C. Doré, T. M. Chlon, C. D. Brown, K. P. White, and J. D. Crispino. Chromatin occupancy analysis reveals genome-wide GATA factor switching during hematopoiesis. *Blood*, 119(16):3724–3733, 2012.

M. Dubinska-Magiera, M. Zaremba-Czogalla, and R. Rzepecki. Muscle development, regeneration and laminopathies: how lamins or lamina-associated proteins can contribute to muscle development, regeneration and disease. *Cellular and molecular life sciences : CMLS*, 70(15):2713–2741, 2013.

E. Eden, D. Lipson, S. Yogev, and Z. Yakhini. Discovering Motifs in Ranked Lists of DNA Sequences. *PLoS Comput Biol*, 3(3):e39, 2007.

G. B. Ehret, P. Reichenbach, U. Schindler, C. M. Horvath, S. Fritz, M. Nabholz, and P. Bucher. DNA binding specificity of different STAT proteins. *Journal of Biological Chemistry*, 276(9):6675–6688, 2001.

P. Eriksson and O. Wrange. Protein-protein contacts in the glucocorticoid receptor homodimer influence its DNA binding properties. *Journal of Biological Chemistry*, 265(6):3535–3542, 1990.

J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

R. Fagin, R. Kumar, and D. Sivakumar. Comparing top *k* lists. *SIAM Journal of Discrete Mathematics*, 17 (1):134–160, 2003.

S. Fields and O.-K. Song. A novel genetic system to detect proteinñprotein interactions. *Nature*, 340(6230): 245–246, 1989.

R. Fisher. *The Design of Experiments.* Edinburgh: Oliver & Boyd, 8th, 1966 edition, 1935.

P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. GirÛn, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino, and S. M. Searle. Ensembl 2014. *Nucleic Acids Res.*, 42(D1):D749–D755, 2014.

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

D. J. Galas and A. Schmitz. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research*, 5(9):3157–3170, 1978.

G. U. Gangenahalli, P. Gupta, D. Saluja, Y. K. Verma, V. Kishore, R. Chandra, R. Sharma, and T. Ravindranath. Stem Cell Fate Specification: Role of Master Regulatory Switch Transcription Factor PU.1 in Differential Hematopoiesis. *Stem Cells and Development*, 14(2):140–152, 2005.

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 Research*, 9(13):3047–3060, 1981.

M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O/'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, and M. Snyder. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012.

J. N. Glover and S. C. Harrison. Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature*, 373(6511):257–261, 1995.

S. Gupta, M. Vingron, and S. A. Haas. T-STAG: resource and web-interface for tissue-specific transcripts and genes. *Nucleic acids research*, 33(Web Server issue), 2005.

P. Hall and M. G. Schimek. Moderate-Deviation-based inference for random degeneration in paired rank lists. *Journal of the American Statistical Association*, 107(498):661–672, 2012.

Heinemeyer, T., Wingender, E., Reuter, I, Hermjakob, H., Kel, A., Kel, O., Ignatieva, E., Ananko, E., Podkolodnaya, O., Kolpakov, F., Podkolodny, N., Kolchanov, and N. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.*, 26(1):362–367, 1998.

Z. Hu and S. Gallo. Identification of interacting transcription factors regulating tissue gene expression in human. *BMC Genomics*, 11(1):49+, 2010.

R. Huber, D. Pietsch, J. Günther, B. Welz, N. Vogt, and K. Brand. Regulation of monocyte differentiation by specific signaling modules and associated transcription factor networks. *Cellular and Molecular Life Sciences*, 71(1):63–92, 2014.

Ingenuity ® Systems, IPA. URL http://www.ingenuity.com.

A. Jankowski, E. Szczurek, R. Jauch, J. Tiuryn, and S. Prabhakar. Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. *Genome Research*, 23(8):1307–1318, 2013.

A. Jankowski, S. Prabhakar, and J. Tiuryn. TACO: a general-purpose tool for predicting cell-type-specific transcription factor dimers. *BMC Genomics*, 15(1):208+, 2014.

S. John, P. J. Sabo, R. E. Thurman, M.-H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager, and J. A. Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43(3):264–268, 2011.

D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, 316(5830):1497–1502, 2007.

M. Kazemian, H. Pham, S. A. Wolfe, M. H. Brodsky, and S. Sinha. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. *Nucleic Acids Research*, 41(17):8237–8252, 2013.

M. Kendall and J. D. Gibbons. *Rank correlation methods*. Oxford University Press, 1990.

H. Klein and M. Vingron. Using transcription factor binding site co-occurrence to predict regulatory regions. *Genome Informatics*, 18:109–118, 2007.

W. Krivan and W. W. Wasserman. A Predictive Model for Regulatory Sequences Directing Liver-Specific Transcription. *Genome Research*, 11(9):1559–1566, 2001.

D. E. Levy and J. E. Darnell. STATs: transcriptional control and biological impact. *Nat Rev Mol Cell Biol*, 3 (9):651–662, 2002.

Q. Li, J. B. Brown, H. Huang, and P. J. Bickel. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, 5(3):1752–1779, 2011a.

X. Y. Li, S. Thomas, P. Sabo, M. Eisen, J. Stamatoyannopoulos, and M. Biggin. The role of chromatin accessibility in directing the widespread, overlapping patterns of drosophila transcription factor binding. *Genome Biology*, 12(4):R34+, 2011b.

C. G. Liew, N. N. Shah, S. J. Briston, R. M. Shepherd, C. P. Khoo, M. J. Dunne, H. D. Moore, K. E. Cosgrove, and P. W. Andrews. PAX4 Enhances β-Cell Differentiation of Human Embryonic Stem Cells. *PLoS ONE*, 3(3):e1783+, 2008.

M. I. Love and H.-R. Chung. Cell-type-specific chromatin accessibility in human fetal tissues. Unpublished manuscript, 2012.

Y. Maeda, V. Dave, and J. A. Whitset. Transcriptional control of lung morphogenesis. *Physiological Reviews*, 87:219–244, 2007.

D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, 39 (suppl 1):D52–D57, 2011.

E. R. Mardis. ChIP-seq: welcome to the new frontier. *Nat Meth*, 4(8):613–614, 2007.

J. H. Martens, D. Rico, K. Downes, D. Richardson, A. Breschi, S. Heath, E. C. de Santa-Pau, S. Saeed, M. Frontini, V. Pancaldi, L. Chen, F. Müller, L. Clarke, H. Kerstens, S. P. Wilder, E. Palumbo, S. Djebali, E. Raineri, A. Merkel, M. Sultan, A. van Bömmel, P. Bertone, M. Gut, M.-L. Yaspo, M. Rubio, J. M. Fernandez, A. Attwood, V. de la Torre, R. Royo, S. Fragkogianni, J. L. GelpÌ, D. Torrents, V. Iotchkova, C. Logie, A. Aghajanirefah, A. Singh, E. Janssen-Megens, K. Berentsen, W. Erber, A. Rendon, M. Konstadima, A. Esteve-Codina, M. van der Ent, A. Kaan, N. Sharifi, D. Paul, D. C. Ifrim, J. Quintin, M. I. Love, D. G. Pisano, F. Burden, N. Foad, S. Farrow, D. R. Zerbino, I. Dunham, T. Kuijpers, H. Lehrach, T. Lengauer, S. Beck, M. G. Netea, P. Flicek, M. Vingron, I. Gut, N. Soranzo, C. Bock, R. Guigo, W. H. Ouwehand, A. Valencia, and H. G. Stunnenberg. Epigenomic plasticity of human neutrophils and monocytes in cord and peripheral blood. *submitted*, 2014.

A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C.-y. Y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, 42(Database issue):gkt997–D147, 2014.

V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(Database issue):D108–D110, 2006.

R. G. Miller. *Simultaneous statistical inference.* McGraw-Hill Book Comp., New York, 1966.

R. Murakami, T. Okumura, and H. Uchiyama. GATA factors as key regulatory molecules in the development of drosophila endoderm. *Development, growth & differentiation*, 47(9):581–589, 2005.

A. Myšičková and M. Vingron. Detection of interacting transcription factors in human tissues using predicted DNA binding affinity. *BMC Genomics*, 13 Suppl 1(Suppl 1):S2+, 2012.

Neph, Shane, Vierstra, Jeff, Stergachis, A. B., Reynolds, A. P., Haugen, Eric, Vernot, Benjamin, Thurman, R. E., John, Sam, Sandstrom, Richard, Johnson, A. K., Maurano, M. T., Humbert, Richard, Rynes, Eric, Wang, Hao, Vong, Shinny, Lee, Kristen, Bates, Daniel, Diegel, Morgan, Roach, Vaughn, Dunn, Douglas, Neri, Jun, Schafer, Anthony, Hansen, R. Scott, Kutyavin, Tanya, Giste, Erika, Weaver, Molly, Canfield, Theresa, Sabo, Peter, Zhang, Miaohua, Balasundaram, Gayathri, Byron, Rachel, MacCoss, M. J., Akey, J. M., Bender, M. A., Groudine, Mark, Kaul, Rajinder, Stamatoyannopoulos, and J. A. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012.

D. E. Newburger and M. L. Bulyk. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic acids research*, 37(Database issue):D77–D82, 2009.

S. Ni and M. Vingron. R2KS: a novel measure for comparing gene expression based on ranked gene lists. *Journal of Computational Biology*, 19(6):766–775, 2012.

G. Nucifora, L. Laricchia-Robbio, and V. Senyuk. EVI1 and hematopoietic disorders: history and perspectives. *Gene*, 368:1–11, 2006.

S. L. Nutt and B. L. Kee. The Transcriptional Regulation of B Cell Lineage Commitment. *Immunity*, 26: 715–725, 2007.

D. T. Odom, N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T. L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. Control of pancreas and liver gene expression by HNF transcription factors. *Science*, 303(5662):1378–1381, 2004.

Y. M. M. Oh, J. K. K. Kim, S. Choi, and J.-Y. Y. Yoo. Identification of co-occurring transcription factor binding sites from DNA sequence using clustered position weight matrices. *Nucleic acids research*, 40(5):e38, 2012.

K. Ohneda and M. Yamamoto. Roles of Hematopoietic Transcription Factors GATA-1 and GATA-2 in the Development of Red Blood Cell Lineage. *Acta Haematol*, 108:237–245, 2002.

U. J. Pape, S. Rahmann, and M. Vingron. Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics*, 24(3):350–357, 2008.

U. J. Pape, H. Klein, and M. Vingron. Statistical detection of cooperative transcription factors with similarity adjustment. *Bioinformatics*, 25(16):2103–2109, 2009.

S.-J. J. Park, T. Umemoto, M. Saito-Adachi, Y. Shiratsuchi, M. Yamato, and K. Nakai. Computational promoter modeling identifies the modes of transcriptional regulation in hematopoietic stem cells. *PloS one*, 9(4), 2014.

H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, N. Kurbatova, M. Lukk, J. Malone, R. Mani, E. Pilicheva, G. Rustici, A. Sharma, E. Williams, T. Adamusiak, M. Brandizi, N. Sklyar, and A. Brazma. ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic acids research*, 39(Database issue):D1002–D1004, 2011.

L. Pinello, J. Xub, S. H. Orkinb, and G.-C. Yuan. Analysis of chromatin-state plasticity identifies cell-type-specific regulators of H3K27me3 patterns. *PNAS*, 111(3):E344?E353, 2014.

R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research*, 21(3): 447–455, 2011.

S. Rahmann, T. Müller, and M. Vingron. On the power of profiles for transcription factor binding site detection. *Statistical Applications in Genetics and Molecular Biology*, 2(1), 2003.

J. Ramirez, K. Lukin, and J. Hagman. From hematopoietic progenitors to b cells: mechanisms of lineage restriction and commitment. *Current Opinion in Immunology*, 22:177:184, 2010.

T. Ravasi, H. Suzuki, C. V. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. Forrest, J. Gough, S. Grimmond, J.-H. H. Han, T. Hashimoto, W. Hide, O. Hofmann, A. Kamburov, M. Kaur, H. Kawaji, A. Kubosaki, T. Lassmann, E. van Nimwegen, C. R. R. MacPherson, C. Ogawa, A. Radovanovic, A. Schwartz, R. D. Teasdale, J. Tegnér, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker, and Y. Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, 2010.

A. Remenyi, H. R. Scholer, and M. Wilmanns. Combinatorial control of gene expression. *Nat Struct Mol Biol*, 11(9):812–815, 2004.

B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science (New York, N.Y.)*, 290(5500):2306–2309, 2000.

G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Seraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotech*, 17(10):1030–1032, 1999.

J. M. Robins, A. van der Vaart, and V. Ventura. Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, 95(452):1143–1156, 2000.

H. G. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23:134–141, 2007.

H. G. Roider, T. Manke, S. O'Keeffe, M. Vingron, and S. A. Haas. PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, 25(4):435–442, 2009.

E. Roulet, S. Busso, A. A. Camargo, A. J. Simpson, N. Mermod, and P. Bucher. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nature biotechnology*, 20(8): 831–835, 2002.

V. Sartorelli and G. Caretti. Mechanisms underlying the transcriptional regulation of skeletal myogenesis. *Current Opinion in Genetics & Development*, 15(5):528–535, 2005.

M. G. Schimek, A. Myšičková, and E. Budinská. An inference and integration approach for the consolidation of ranked lists. *Communications in Statistics - Simulation and Computation*, 41:1152–1166, 2012.

J. Schlesinger, M. Schueler, M. Grunert, J. J. Fischer, Q. Zhang, T. Krueger, M. Lange, M. Tönjes, I. Dunkel, and S. R. Sperling. The Cardiac Transcription Network Modulated by Gata4, Mef2a, Nkx2.5, Srf, Histone Modifications, and MicroRNAs. *PLoS Genet*, 7(2):e1001313+, 2011.

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

B. A. Shoemaker and A. R. Panchenko. Deciphering Protein-Protein interactions. part i. experimental techniques and databases. *PLoS Comput Biol*, 3(3):e42+, 2007.

R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

A. Smit, R. Hubley, and P. Green. RepeatMasker Open-3.0., 1996-2010. URL http://www.repeatmasker.org.

A. D. Smith, P. Sumazin, and M. Q. Zhang. Tissue-specific regulatory elements in mammalian promoters. *Mol Syst Biol*, 3, 2007.

L. Song, Z. Zhang, L. L. Grasfeder, A. P. Boyle, P. G. Giresi, B.-K. Lee, N. C. Sheffield, S. Gräf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Wang, D. Winter, Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer, G. E. Crawford, J. D. Lieb, and T. S. Furey. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, 21(10):gr.121541.111–1767, 2011.

C. Spearman. 'Footrule' for measuring correlation. *British Journal of Psychology*, 2 (1):89–108, 1906.

C. Stark, B.-J. Breitkreutz, A. Chatr-aryamontri, L. Boucher, R. Oughtred, M. S. Livstone3, J. Nixon, K. V. Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers. The biogrid interaction database: 2011 update. *Nucleic Acids Res.*, 39 (Suppl. 1):D698–D704, 2011.

G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the 'perceptron' algorithm to distinguish translational initiation sites in e. coli. *Nucleic Acids Research*, 10(9):2997–3011, 1982.

A. Stuart, K. Ord, and S. Arnold. *Kendall's Advanced Theory of Statistics: Volume 2A, Classical Inference and the Linear Model*. Wiley, 6th edition, 2008.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

H. Suzuki, Y. Fukunishi, I. Kagawa, R. Saito, H. Oda, T. Endo, S. Kondo, H. Bono, Y. Okazaki, and Y. Hayashizaki. Protein-protein interaction panel using mouse full-length cDNAs. *Genome research*, 11 (10):1758–1765, 2001.

N. Tanaka, T. Kawakami, and T. Taniguchi. Recognition DNA sequences of interferon regulatory factor 1 (IRF-1) and IRF-2, regulators of cell growth and the interferon system. *Molecular and cellular biology*, 13 (8):4531–4538, 1993.

A. Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Research*, 16(8): 962–972, 2006.

W. D. Tembe, J. V. Pearson, N. Homer, J. Lowey, E. Suh, and D. W. Craig. Statistical comparison framework and visualization scheme for ranking-based algorithms in high-throughput genome-wide studies. *Journal of Computational Biology*, 16(4):565–577, 2009.

D. G. Tenen, R. Hromas, J. D. Licht, and D.-E. Zhang. Transcription Factors, Normal Myeloid Development, and Leukemia. *Blood*, 90(2):489–519, 1997.

The ENCODE Project consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

M. Thomas-Chollier, M. Defrance, A. Medina-Rivera, O. Sand, C. Herrmann, D. Thieffry, and J. van Helden. RSAT 2011: regulatory sequence analysis tools. *Nucleic acids research*, 39(Web Server issue):W86–W91, 2011.

N. Treiber, T. Treiber, G. Zocher, and R. Grosschedl. Structure of an Ebf1:DNA complex reveals unusual DNA recognition and structural homology with rel proteins. *Genes & Development*, 24(20):2270–2275, 2010.

UniProt Consortium. UniProt knowledgebase: a hub of integrated protein data. *Database*, 2011:bar009+, 2011.

A. Vandenbon, Y. Kumagai, S. Akira, and D. Standley. A novel unbiased measure for motif co-occurrence predicts combinatorial regulation of transcription. *BMC Genomics*, 13(Suppl 7):S11+, 2012.

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, 2009.

D. Wang, I. Garcia-Bassets, C. Benner, W. Li, X. Su, Y. Zhou, J. Qiu, W. Liu, M. U. Kaikkonen, K. A. Ohgi, C. K. Glass, M. G. Rosenfeld, and X.-D. D. Fu. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, 474(7351):390–394, 2011.

J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22(9):1798–1812, 2012.

H. Weintraub and M. Groudine. Chromosomal subunits in active genes have an altered conformation. *Science*, 193(4256):848–856, 1976.

A. West, P. Shore, and A. Sharrocks. DNA binding by MADS-box transcription factors: a molecular mechanism for differential DNA bending. *Molecular and Cellular Biology*, 17(5):2876?2887, 1997.

T. Whitfield, J. Wang, P. Collins, E. C. Partridge, S. Aldred, N. Trinklein, R. Myers, and Z. Weng. Functional analysis of transcription factor binding sites in human promoters. *Genome Biology*, 13(9):R50+, 2012.

T. Whitington, M. C. Frith, J. Johnson, and T. L. Bailey. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Research*, 39(15):e98, 2011.

N. K. Wilson, S. D. Foster, X. Wang, K. Knezevic, J. Sch¸tte, P. Kaimakis, P. M. Chilarska, S. Kinston, W. H. Ouwehand, E. Dzierzak, J. E. Pimanda, M. F. de Bruijn, and B. Gˆttgens. Combinatorial Transcriptional Control In Blood Stem/Progenitor Cells: Genome-wide Analysis of Ten Major Transcriptional Regulators. *Cell Stem Cell*, 7(4):532–544, 2010.

H. Xi, H. P. Shulha, J. M. Lin, T. R. Vales, Y. Fu, D. M. Bodine, R. D. McKay, J. G. Chenoweth, P. J. Tesar, T. S. Furey, B. Ren, Z. Weng, and G. E. Crawford. Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS genetics*, 3(8):e136+, 2007.

Z. Xie, S. Hu, J. Qian, S. Blackshaw, and H. Zhu. Systematic characterization of protein-DNA interactions. *Cellular and molecular life sciences : CMLS*, 68(10):1657–1668, 2011.

J.-C. Yeo and H.-H. Ng. The transcriptional regulation of pluripotency. *Cell Research*, 23:20–32, 2013.

C. Yoshida, F. Yoshida, D. E. Sears, S. M. Hart, D. Ikebe, A. Muto, S. Basu, K. Igarashi, and J. V. Melo. Bcr-Abl signaling through the PI-3/S6 kinase pathway inhibits nuclear translocation of the transcription factor bach2, which represses the antiapoptotic factor heme oxygenase-1. *Blood*, 109(3):1211–1219, 2007.

Yu, Xueping, Lin, Jimmy, Zack, D. J., Qian, and Jiang. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Research*, 34(17):4925–4936, 2006.

Y. Zhao, D. Granas, and G. D. Stormo. Inferring Binding Energies from Selected Binding Sites. *PLoS Comput Biol*, 5(12):e1000590+, 2009.

List of Abbreviations

A adenine

C cytosine

G guanine

T thymine

bp base pair

ChIP chromatin immunoprecipitation

CTS-DHS cell-type specific DNase hypersensitive site

DHS DNase I hypersensitive site

DNA deoxyribonucleic acid

ESC embryonic stem cell

FDR false discovery rate

IDR irreproducible discovery rate

ML information content

IC maximum likelihood

PBM protein binding microarray

PPI protein-protein interaction

PSFM position-specific frequency matrix

PWM position weight matrix

RNA ribonucleic acid

TF transcription factor

TFBS transcription factor binding site

TSS transcription start site

TRAP transcription factor affinity prediction tool from Roider et al. (2007)

Y2H yeast-two-hybrid system

Notations

S	sequence
S_j	position j in sequence S
l	length of sequence
$F(4 \times l)$	position-specific frequency matrix of length l
$M(4 \times l)$	position weight matrix of length l
r	Pearson's correlation coefficient
$r_x(i)$	rank of object i according to quantity x
R_x	ranked list ordered according to quantity x
$\|R_x\|$	length of ranked list R_x
ρ_{Sp}	Spearman's correlation
R_{Sp}	Spearman's footrule
τ_K	Kendall's τ
R^*	R2KS score
n_{jk}	observed cell count in a 2-way contingency table
$S^{\mathsf{max}}(M_i, M_j)$	MOSTA similarity measure between PWMs M_i and M_j
H_0	null hypothesis
H_a	alternative hypothesis
$Z_t(i)$	binary variable indicating the specificity of gene i in tissue/cell type t

$X_k(i)$	binary variable indicating whether gene i is ranked among the top k genes
π_{jkl}	cell probability in a 3-way contingency table
n_{jkl}	observed cell count in a 3-way contingency table
μ_{jkl}	expected cell frequency in a 3-way contingency table
$\widehat{\mu}_{jkl}$	fitted cell frequency in a 3-way contingency table
$X \perp\!\!\!\perp Y$	random variable X is statistically independent of random variable Y
w	genomic window of a constant length
C_{wi}	normalized log read count of sample i in window w
$\overline{C}_{wl}$	average log read count of cell type l in window w
$\overline{C}_{wG}$	global average log read count over all cell types in window w
t_{wl}	Student's t-statistic for cell type l in window w
L_l	log ratio of p-value obtained from partial table in cell type l and of p-value obtained from ubiquitous partial table

A Appendix

A.1 Supplementary Figures

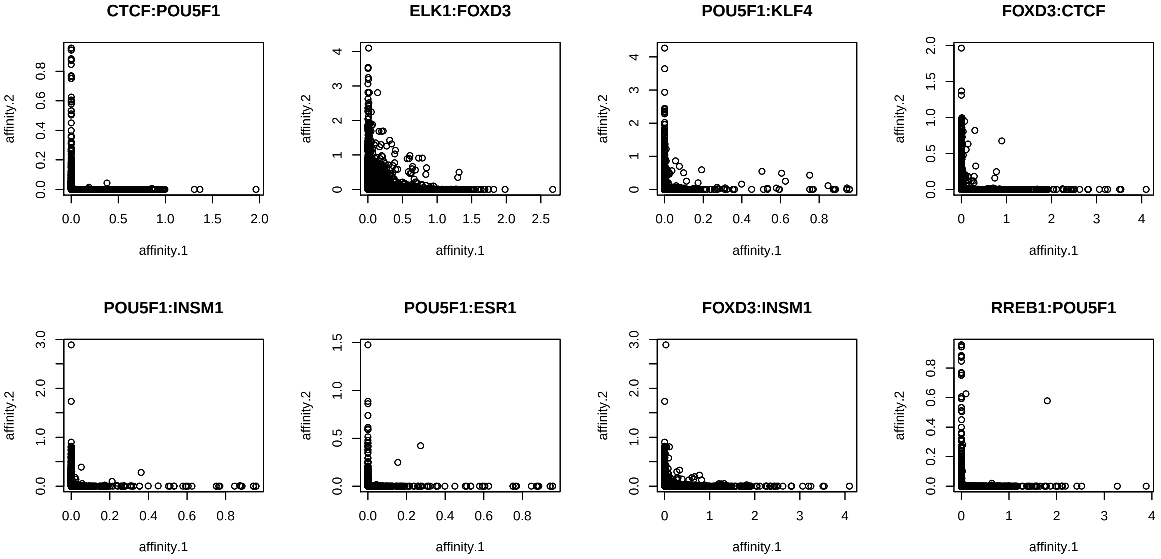


Figure A.1: Scatterplots of binding affinities for 8 TF pairs with large proportion of promoters with low irreducible discovery rate (IDR$< 10^{-15}$).

A.2 Supplementary Tables

Table A.1: DNase data from The ENCODE Project consortium from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/ used in Chapter 5 with corresponding cell line identifier, assigned cell type, tissue and type of cell.

No.	File name	Cell line	Cell type	Tissue	Type
1	wgEncodeUwDnaseAg04449-AlnRep1	AG04449	fetal thigh fibroblast	skin	fibroblast
2	wgEncodeUwDnaseAg04449-AlnRep2	AG04449	fetal thigh fibroblast	skin	fibroblast
3	wgEncodeUwDnaseAg04450-AlnRep1	AG04450	fetal lung fibroblast	lung	fibroblast
4	wgEncodeUwDnaseAg04450-AlnRep2	AG04450	fetal lung fibroblast	lung	fibroblast
5	wgEncodeUwDnaseAg09309-AlnRep1	AG09309	toe fibroblast	skin	fibroblast
6	wgEncodeUwDnaseAg09309-AlnRep2	AG09309	toe fibroblast	skin	fibroblast
7	wgEncodeUwDnaseAg09319-AlnRep1	AG09319	gum fibroblast	gingival	fibroblast
8	wgEncodeUwDnaseAg09319-AlnRep2	AG09319	gum fibroblast	gingival	fibroblast
9	wgEncodeUwDnaseAg10803-AlnRep1	AG10803	abdominal fibroblast	skin	fibroblast
10	wgEncodeUwDnaseAg10803-AlnRep2	AG10803	abdominal fibroblast	skin	fibroblast
11	wgEncodeUwDnaseAoafAlnRep1	AoAF	aortic fibroblast	blood vessel	fibroblast
12	wgEncodeUwDnaseAoafAlnRep2	AoAF	aortic fibroblast	blood vessel	fibroblast
13	wgEncodeUwDnaseBjAlnRep1	BJ	skin fibroblast	skin	fibroblast
14	wgEncodeUwDnaseBjAlnRep2	BJ	skin fibroblast	skin	fibroblast
15	wgEncodeUwDnaseCd34mobilized-AlnRep1	CD34+ Mobilized	hematopoietic progenitor	blood	white blood
16	wgEncodeUwDnaseCd4-naivewb11970640AlnRep1	CD4+ Naive Wb11970640	T-cell	blood	white blood
17	wgEncodeUwDnaseCd4-naivewb78495824AlnRep1	CD4+ Naive Wb78495824	T-cell	blood	white blood
18	wgEncodeUwDnaseGm04503-AlnRep1	GM04503	twin fibroblast	skin	fibroblast
19	wgEncodeUwDnaseGm04503-AlnRep2	GM04503	twin fibroblast	skin	fibroblast
20	wgEncodeUwDnaseGm04504-AlnRep1	GM04504	twin fibroblast	skin	fibroblast
21	wgEncodeUwDnaseGm04504-AlnRep2	GM04504	twin fibroblast	skin	fibroblast
22	wgEncodeUwDnaseGm06990-AlnRep1	GM06990	B-lymphocyte	blood	white blood
23	wgEncodeUwDnaseGm06990-AlnRep2	GM06990	B-lymphocyte	blood	white blood
24	wgEncodeUwDnaseGm12864-AlnRep1	GM12864	B-lymphocyte	blood	white blood

25	wgEncodeUwDnaseGm12865-AlnRep1	GM12865	B-lymphocyte	blood	white blood
26	wgEncodeUwDnaseGm12865-AlnRep2	GM12865	B-lymphocyte	blood	white blood
27	wgEncodeUwDnaseGm12878-AlnRep1	GM12878	B-lymphocyte	blood	white blood
28	wgEncodeUwDnaseGm12878-AlnRep2	GM12878	B-lymphocyte	blood	white blood
29	wgEncodeUwDnaseH1hescAlnRep1	H1-hESC	ESC	ESC	SC
30	wgEncodeUwDnaseH7esAlnRep1	H7-hESC	undiff ESC	ESC	SC
31	wgEncodeUwDnaseH7esAlnRep2	H7-hESC	undiff ESC	ESC	SC
32	wgEncodeUwDnaseH7esDiffa14d-AlnRep1	H7-hESC diff14d	diff ESC	ESC	SC
33	wgEncodeUwDnaseH7esDiffa14d-AlnRep2	H7-hESC diff14d	diff ESC	ESC	SC
34	wgEncodeUwDnaseH7esDiffa2d-AlnRep1	H7-hESC diff2d	diff ESC	ESC	SC
35	wgEncodeUwDnaseH7esDiffa5d-AlnRep1	H7-hESC diff5d	diff ESC	ESC	SC
36	wgEncodeUwDnaseH7esDiffa5d-AlnRep2	H7-hESC diff5d	diff ESC	ESC	SC
37	wgEncodeUwDnaseH7esDiffa9d-AlnRep1	H7-hESC diff9d	diff ESC	ESC	SC
38	wgEncodeUwDnaseHacAlnRep1	HAc	astrocytes cerebellar	brain	astrocytes
39	wgEncodeUwDnaseHacAlnRep2	HAc	astrocytes cerebellar	brain	astrocytes
40	wgEncodeUwDnaseHaeAlnRep1	HAEpiC	amniotic epithelial	epithelium	epithelium
41	wgEncodeUwDnaseHaeAlnRep2	HAEpiC	amniotic epithelial	epithelium	epithelium
42	wgEncodeUwDnaseHahAlnRep1	HA-h	astrocytes hippocampal	brain	astrocytes
43	wgEncodeUwDnaseHahAlnRep2	HA-h	astrocytes hippocampal	brain	astrocytes
44	wgEncodeUwDnaseHaspAlnRep1	HA-sp	astrocytes spinal cord	brain	astrocytes
45	wgEncodeUwDnaseHaspAlnRep2	HA-sp	astrocytes spinal cord	brain	astrocytes
46	wgEncodeUwDnaseHbmecAlnRep1	HBMEC	brain microvascular endothelial	blood vessel	endothelial
47	wgEncodeUwDnaseHbmecAlnRep2	HBMEC	brain microvascular endothelial	blood vessel	endothelial
48	wgEncodeUwDnaseHbvpAlnRep1	HBVP	brain vascular	blood vessel	endothelial
49	wgEncodeUwDnaseHbvsmcAlnRep1	HBVSMC	brain vascular	blood vessel	myoblast
50	wgEncodeUwDnaseHbvsmcAlnRep2	HBVSMC	brain vascular	blood vessel	myoblast
51	wgEncodeUwDnaseHcfAlnRep1	HCF	cardiac fibroblasts	heart	fibroblast
52	wgEncodeUwDnaseHcfAlnRep2	HCF	cardiac fibroblasts	heart	fibroblast
53	wgEncodeUwDnaseHcfaaAlnRep1	HCFaa	cardiac fibroblasts atrial	heart	fibroblast
54	wgEncodeUwDnaseHcfaaAlnRep2	HCFaa	cardiac fibroblasts atrial	heart	fibroblast

55	wgEncodeUwDnaseHcmAlnRep1	HCM	cardiac myocytes	heart	myoblast
56	wgEncodeUwDnaseHcmAlnRep2	HCM	cardiac myocytes	heart	myoblast
57	wgEncodeUwDnaseHconfAlnRep1	HConF	conjunctival fibroblast	eye	fibroblast
58	wgEncodeUwDnaseHconfAlnRep2	HConF	conjunctival fibroblast	eye	fibroblast
59	wgEncodeUwDnaseHcpeAlnRep1	HCPEpiC	plexus epithelial	epithelium	epithelium
60	wgEncodeUwDnaseHcpeAlnRep2	HCPEpiC	plexus epithelial	epithelium	epithelium
61	wgEncodeUwDnaseHeeAlnRep1	HEEpiC	esophageal epithelial	epithelium	epithelium
62	wgEncodeUwDnaseHeeAlnRep2	HEEpiC	esophageal epithelial	epithelium	epithelium
63	wgEncodeUwDnaseHelas3AlnRep1	HeLa-S3	cervical carcinoma	urogenital	cancer
64	wgEncodeUwDnaseHelas3AlnRep2	HeLa-S3	cervical carcinoma	urogenital	cancer
65	wgEncodeUwDnaseHffAlnRep1	HFF	foreskin fibroblast	urogenital	fibroblast
66	wgEncodeUwDnaseHffAlnRep2	HFF	foreskin fibroblast	urogenital	fibroblast
67	wgEncodeUwDnaseHffmycAlnRep1	HFF-Myc	foreskin fibroblast	urogenital	fibroblast
68	wgEncodeUwDnaseHffmycAlnRep2	HFF-Myc	foreskin fibroblast	urogenital	fibroblast
69	wgEncodeUwDnaseHgfAlnRep1	HGF	gingival fibroblasts	gingival	fibroblast
70	wgEncodeUwDnaseHgfAlnRep2	HGF	gingival fibroblasts	gingival	fibroblast
71	wgEncodeUwDnaseHipeAlnRep1	HIPEpiC	pigment epithelial	epithelium	epithelium
72	wgEncodeUwDnaseHipeAlnRep2	HIPEpiC	pigment epithelial	epithelium	epithelium
73	wgEncodeUwDnaseHmecAlnRep1	HMEC	mammary epithelial	epithelium	epithelium
74	wgEncodeUwDnaseHmecAlnRep2	HMEC	mammary epithelial	epithelium	epithelium
75	wgEncodeUwDnaseHmfAlnRep1	HMF	mammary fibroblasts	urogenital	fibroblast
76	wgEncodeUwDnaseHmfAlnRep2	HMF	mammary fibroblasts	urogenital	fibroblast
77	wgEncodeUwDnaseHmvecdad-AlnRep1	HMVEC-dAd	lymph microvascular endothelial dermal	blood vessel	endothelial
78	wgEncodeUwDnaseHmvecdad-AlnRep2	HMVEC-dAd	lymph microvascular endothelial dermal	blood vessel	endothelial
79	wgEncodeUwDnaseHmvecdblad-AlnRep1	HMVEC-dBl-Ad	blood microvascular endothelial dermal	blood vessel	endothelial
80	wgEncodeUwDnaseHmvecdblad-AlnRep2	HMVEC-dBl-Ad	blood microvascular endothelial dermal	blood vessel	endothelial

81	wgEncodeUwDnaseHmvecdblneo-AlnRep1	HMVEC-dBl-Neo	neonatal blood microvascular endothelial	blood vessel	endothelial
82	wgEncodeUwDnaseHmvecdblneo-AlnRep2	HMVEC-dBl-Neo	neonatal blood microvascular endothelial	blood vessel	endothelial
83	wgEncodeUwDnaseHmvecdlyad-AlnRep1	HMVEC-dLy-Ad	lymph microvascular endothelial dermal	blood vessel	endothelial
84	wgEncodeUwDnaseHmvecdlyad-AlnRep2	HMVEC-dLy-Ad	lymph microvascular endothelial dermal	blood vessel	endothelial
85	wgEncodeUwDnaseHmvecdlyneo-AlnRep1	HMVEC-dLy-Neo	lymph microvascular endothelial dermal	blood vessel	endothelial
86	wgEncodeUwDnaseHmvecdlyneo-AlnRep2	HMVEC-dLy-Neo	lymph microvascular endothelial dermal	blood vessel	endothelial
87	wgEncodeUwDnaseHmvecdneo-AlnRep1	HMVEC-dNeo	lymph microvascular endothelial dermal	blood vessel	endothelial
88	wgEncodeUwDnaseHmvecdneo-AlnRep2	HMVEC-dNeo	lymph microvascular endothelial dermal	blood vessel	endothelial
89	wgEncodeUwDnaseHmveclbl-AlnRep1	HMVEC-LBl	blood microvascular endothelial lung	blood vessel	endothelial
90	wgEncodeUwDnaseHmveclbl-AlnRep2	HMVEC-LBl	blood microvascular endothelial lung	blood vessel	endothelial
91	wgEncodeUwDnaseHmveclly-AlnRep1	HMVEC-LLy	lymph microvascular endothelial lung	blood vessel	endothelial
92	wgEncodeUwDnaseHmveclly-AlnRep2	HMVEC-LLy	lymph microvascular endothelial lung	blood vessel	endothelial
93	wgEncodeUwDnaseHnpceAlnRep1	HNPCEpiC	ciliary epithelial	epithelium	epithelium
94	wgEncodeUwDnaseHnpceAlnRep2	HNPCEpiC	ciliary epithelial	epithelium	epithelium
95	wgEncodeUwDnaseHpaecAlnRep1	HPAEC	pulmonary artery endothelial	blood vessel	endothelial
96	wgEncodeUwDnaseHpafAlnRep1	HPAF	pulmonary artery fibroblasts	blood vessel	fibroblast
97	wgEncodeUwDnaseHpafAlnRep2	HPAF	pulmonary artery fibroblasts	blood vessel	fibroblast
98	wgEncodeUwDnaseHpdlfAlnRep1	HPdLF	ligament fibroblasts	epithelium	fibroblast
99	wgEncodeUwDnaseHpdlfAlnRep2	HPdLF	ligament fibroblasts	epithelium	fibroblast
100	wgEncodeUwDnaseHpfAlnRep1	HPF	pulmonary fibroblasts	lung	fibroblast

101	wgEncodeUwDnaseHpfAlnRep2	HPF	pulmonary fibroblasts	lung	fibroblast
102	wgEncodeUwDnaseHrceAlnRep1	HRCEpiC	renal epithelial	epithelium	epithelium
103	wgEncodeUwDnaseHrceAlnRep2	HRCEpiC	renal epithelial	epithelium	epithelium
104	wgEncodeUwDnaseHreAlnRep1	HRE	renal epithelial	epithelium	epithelium
105	wgEncodeUwDnaseHreAlnRep2	HRE	renal epithelial	epithelium	epithelium
106	wgEncodeUwDnaseHrgecAlnRep1	HRGEC	renal glomerular endothelial	kidney	endothelial
107	wgEncodeUwDnaseHrgecAlnRep2	HRGEC	renal glomerular endothelial	kidney	endothelial
108	wgEncodeUwDnaseHrpeAlnRep1	HRPEpiC	retinal epithelial	epithelium	epithelium
109	wgEncodeUwDnaseHrpeAlnRep2	HRPEpiC	retinal epithelial	epithelium	epithelium
110	wgEncodeUwDnaseHs27aAlnRep1	Hs27	marrow stromal	bone marrow	white blood
111	wgEncodeUwDnaseHs5AlnRep1	Hs5	marrow stromal	bone marrow	white blood
112	wgEncodeUwDnaseHsmmAlnRep1	HSMM	muscle myoblast	muscle	myoblast
113	wgEncodeUwDnaseHsmmAlnRep2	HSMM	muscle myoblast	muscle	myoblast
114	wgEncodeUwDnaseHsmmtAlnRep1	HSMMtube	muscle myoblast	muscle	myoblast
115	wgEncodeUwDnaseHsmmtAlnRep2	HSMMtube	muscle myoblast	muscle	myoblast
116	wgEncodeUwDnaseHuvecAlnRep1	HUVEC	umbilical vein endothelial	blood vessel	endothelial
117	wgEncodeUwDnaseHuvecAlnRep2	HUVEC	umbilical vein endothelial	blood vessel	endothelial
118	wgEncodeUwDnaseHvmfAlnRep1	HVMF	mesenchymal fibroblast	urogenital	fibroblast
119	wgEncodeUwDnaseHvmfAlnRep2	HVMF	renal epithelial	urogenital	epithelium
120	wgEncodeUwDnaseK562AlnRep1	K562	leukemia	blood	cancer
121	wgEncodeUwDnaseK562AlnRep2	K562	leukemia	blood	cancer
122	wgEncodeUwDnaseLhcnm2AlnRep1	LHCN-M2	skeletal myoblasts	muscle	myoblast
123	wgEncodeUwDnaseLhcnm2AlnRep2	LHCN-M2	skeletal myoblasts	muscle	myoblast
124	wgEncodeUwDnaseLhcnm2Diff4d-AlnRep1	LHCN-M2 Diff4	skeletal myoblasts	muscle	myoblast
125	wgEncodeUwDnaseLhcnm2Diff4d-AlnRep2	LHCN-M2 Diff4	skeletal myoblasts	muscle	myoblast
126	wgEncodeUwDnaseMonocd14-AlnRep1	Monocytes-CD14+	monocyte	blood	white blood
127	wgEncodeUwDnaseMonocd14ro1746-AlnRep1V2	Monocytes-CD14+ RO01746	monocyte	blood	white blood
128	wgEncodeUwDnaseMonocd14ro1746-AlnRep2	Monocytes-CD14+ RO01746	monocyte	blood	white blood
129	wgEncodeUwDnaseMscAlnRep1	MSC	MSC	adipose	SC
130	wgEncodeUwDnaseMscAlnRep2	MSC	MSC	adipose	SC
131	wgEncodeUwDnaseNhaAlnRep1	NH-A	astrocyte	brain	astrocytes
132	wgEncodeUwDnaseNhaAlnRep2	NH-A	astrocyte	brain	astrocytes
133	wgEncodeUwDnaseNhberaAlnRep1	NHBE RA	bronchial epithelial	epithelium	epithelium
134	wgEncodeUwDnaseNhberaAlnRep2	NHBE RA	bronchial epithelial	epithelium	epithelium
135	wgEncodeUwDnaseNhdfadAlnRep1	NHDF-Ad	dermal fibroblasts	skin	fibroblast

136	wgEncodeUwDnaseNhdfadAlnRep2	NHDF-Ad	dermal fibroblasts	skin	fibroblast
137	wgEncodeUwDnaseNhdfneo-AlnRep1	NHDF-neo	neonatal dermal fibroblasts	skin	fibroblast
138	wgEncodeUwDnaseNhdfneo-AlnRep2	NHDF-neo	neonatal dermal fibroblasts	skin	fibroblast
139	wgEncodeUwDnaseNhlfAlnRep1	NHLF	lung fibroblast	lung	fibroblast
140	wgEncodeUwDnaseNhlfAlnRep2	NHLF	lung fibroblast	lung	fibroblast
141	wgEncodeUwDnasePrecAlnRep1	PrEC	epithelial	epithelium	epithelium
142	wgEncodeUwDnasePrecAlnRep2	PrEC	epithelial	epithelium	epithelium
143	wgEncodeUwDnaseRptecAlnRep1	RPTEC	renal tubule epithelial	epithelium	epithelium
144	wgEncodeUwDnaseRptecAlnRep2	RPTEC	renal tubule epithelial	epithelium	epithelium
145	wgEncodeUwDnaseSaecAlnRep1	SAEC	epithelial	epithelium	epithelium
146	wgEncodeUwDnaseSaecAlnRep2	SAEC	epithelial	epithelium	epithelium
147	wgEncodeUwDnaseSkmcAlnRep1	SKMC	skeletal striated muscle	muscle	myoblast
148	wgEncodeUwDnaseSkmcAlnRep2	SKMC	skeletal striated muscle	muscle	myoblast
149	wgEncodeUwDnaseTh17AlnRep1	Th17	T-cell	blood	white blood
150	wgEncodeUwDnaseTh1AlnRep1	Th1	T-cell primary	blood	white blood
151	wgEncodeUwDnaseTh1AlnRep2	Th1	T-cell primary	blood	white blood
152	wgEncodeUwDnaseTh1wb33676984-AlnRep1	Th1 Wb33676984	T-cell	blood	white blood
153	wgEncodeUwDnaseTh1wb54553204-AlnRep1	Th1 Wb54553204	T-cell	blood	white blood
154	wgEncodeUwDnaseTh1wb54553204-AlnRep2	Th1 Wb54553204	T-cell	blood	white blood
155	wgEncodeUwDnaseTh2AlnRep1	Th2	T-cell primary	blood	white blood
156	wgEncodeUwDnaseTh2AlnRep2	Th2	T-cell	blood	white blood
157	wgEncodeUwDnaseTh2wb33676984-AlnRep1	Th2 Wb33676984	T-cell	blood	white blood
158	wgEncodeUwDnaseTh2wb54553204-AlnRep1	Th2 Wb54553204	T-cell	blood	white blood
159	wgEncodeUwDnaseTreg-wb78495824AlnRep1	Treg Wb78495824	T-cell regulatory	blood	white blood
160	wgEncodeUwDnaseTreg-wb83319432AlnRep1	Treg Wb78495824	T-cell regulatory	blood	white blood
161	wgEncodeUwDnaseWi38AlnRep1	WI-38	embryonic lung fibroblast	lung	fibroblast
162	wgEncodeUwDnaseWi38AlnRep2	WI-38	embryonic lung fibroblast	lung	fibroblast
163	wgEncodeUwDnaseWi38Ohtam-AlnRep1	WI-38-Ohtam	embryonic lung fibroblast	lung	fibroblast
164	wgEncodeUwDnaseWi38Ohtam-AlnRep2	WI-38-Ohtam	embryonic lung fibroblast	lung	fibroblast

Table A.2: Transcription factor binding motifs and corresponding transcription factors or transcription factor groups used in Chapter 5.

TF binding motif	TF/TF group
AHR_01	AHR
AHR_Q5	AHR
AHRARNT_01	AHR:ARNT
AHRARNT_02	AHR:ARNT
AHRHIF_Q6	AHR:ARNT:HIF1A
AIRE_01	AIRE
AIRE_02	AIRE
ALPHACP1_01	NFYA
ALX4_01	ALX4
AMEF2_Q6	MEF2A
AML_Q6	RUNX
AML1_01	RUNX1
AML1_Q6	RUNX1
AP1_01	FOS:JUN:AP1
AP1_Q2_01	FOS:JUN:AP1
AP1_Q4_01	FOS:JUN:AP1
AP1_Q6_01	FOS:JUN:AP1
AP1FJ_Q2	FOS:JUN:AP1
AP2_Q3	TFAP2
AP2_Q6	TFAP2
AP2_Q6_01	TFAP2
AP2ALPHA_01	TFAP2A
AP2ALPHA_02	TFAP2A
AP2ALPHA_03	TFAP2A
AP2GAMMA_01	TFAP2C
AP2REP_01	KLF12
AP4_01	TFAP4
AP4_Q5	TFAP4
AP4_Q6	TFAP4
AP4_Q6_01	TFAP4
AR_01	AR
AR_02	AR
AR_03	AR
AR_Q2	AR
AR_Q6	AR
AREB6_01	ZEB1
AREB6_02	ZEB1
AREB6_03	ZEB1
AREB6_04	ZEB1
ARNT_01	ARNT
ARNT_02	ARNT
ARP1_01	NR2F2
ATF1_Q6	ATF
ATF3_Q6	ATF
ATF4_Q2	ATF
ATF6_01	ATF
BACH1_01	BACH1
BACH2_01	BACH2
BLIMP1_Q6	PRDM1
BRACH_01	T:BRACH
BRN2_01	POU3F2

CACCCBINDINGFACTOR_Q6	ZNF148
CART1_01	ALX1
CDC5_01	CDC5L
CDP_01	CUX1
CDP_02	CUX1
CDPCR1_01	CUX1
CDPCR3_01	CUX1
CDPCR3HD_01	CUX1
CDX_Q5	CDX
CDX2_Q5	CDX
CEBP_01	CEBP
CEBP_C	CEBP
CEBP_Q2	CEBP
CEBP_Q2_01	CEBP
CEBP_Q3	CEBP
CEBPA_01	CEBP
CEBPB_01	CEBPB
CEBPB_02	CEBPB
CETS1P54_01	ETS1:P54
CHOP_01	CEBPA:DDIT3
CHX10_01	ABCD4:VSX2
CIZ_01	ZNF384
CMYB_01	MYB
COUP_01	NR2F1:HNF4A
COUP_DR1_Q6	NR2F
COUPTF_Q6	NR2F
CP2_01	TFCP2
CP2_02	TFCP2
CREB_01	CREB1
CREB_02	CREB1
CREB_Q2	CREB1
CREB_Q2_01	CREB1:CREM
CREB_Q3	ATF:CREB
CREB_Q4	CREB1
CREB_Q4_01	CREB1:CREM
CREBATF_Q6	ATF:CREB
CREBP1_01	ATF:CREB
CREBP1_Q2	ATF:CREB
CREBP1CJUN_01	ATF:JUN
CREL_01	REL
CRX_Q4	CRX:RAX
DBP_Q6	DBP
DEC_Q1	BHLHB
DR1_Q3	HNF4:NR2
DR3_Q4	RXR:VDR:NR1I
DR4_Q2	RAR:RXR:NR2F:NR1
E12_Q6	TCF3
E2A_Q2	MYF:MYOD:TCF
E2A_Q6	MYF:MYOD:TCF
E2F_01	E2F1
E2F_02	E2F
E2F_03	E2F1
E2F_Q2	E2F
E2F_Q3	E2F1
E2F_Q3_01	E2F
E2F_Q4	E2F1

E2F_Q4_01	E2F
E2F_Q6	E2F1
E2F_Q6_01	E2F
E2F1_Q3	E2F1
E2F1_Q3_01	E2F1
E2F1_Q4	E2F1
E2F1_Q6	E2F1
E2F1_Q6_01	E2F1
E47_01	TCF3
E47_02	TCF3
E4BP4_01	NFIL3
EBF_Q6	EBF
EBOX_Q6_01	EBOX:MYC:MYF:MYOD:TAL:USF
EFC_Q6	RFX1
EGR_Q6	EGR
EGR1_01	EGR1
EGR2_01	EGR2
EGR3_01	EGR3
ELF1_Q6	ELF1
ELK1_01	ELK1
ELK1_02	ELK1
EN1_01	EN1
ER_Q6	ESR
ER_Q6_02	ESR
ERR1_Q2	ESRRA
ETF_Q6	ETFA:TEAD2
ETS_Q4	ELF:ELK:ETS:FLI1:GABP
ETS_Q6	ELF:ELK:ETS:FLI1:GABP
ETS1_B	ETS1
EVI1_01	EVI1
EVI1_02	EVI1
EVI1_03	EVI1
EVI1_04	EVI1
EVI1_05	EVI1
EVI1_06	EVI1
FAC1_01	BPTF
FOX_Q2	FOX
FOXD3_01	FOXD3
FOXJ2_01	FOXJ2
FOXJ2_02	FOXJ2
FOXM1_01	FOXM1
FOXO1_01	FOXO1
FOXO1_02	FOXO1
FOXO3_01	FOXO3
FOXO4_01	FOXO4
FOXO4_02	FOXO4
FOXP1_01	FOXP1
FOXP3_Q4	FOXP3
FREAC2_01	FOXF2
FREAC3_01	FOXC1
FREAC4_01	FOXD1
FREAC7_01	FOXL1
FXR_Q3	NR1H4
GABP_B	GABP
GATA_C	GATA
GATA_Q6	GATA

GATA1_01	GATA1
GATA1_02	GATA1
GATA1_03	GATA1
GATA1_04	GATA1
GATA1_05	GATA1
GATA2_01	GATA2
GATA3_01	GATA3
GATA4_Q3	GATA4
GATA6_01	GATA6
GCM_Q2	GCM
GCNF_01	NR6A1
GFI1_01	GFI1
GFI1_Q6	GFI1
GFI1B_01	GFI1B
GKLF_02	KLF4
GKLF_Q4	KLF4
GLI_Q2	GLI
GR_01	NR3C1
GR_Q6	NR3C1
GR_Q6_01	NR3C1
GRE_C	NR3C1
GZF1_01	GZF1
HAND1E47_01	TCF3:HAND1
HEB_Q6	TCF12
HELIOSA_01	IKZF2
HELIOSA_02	IKZF2
HES1_Q2	HES1
HFH1_01	FOXQ1
HFH3_01	FOXI1
HFH4_01	FOXF1:FOXJ1
HFH8_01	FOXF1
HIC1_02	HIC1
HIC1_03	HIC1
HIF1_Q3	HIF1A
HIF1_Q5	HIF1A
HLF_01	HLF
HMEF2_Q6	MEF2A
HMGIY_Q3	HMGA
HMGIY_Q6	HMGA
HMX1_01	HMX3
HNF1_01	HNF1A
HNF1_C	HNF1A
HNF1_Q6	HNF1
HNF1_Q6_01	HNF1
HNF3_Q6	FOXA
HNF3_Q6_01	FOXA
HNF3ALPHA_Q6	FOXA
HNF3B_01	FOXA
HNF4_01	HNF4A
HNF4_01_B	HNF4A
HNF4_DR1_Q3	HNF4A
HNF4_Q6	HNF4:NR2F
HNF4_Q6_01	HNF4A
HNF4_Q6_02	HNF4A
HNF4_Q6_03	HNF4A
HNF4ALPHA_Q6	HNF4A

HNF6_Q6	ONECUT
HOX13_01	HOXA5
HOXA3_01	HOXA3
HOXA4_Q2	HOXA4
HSF_Q6	HSF1
HSF1_01	HSF1
HSF1_Q6	HSF1
HSF2_01	HSF2
HTF_01	XBP1
ICSBP_Q6	IRF8
IK1_01	IKZF1
IK2_01	IKZF1
IK3_01	IKZF1
IPF1_Q4	PDX1
IPF1_Q4_01	PDX1
IRF_Q6	IRF
IRF_Q6_01	IRF
IRF1_01	IRF1
IRF1_Q6	IRF1
IRF2_01	IRF2
IRF7_01	IRF7
KROX_Q6	EGR
LEF1_Q2	LEF1:TCF
LEF1_Q2_01	LEF1
LEF1TCF1_Q4	LEF1:TCF
LFA1_Q6	ITGAL
LHX3_01	LHX3
LMO2COM_01	LMO2
LMO2COM_02	LMO2
LUN1_01	TOPORS
LYF1_01	IKZF1
MAF_Q6	MAF
MAF_Q6_01	MAF:NFE2:BACH
MAX_01	MAX
MAZ_Q6	KIF22:MAZ
MAZR_01	PATZ1
MEF2_01	MEF2A
MEF2_02	MEF2A
MEF2_03	MEF2A
MEF2_04	MEF2A
MEF2_Q6_01	MEF2
MEIS1_01	MEIS1
MEIS1AHOXA9_01	HOXA9:MEIS1
MEIS1BHOXA9_02	HOXA9:MEIS1
MMEF2_Q6	MEF2A
MRF2_01	ARID5B
MSX1_01	MSX1
MTF1_Q4	MTF1
MYB_Q3	MYB
MYB_Q5_01	MYB
MYB_Q6	MYB
MYC_Q2	MYC:MAX
MYCMAX_01	MYC:MAX
MYCMAX_02	MYC:MAX
MYCMAX_03	MYC:MAX
MYCMAX_B	MYC:MAX

MYOD_01	MYOD1
MYOD_Q6	MYOD1
MYOGENIN_Q6	MYOG
MYOGNF1_01	NFI
MZF1_01	MZF1
MZF1_02	MZF1
NANOG_01	NANOG
NANOG_02	NANOG
NCX_01	TLX2
NERF_Q2	ELF2
NF1_Q6	NFI
NF1_Q6_01	NFI
NFAT_Q4_01	NFATC
NFAT_Q6	NFATC
NFE2_01	NFE2
NFKAPPAB_01	NFKB:RELA
NFKAPPAB50_01	NFKB1
NFKAPPAB65_01	RELA
NFKB_C	NFKB
NFKB_Q6	NFKB1
NFKB_Q6_01	NFKB:RELA
NFY_01	NFY
NFY_C	NFY
NFY_Q6	NFY
NFY_Q6_01	NFY
NGFIC_01	EGR4
NKX22_01	NKX2-2
NKX25_01	NKX2-5
NKX25_02	NKX2-5
NKX25_Q5	NKX2-5
NKX3A_01	NKX3-1
NKX61_01	NKX6-1
NKX62_Q2	NKX6-2
NMYC_01	MYCN
NRF1_Q6	NRF1
NRF2_Q4	NFE2L2
NRSE_B	REST
NRSF_01	REST
NRSF_Q4	REST
OCT_C	OCT:POU2F
OCT_Q6	OCT:POU2F
OCT1_01	POU2F1
OCT1_02	POU2F1
OCT1_03	POU2F1
OCT1_04	POU2F1
OCT1_05	POU2F1
OCT1_06	POU2F1
OCT1_07	POU2F1
OCT1_B	POU2F1
OCT1_Q5_01	POU2F1
OCT1_Q6	POU2F1
OCT4_01	OCT4
OCT4_02	OCT4
OSF2_Q6	RUNX2
P300_01	EP300
P53_01	TP53

P53_02	TP53
P53_DECAMER_Q2	TP53
PAX_Q6	PAX
PAX1_B	PAX1
PAX2_01	PAX2
PAX2_02	PAX2
PAX3_01	PAX3
PAX3_B	PAX3
PAX4_01	PAX4
PAX4_02	PAX4
PAX4_03	PAX4
PAX4_04	PAX4
PAX5_01	PAX5
PAX5_02	PAX5
PAX6_01	PAX6
PAX6_Q2	PAX6
PAX8_01	PAX8
PAX8_B	PAX8
PBX_Q3	PBX
PBX1_01	PBX1
PBX1_02	PBX1
PBX1_03	PBX1
PEA3_Q6	ETV4
PEBP_Q6	CBFB:RUNX
PIT1_Q6	POU1F1
PITX2_Q2	PITX2
PLZF_02	ZBTB16
POU1F1_Q6	POU1F1
POU3F2_01	POU3F2
POU3F2_02	POU3F2
POU5F1_01	OCT4
POU6F1_01	POU6F1
PPAR_DR1_Q2	PPAR
PPARA_01	PPARA:RXRA
PPARA_02	PPARA:RXRA
PPARG_01	PPARG
PPARG_02	PPARG
PPARG_03	PPARG
PR_01	PGR
PR_02	PGR
PR_Q2	NR3C1:PGR
PU1_Q6	SPI1
PXR_Q2	NR1
RFX_Q6	RFX
RFX1_01	RFX1
RFX1_02	RFX1
ROAZ_01	ZNF423
RORA1_01	RORA
RORA2_01	RORA
RP58_01	ZNF238
RREB1_01	RREB1
RSRFC4_01	MEF2A
RSRFC4_Q2	MEF2A
S8_01	PRRX2
SF1_Q6	NR5A1
SMAD_Q6	SMAD

SMAD_Q6_01	SMAD
SMAD3_Q6	SMAD3
SMAD4_Q6	SMAD4
SOX_Q6	SOX:SRY
SOX2_Q6	SOX2
SOX5_01	SOX5
SOX9_B1	SOX3
SP1_01	SP1
SP1_Q2_01	SP1
SP1_Q4_01	SP1
SP1_Q6	SP1
SP1_Q6_01	SP1
SP3_Q3	SP3
SPZ1_01	SPZ1
SREBP_Q3	SREBF
SREBP1_01	SREBF1
SREBP1_02	SREBF1
SREBP1_Q6	SREBF1
SRF_01	SRF
SRF_C	SRF
SRF_Q4	SRF
SRF_Q5_01	SRF
SRF_Q5_02	SRF
SRF_Q6	SRF
SRY_01	SRY
SRY_02	SRY
STAT_01	STAT
STAT_Q6	STAT
STAT1_01	STAT1
STAT1_02	STAT1
STAT1_03	STAT1
STAT3_01	STAT3
STAT3_02	STAT3
STAT4_01	STAT4
STAT5A_01	STAT5A
STAT5A_02	STAT5A
STAT5A_03	STAT5A
STAT5A_04	STAT5A
STAT5B_01	STAT5B
STAT6_01	STAT6
STAT6_02	STAT6
STRA13_01	BHLHB
T3R_Q6	RAR:RXR:THR
TAL1_Q6	TAL1
TAL1ALPHAE47_01	TAL1:TCF3
TAL1BETAE47_01	TAL1:TCF3
TAL1BETAITF2_01	TAL1:TCF4
TATA_01	TBP
TATA_C	TBP
TAXCREB_01	CREB1
TAXCREB_02	CREB1
TBP_01	TBP
TBP_Q6	TBP
TBX5_01	TBX5
TBX5_02	TBX5
TBX5_Q5	TBX5

TCF11_01	NFE2L1
TCF11MAFG_01	NFE2L1
TCF4_Q5	TCF7L2
TEF_Q6	TEF
TEF1_Q6	TEAD1
TEL2_Q6	ETV7
TFE_Q6	MDFI:MITF:TFE
TFIIA_Q6	GTF2A
TFIII_Q6	GTF2I
TGIF_01	TGIF1
TITF1_Q3	NKX2-1
TST1_01	POU3F1
TTF1_Q6	NKX2-1
USF_01	USF1
USF_02	USF1
USF_C	USF1
USF_Q6	USF
USF_Q6_01	USF
USF2_Q6	USF2
VDR_Q3	VDR
VDR_Q6	VDR
WHN_B	FOXN1
XBP1_01	XBP1
YY1_01	YY1
YY1_02	YY1
YY1_Q6	YY1
YY1_Q6_02	YY1
ZF5_01	ZFP161
ZF5_B	ZFP161
ZIC1_01	ZIC1
ZIC2_01	ZIC2
ZIC3_01	ZIC3
ZID_01	ZBTB6

Zusammenfassung

Eine der wichtigsten ungelösten Fragen in der Molekularbiologie ist wie die Zellen eines höheren Organismus mit einer identischen genetischen Information in eine grosse Vielfalt von unterschiedlichen Zelltypen differenzieren. Die Zelldifferenzierung wird durch die zellspezifische Regulierung von Genen gesteuert. Dabei wird nur ein bestimmter Teil der genetischen Information aktiviert, sodass nur die benötigte RNA und Proteine produziert werden. Eine der wichtigsten Komponenten in der Genregulation sind die Transkriptionsfaktoren. Diese DNA-bindende Proteine können die Expression ihrer Zielgene aktivieren bzw. unterdrücken. Die Transkriptionsfaktoren agieren jedoch selten einzeln sondern wirken mit anderen Transkriptionsfaktoren zusammen, um eine hohe kombinatorische Vielfältigkeit und Gewebespezifität zu erreichen. Das kombinatorische Zusammenspiel zwischen Transkriptionsfaktoren experimentell nachzuweisen ist jedoch sehr kompliziert und sogar für viele Proteine nicht durchführbar.

Das Ziel dieser Arbeit ist das kombinatorische Auftreten von Paaren von Transkriptionsfaktoren in den regulatorischen Abschnitten der DNA vorherzusagen. Als Information werden dafür die zugrundeliegende DNA-Sequenz und die berechnete Bindungsaffinität der Faktoren zu der Sequenz verwendet. Für die Vorhersage wird jeder Transkriptionsfaktor als eine Liste der regulatorischen Abschnitte, die gemäß der Bindungsaffinität geordnet ist, repräsentiert. Mit Hilfe dieser Listen können rangbasierte Maße für die Assoziationsbestimmung verwendet werden. Diese Darstellung als geordnete Liste und die Anwendung der rangbasierten Maße wird in dem ersten Teil der Arbeit (Kapitel 2 und 3) diskutiert.

Im zweiten Teil der Arbeit wird dann das gemeinsame Vorkommen der Transkriptionsfaktorpaare in gewebe- und zelltypspezifischer Weise vorhergesagt. Durch die Gewebebzw. Zelltypinformation wird in die Analyse eine dritte Dimension eingeführt. Um die assoziierten Transkriptionsfaktorpaare in den gewebespezifischen Promotoren zu finden, werden die dreidimensionale Kontingenztabellen und die dazugehörigen statistische Tests verwendet (Kapitel 4). Die Ergebnisse aus diesem Kapitel wurden im Januar 2012 veröffentlicht (Myšičková and Vingron, 2012). Im Kapitel 5 werden dann die

assoziierten Transkriptionsfaktorpaare in den zelltypspezifisch offenen regulatorischen Abschnitten, die mit Hilfe der DNase I-Verdauung und darauffolgender Sequenzierung (DNase-seq) bestimmt wurden, vorhergesagt. Die neuartigen DNase-seq Daten erfordern eine neue Methode um die assoziierten Transkriptionsfaktorpaare zu finden. Dazu wurde ein Verhältnis von zwei p-Werten des exakten Fisher-Tests definiert: der erste p-Wert ist abgleitet von der Kontingenztabelle auf den zelltypspezifisch offenen Abschnitten und der zweite p-Wert ist abgeleitet von der Kontingenztabelle auf den ubiquitär offenen Abschnitten. Transkriptionsfaktorpaare mit einem signifikant hohen Verhältnis treten dann viel wahrscheinlicher gemeinsam in den zelltypspezifischen Abschnitten auf, als in den ubiquitär offenen Abschnitten. Die vorhergesagten gewebe- und zelltypspezifischen Transkriptionsfaktorpaare stimmen mit Vorhersagen von anderen computergestützten Methoden überein. Zudem sind sie angereichert mit bekannten Protein-Protein-Interaktionen. Darüber hinaus ist der Großteil der vorhergesagten Transkriptionsfaktoren in dem jeweiligen Gewebe bzw. Zelltyp exprimiert und etwa ein Drittel der Faktoren hat eine bekannte Funktion in dem jeweiligen Gewebe oder Zelltyp. Das deutet darauf hin, dass die vorhergesagten Transkriptionsfaktorpaare tatsächlich eine regulatorische Funktion in dem jeweiligen Zelltyp haben.

Zusammengefasst liefert diese Arbeit neue Erkenntnisse über die kombinatorische Genregulation durch Transkriptionsfaktoren und präsentiert neue Anwendung der rangbasierten Methoden um assoziierte Transkriptionsfaktorpaare in gewebespezifischer Weise vorherzusagen.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Alena van Bömmel

Berlin, Juli 2014