Computational Characterization of Genome-wide DNA-binding Profiles

Christian Rödelsperger

Dezember 2011

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

Gutachter:
Prof. Dr. Martin Vingron
PD. Dr. Peter N. Robinson

1. Referent: Prof. Dr. Martin Vingron
2. Referent: PD. Dr. Peter N. Robinson

Tag der Promotion: 19.10.2011

Preface

The work and data that is presented in this thesis is part of a collaborative project that is funded by the Berlin Center for Regenerative Therapies. A number of people have contributed to this work and for clarity I will now mention the individual contributions. Stefan Mundlos, Peter N. Robinson and Jochen Hecht designed this project with the purpose of studying bone development using ChIP-seq in a chicken model. Jochen Hecht and Asita Stiege established the ChIP-seq protocol and together with Daniel Ibrahim, Hendrikje Hein, and Catrin Janetzky carried out the immunoprecipitations and sequencing. Peter Krawitz was responsible for the data processing that involved base calling and basic quality control. Daniel Ibrahim contributed to the analysis on the Hox proteins identifying the Q317K mutant to be related to Pitx1 and Obox family members. Sebastian Köhler and Sebastian Bauer carried out the computation of the Gene Ontology similarity data and random walk distances that I used for the target gene assignments in chapter 5. The results for the EMSA experiments that are shown in chapter three has been carried out by Asita Stiege.

The work on target gene assignment that is presented in chapter 5 has been published in *Nucleic Acids Research* [1]. All the remaining methods, data and the experimental results will be partially be included in future publications by Ibrahim *et al.* and Hein *et al.*.

Acknowledgements

I would like to thank both my supervisors Peter N. Robinson and Martin Vingron for giving me the opportunity to do this thesis as a joint project between experimental biology and bioinformatics. I would further like to thank Stefan Mundlos for giving me the opportunity to collaborate with a number of people in his department, especially with Jochen Hecht and his group members Hendrike Hein, Daniel Ibrahim, Ulrike Wille, and Asita Stiege. The close interaction and exchange with the people who generated the data, worked on experimental validation, and with whom I had numerous discussions on the interpretation of results has constantly motivated me to gain more insights into the underlying biology and to improve the analysis methods.

A lot of thanks goes to Peter Krawitz, Marcel H. Schulz, Pablo Villavicencio-Lorini, Pia Kuss, Morgane Thomas-Chollier and all the people who attended my seminar talks for sharing the interest in gene regulation, for all the helpful suggestions and inspiring discussions. The aforementioned group partially overlaps with the people with whom I shared offices and whom I have to thank for the nice working atmosphere, these people are Akdes Serin, Alena Mysickova, Jonathan Goeke, Paz Polak, Rosa Karlic, Federico Squartini, Sebastian Bauer, Sebastian Köhler, Marten Jäger and Gao Guo. The rest of credit goes to all the members of Computational Biology Department and the Development and Disease Group at the MPI for Molecular Genetics and the Institute for Medical Genetics at the Charité and finally to those people I forgot to mention.

Contents

Chapter 1

Introduction

1.1 Gene expression-profiling for identification of regulatory interactions

The question how complex organisms with hundreds of tissues evolve from a single cell progenitor is inherently linked to the question how genes are activated and repressed. Most vertebrate genomes contain roughly 20,000-30,000 protein-coding genes. Approximately 10% of these genes encode transcription factors (TFs) that are able to regulate the expression level of a number of target genes by binding to specific recognition sites in their promoter regions and interacting with the core transcription initiation machinery. The complete set of genes within a genome can be seen as a gene-regulatory network that coordinates proliferation and differentiation processes throughout development of an organism.

For more than a decade, gene expression-profiling using microarrays and more recently new sequencing techniques, has been the method of choice in order to elucidate gene-regulatory networks. Typically one component of the network is modified and the network response is measured as the difference in gene expression relative to a control experiment, for instance cells can be treated by drugs, hormones, or growth factors. Alternatively genes can be transiently or stably knocked out [2, 3] or overexpressed. Genes that change their expression level as a consequence of variations in the activity of a given transcription factor are reasonable candidates for being target genes for this transcription factor.

However gene expression-profiling does not allow primary from secondary effects to be distinguished, i.e. an observed upregulation of a putative target gene may be due to direct or indirect regulation. Likewise it may be, that the transcription factor represses a different gene, which normally acts as a transcription activator. Thus the observed upregulation would just be a secondary effect. Even other mechanisms are conceivable that are based on more complex network models.

One necessary but not sufficient criterion for direct regulation is the binding of a candidate transcription factor in the proximal promoter region of a target gene. Although numerous studies have shown the regulatory role of distant enhancer and silencer regions that may act over hundreds of kilobases, studying the proximal promoter region of a single gene may reveal important features of the underlying regulatory mechanism. This is true for the experimental analysis of single promoters [4, 5] but holds also on genome-wide analyses that correlate gene expression patterns with motifs in proximal promoter regions [6]. In the next two sections we will describe computational models for transcription factor binding as well as experimental approaches for validation. Both strategies have

been successfully applied to either study global associations between transcription factors and their predicted target genes or to closely investigate the regulatory mechanisms within single promoters.

Genome-wide computational approaches have the drawback of a high number of false positive predictions and reporter gene and binding assays are not scalable to large numbers of candidate sequences. This emphasizes the importance of experimental approaches for genome-wide mapping of protein-DNA interactions.

1.2 Computational models for transcription factor binding

Transcription factors are DNA binding proteins that are able to regulate the gene expression of target genes by interacting with the core transcription factor machinery that is comprised of transcription initiation and elongation factors such as RNA polymerase II and Taf proteins. The transcription factor-DNA interaction is sequence-dependent. Transcription factors have recognition motifs, that is a set of similar DNA sequences typically of length 5 to 15 bp. Depending on the actual DNA sequence the binding affinity of the transcription factor to the DNA can be modulated. Binding motifs can be determined by gel shift experiments coupled with mutational analysis [7] but also by high-throughput assays that use protein binding on double stranded DNA immobilized on a microarray [8] and "Systematic Evolution of Ligands by Exponential Enrichment" (SELEX) assays [9].

If only a handful of bound sites is known, the motif is conveniently represented as the set of bound sequences and screening a candidate promoter sequence for putative binding sites, equals to the detection of one of these sequences.

However, usually the number of bound sequences is too large to follow this approach and a more common way is to align the sequences and construct a position specific frequency matrix (PSFM). In a PSFM the rows correspond to the position within the motif and columns correspond to the four possible DNA bases. An entry in the PSFM indicates the observed frequency of the corresponding base at this position. The probability of a query sequence with respect to the PSFM model can be calculated as the product of the matrix base frequencies of the individual positions in the sequence. Typically a cutoff is defined to determine, whether a certain probability value is regarded as a match or not[10].

Alternative approaches have chosen a biophysical approach to motivate their computational model [11, 12]. The combination of a binding sequence S and a transcription factor TF can exist in two states, the bound and unbound state, respectively. This can be formalized as

$$TF + S \underset{}{\overset{K_d(S)}{\rightleftharpoons}} TF \cdot S \tag{1.1}$$

whereby $K_d(S)$ denotes the sequence specific equilibrium dissociation constant, which is a measure of the affinity of the transcription factor to the site S and is directly related to the ΔG Gibbs free energy of binding per mole [11]. If we consider the case that the TF and the sequence exist at given concentrations, $K_d(S)$ determines the ratio of unbound vs. bound sites at the equilibrium.

$$K_d(S) = \frac{[TF] + [S]}{[TF \cdot S]} \tag{1.2}$$

The parentheses refer to the concentrations of transcription factor and sequence that are either bound or unbound. The fraction of sequences that are bound by the TF can be calculated as the ratio of bound sequences relative to all sequences.

$$N(S) = \frac{[TF \cdot S]}{[TF \cdot S] + [S]} = \frac{[TF]}{[TF] + K_d(S)} = \frac{K_d(S)[TF]}{1 + K_d(S)[TF]} \tag{1.3}$$

Typically TFs bind various degenerate motifs, thus the model has to take into account changes in $K_d(S)$ with respect to changes in the sequence. This can be done by setting the free energy $\Delta G = 0$ for the optimal binding site and subsequently quantifying the energy for an arbitrary sequence S relative to the optimal site S_0.

$$K(S) = K(S_0)e^{-\frac{\Delta\Delta G}{RT}} \tag{1.4}$$

$\Delta\Delta G$ refers to the change in free energy or equivalently the mismatch energy, R is the gas constant and T is the temperature. The mismatch energy can be modeled by summing up the probabilities in the TF-specific matrix M over all positions in the sequence S.

$$\frac{\Delta\Delta G}{RT} = \frac{1}{\lambda} \sum_{i=1} \sum_{\alpha=A,C,G,T} S_i^{\alpha} log(\frac{m_{i,max}}{m_{i,\alpha}} b_{i,\alpha}) \tag{1.5}$$

$S_i^{\alpha} = 1$ if the base at position i in sequence S is equal to α and 0 otherwise, $b_{i,\alpha}$ denotes the ratio of the background frequencies of $m_{i,max}$ and $m_{i,\alpha}$. If $S = S_0$, $\frac{\Delta\Delta G}{RT} = 0$ and it follows $K(S) = K(S_0)e^0$. The λ parameter is used to scale the mismatch energies in units of thermal energy. A second parameter $R_0 = K(S_0)[TF]$ summarizes the TF specific binding affinity for the optimal sequence S_0 and its concentration. By running this calculation on a contiguous sequence such as the promoter region, this model can be used to calculate an affinity measure which can be interpreted as the expected number of TFs bound to the promoter region [12].

1.3 Experimental validation of target genes

In order to validate a candidate target that was identified by gene expression-profiling after transcription factor knockouts [2, 3], luciferase assays can be employed to confirm the differential expression, as well as to further map the position of a protein-DNA contact in the promoter region [4, 5, 13]. In a first step, candidate constructs of varying size and location in the presumptive promoter region are cloned in front of a luciferase reporter gene with minimal promoter sequence and transfected with a second expression vector carrying the candidate regulatory TF into a cell line. The minimal promoter sequence alone is not enough to activate the luciferase which could be quantified as a luminescense signal. Only if the cloned promoter sequence has regulatory activity, a signal can be detected. The construct resulting in the strongest luciferase activity is then taken for further analysis. The size of the resulting promoter region may still be in the range of 100 bp. To further pin down the actual binding site of the transcription factor, the sequence can be searched for matches with a known consensus motif of which a mutation should influence the activity of the luciferase. However in some cases, the factor does not directly bind DNA but rather interacts with a different transcription factor which possibly binds a totally different motif [4]. The described experiments require a substantial amount of work and their success cannot be guaranteed. Thus in order to further screen lists of differentially expressed genes for direct interactions on a genome-wide scale, a different kind of assay, chromatin-immunoprecipitation (ChIP), can be used if specific antibodies for the transcription factor are available. ChIP was originally applied on a single gene level, however it may also be combined with high-throughput quantification methods

such as microarrays (ChIP-chip) and massively parallel sequencing (ChIP-seq). This constitutes a powerful approach to study protein-DNA interactions on a genome-wide scale.

In the following we will describe the experimental protocol and basic computational analysis of ChIP-seq experiments in order to better understand the subsequent discussion of ChIP-seq related questions and topics that are under current research.

1.4 Chromatin-Immunoprecipitation

Chromatin-immunoprecipitation is a method for selective enrichment of protein-bound DNA. The addition of formaldehyde [14] covalently links protein-protein as well as protein-DNA interactions (cross-linking). In a cross-linked state, a number of experimental steps can be performed to separate bound from unbound DNA without dissociation of the proteins. The second step after cross-linking consists in fragmentation by ultrasound and size selection on a gel. This results in a set of fragments with a defined length that still contain bound as well as unbound fragments. In order to separate these two classes from each other, the bound fraction is captured by protein-specific antibodies that are attached to magnetic beads. The cross-linking can be reversed and captured DNA may be detected by qPCR, hybridization to microarrays, or more recently by high-throughput sequencing.

1.4.1 Alignment

Second generation sequencing platforms such as the Illumina GA and the ABI SOLiD platforms produce millions of reads for a single experiment. The massive amount of sequence data as well as the decreased read length (36-100 bp *vs.* 700-800 bp) in comparison with traditional Sanger sequencing has posed new challenges for alignment algorithms. The classical Smith-Waterman algorithm for local alignment as well as existing seed and extend approaches did not scale up to the amount of sequencing data produced by the second generation sequencing platforms. One of the earliest short read alignment programs was MAQ [15]. Similar to the Eland algorithm, which is one part of the Illumina GA pipeline, MAQ relies on the so called 'pigeon hole' lemma, which states that in any pair of sequences within distance k (number of mismatches), of any partition of the first sequence into $k+1$ parts, at least one of the parts must be found in the other sequences. MAQ builds multiple hashtables to index the reads and then scans the reference genome with this index. It guarantees to find all matches within distance 2 in the first 28 high quality bases of the read. Whereas the 'pigeon hole' lemma defines a property on arbitrary partitions of a sequence, the q-gram lemma defines a lower boundary on the number of common substrings between two sequences [16, 17]. The q-gram lemma states that two sequences of length n with Hamming distance k share at least $t = n + 1 - (k+1)q$ common substrings of length q, so-called q-grams. The program RazorS employs q-gram counting techniques using an extension of this lemma enabling it to operate on Hamming distances (mismatches only) as well as edit distances (mismatches, insertions and deletions) [18].

A third strategy is based on a Burrows-Wheeler transformed reference sequence index. The Burrows-Wheeler transformation constructs a matrix of all lexicographically ordered cyclic rotations of the reference sequence. This matrix has a property called 'last first (LF) mapping', which means that the i^{th} occurrence of character x in the last column corresponds to the i^{th} occurrence of x in the first column. This property can be used to iteratively search for all occurrences of a query sequence in the reference [19]. This approach can be extended to take mismatches into account.

The program Bowtie was the first to apply the Burrows-Wheeler transformation to the mapping of second generation sequencing reads.

1.4.2 Peak-calling

Within a single cell, a genomic location can either be bound or not and one binding event can give rise to at most one immunprecipitated fragment. However for a tissue sample or cell lines, signals from multiple cells are integrated and the number of reads that map at a unique position, may be taken as an indicator of the fraction of cells in which a particular location is bound by the transcription factor of interest. Thus, regions that are enriched in number of reads, so called 'peaks' are candidate regions for binding events. If the reads were randomly distributed over the genome, then the probability of observing a peak with a read depth of at least H can be computed by a Poisson distribution [20, 21]

$$1 - \sum_{k=0}^{H-1} \frac{e^{-\lambda}\lambda^k}{k!},$$

whereby λ is the expected number of reads in a given region.

Unfortunately control experiments show read distributions that are unlikely to be explained by a uniform distribution [22] and in addition show enrichments near transcription start sites [23]. This deviation from the uniform distribution may be explained by local variations in chromatin structure but also by regions that do not allow unambiguous alignment of short reads [23].

In a recent evaluation of peak-calling algorithms, multiple programs were evaluated in terms of total number and similarity of peak calls, agreement with qPCR validation experiments, and motif enrichments [24]. Among the best programs with respect to the qPCR data were PeakSeq [23] and MACS [25]. PeakSeq uses a two-pass approach for scoring ChIP-seq signal relative to controls. In the first phase, reads are extended into 3' direction (∼200bp) to model DNA fragment sizes and base pair coverage profiles are computed. Locally enriched regions are detected by simulating random read placement in 1Mb intervals taking the mappability of reads into account. In a second phase the ChIP-seq signal is normalized against the control data by computing a linear regression between read counts in 10kb windows. The slope α from the linear regression is subsequently used as a scaling factor for the control. In this process, the enriched regions from the first phase are excluded from the linear regression in order to avoid overestimation of α. A candidate region r is then judged by comparing the observed read count in the ChIP-seq sample n_r^{sample} against the scaled read count $k = \alpha \times N_r^{\text{control}}$ of the control sample. Under the null hypothesis that the region is not enriched, the reads should be distributed evenly between ChIP-seq sample and control with P=0.5. Thus the probability that we observe k or fewer reads in the control would be binomially distributed.

$$F(k, n, P) = \sum_{j=0}^{|k|} \left(\begin{array}{c} n \\ j \end{array} \right) P^j (1-P)^{n-j},$$

whereby $n = n_r^{\text{sample}} + k$.

The program MACS [25] also implements a two-step approach that in a first phase evaluates windows of a fixed size and defines enriched clusters as windows with a minimal fold enrichment relative to a random genome-wide read distribution. MACS samples 1000 of these clusters and

aligns the clusters by their midpoint and estimates the fragment size d as the distance between the summits of the separate read coverage distributions on the watson and crick strands. All reads are then shifted by $d/2$ in 3' direction to the most probable site of protein-DNA interaction. In a second run, MACS slides $2d$ windows across the genome and calculates p-values using a Poisson distribution with local λ values for the ChIP sample and the control in order to model effects due to local chromatin structures or copy number variations. MACS also accounts for amplification artifacts by removing duplicate reads that have been sequenced repeatedly. An empirical false discovery rate (FDR) is defined for each peak as the fraction of control peaks over ChIP-seq peaks at a given p-value after sample swap.

A number of alternative approaches exist that exploit motif occurrence or strand-specific scoring of peaks, however in the evaluation by Wilbanks and Facciotti [24], most peak-calling algorithms performed equally well.

1.4.3 Comparison between ChIP-chip and ChIP-seq

In theory ChIP-seq has a number of advantages over ChIP-chip, that include no limitation on the genome size and by-passing artifacts due to cross-hybridization on the array platform. So far, there has been limited research on the comparison between ChIP-chip and ChIP-seq approaches. Euskirchen and Rozowsky *et al.* compared Stat1 binding in 1% of the genome as defined by the ENCODE project. They used ChIP-chip as well as ChIP-PET, which is a variant of ChIP-seq to identify target regions for Stat1. They found that the highest ranked targets agree well across the two platforms, but at lower ranks the level of overlap decreases. They validated regions that were identified by only one of the methods and observed that 6 out of 10 regions, that were missed by the ChIP-chip approach showed interspersed repetitive sequences that were partially removed from the array design. 7 out of 15 confirmed ChIP-chip targets that were missed by ChIP-PET were found near ChIP-PET targets but were located at the shoulders relative to the site showing the highest signal. Four additional targets that were unique to ChIP-chip overlapped weak ChIP-PET targets and should likely be detected with increasing sequencing depth [26].

In a more recent study protein DNA-interactions of Smad 1/5 (same antibody) and Smad 4 were mapped in mouse embryonic stem cells by ChIP-chip. The microarray which was used in this study, tiled mouse promoter regions from 5.5 kb upstream to 2.5kb downstream of the TSS. In order to confirm the identified regions, immunoprecipitated DNA was sequenced on a Illumina GA II. 62.5% of the 562 SMAD 1/5 bound regions and 40.5% of 2518 Smad4 associated regions could be confirmed by ChIP-seq [27], but no closer investigation was carried out to elucidate the reasons for this discrepancy. Thus it is not clear whether the results of ChIP-seq have a clear advantage over the ChIP-chip data. Euskirchen and Rozowsky *et al.* proposed that both methods should be rather regarded as complementary approaches [26].

1.5 Related questions

Albeit its appeal for biologists, the integration of TF binding and expression data in order to understand gene-regulatory networks is a challenging task; and a number of related questions should be asked that allow a better understanding of the general framework for transcriptional regulation within cells.

How many binding sites does a transcription factor have?

Based on the results of gene expression-profiling experiments following transcription factor knockout or overexpression [2], one would expect that the number of targets for one individual transcription factor would lie in the range of a few hundred genes. This is because gene expression-profiling studies identified in the range of hundreds to thousand of differentially expressed genes and a fraction of these genes are likely to be differentially expressed due to secondary effects. Thus, the differentially expressed genes form a superset of the direct targets.

Although loss of transcription factor activity may be compensated by paralogous genes, and only rarely loss of complete sets of paralogous groups are analyzed [28], the biologically estimated number of target genes stands in strong contrast to the ten thousands of target regions that are detected in ChIP-seq experiments [29, 23]. Rozowsky *et al.* looked at the number of enriched regions relative to controls as a function of sequencing depth (number of reads). They concluded for RNApolII binding, that the number of enriched regions approaches $\sim$ 25,000 target regions. However for the transcription factor STAT1 they found only a slight trend towards saturation at a level around 30,000 regions for a maximal read number of 22.5 million.

What proportion of the binding sites is functional?

If transcription factors really bind in the range of 30,000 regions but the number of target genes lies in the range of hundreds, then it is unlikely that every binding event is functional in a way that it directly affects the transcription of a target gene [30]. Even in the 120 Mb size genome of *Drosophila melanogaster* ChIP-chip experiments identified up to 10,000 bound regions in the early embryo for a single transcription factor [31]. Further analysis showed that regions with a strong signal are usually found in non-coding intergenic parts of the genome and overlap well-known regulatory elements. In contrast regions with weaker ChIP-chip signal are enriched in protein-coding sequences. It has been suggested that these regions might be active at later developmental stages but also that a substantial proportion is not functional at all [31].

What percentage of binding events is conserved across species?

The high degree of conservation in non-coding parts of genomes has brought up the idea that many of these highly conserved sequences may act as developmental enhancers that may act over several hundreds of kilobases on their target genes [32]. This implies that a number of regulatory sequences and consequently binding events are shared across species. Using ChIP-chip for four different transcription factors Odom and Dowell *et al.* showed that binding profiles between human and mouse hepatocytes differed significantly across the two species. Roughly between 41% and 89% of binding events were species specific. Although one can argue that such differences on the molecular level in cross-species comparisons may be due to environmental, lifestyle, or nutritional aspects, Wilson and Barbosa-Morais *et al.* found a way to circumvent this argument. They repeated these experiments in mice carrying a copy of human chromosome 21 and they were able to show that almost all human binding events were recapitulated across the entire chromosome in the mouse nucleus [33].

To what degree does sequence variation affect transcription factor binding?

One particular strength of ChIP-assays is that they allow the identification of *in vivo* motifs. Many transcription factors are known to bind in complexes, e.g. Hoxa9 and Pbx1 or Fos and Jun which bind as heterodimers. Binding as a complex may change the sequence preference of a transcription factor and thus the *in vivo* motifs may differ from motifs that have been identified in *in vitro* assays like protein-binding microarray or SELEX [8, 9]. One interesting question would be how changes in the sequence will affect the *in vivo* binding behavior of the transcription factors. Recently Kasowski and Grubert *et al.* compared ChIP-seq profiles for NfκB and RNA polymerase II in ten HapMap individuals and correlated differences in binding with SNPs and structural variations [34]. They were able to show that changes in binding across individuals follow a pattern as would be expected by SNPs that either increase or decrease the binding of the transcription factor. Similar trends could be observed for SNPs that affect Stat1 binding, a known cofactor of NfκB [34].

1.6 Objective and outline of the thesis

Objective

The goal of this thesis is to develop computational methods that allow the interpretation of ChIP-seq data with respect to binding affinities of the immunoprecipitated transcription factor and to develop methods that can be used to integrate and compare genome-wide binding profiles from multiple experiments. In a second step these binding events have to be classified into functional and non-functional binding events and have to be assigned to their target genes. Ultimately these methods serve to extend gene-regulatory networks by introducing links to target genes, that integrate combinatorial binding events on *cis*-acting regulatory regions that act over large intergenic distances.

Outline

The first chapter has explained the use of gene expression-profiling in the investigation of gene-regulatory networks and has introduced computational models for transcription factor binding site prediction. In addition, the first chapter motivated the importance of mapping transcription factor-DNA interactions on a genome-wide scale in order to extend our network models and explained the basic steps of the ChIP-seq experiment and primary data analysis that are used as a starting point throughout this thesis.

Chapter 2-5 are each divided into a separate Introduction, Methods, Results and Discussion section in order to keep the individual topics in a rather compact form.

The second chapter will investigate more closely the quantitative aspect of the ChIP-seq data. This serves to assess how well the ChIP-seq data actually reflects the sequence dependent binding affinity of a transcription factor and to gain insights into biologically processes that may explain the sequence independent variation of the ChIP-seq data.

The third chapter shows how ChIP-seq experiments can be used to compare different binding profiles from wildtype and mutant proteins and how motif analysis can be used to detect changes in recognition motifs as well as cofactor interactions. The last section of chapter three describes an analysis of experimental data for a putative cofactor of the HOXD13 protein and shows how peak

overlaps can be used as a similarity measure between experiments in order to confirm the results of the motif analysis.

The fourth chapter describes a method that can be used to analyze ChIP-seq profiles for multiple transcription factors in order to rank putative *cis*-regulatory sequences with respect to their impact on gene expression. The method divides the genome-wide binding data into classes of putative *cis*-regulatory sequences that are defined by a pattern of colocalized binding events and uses an approach that is based on gene set enrichment analysis (GSEA) and takes into account biases that arise from the analysis of the genome-wide distribution of binding events.

The fifth chapter aims to assign functional binding events to their target genes. Based on combined data sets for transcription factor binding and differentially expressed genes, we will describe a machine learning approach that integrates data from various biological data sources and is able to predict the correct target genes for a majority of binding events.

The final chapter summarizes and discusses the methods and results of this thesis. This includes the discussion of the potential and limits of the used methodology and suggestion of further extensions. The last chapter serves also to discuss the results of the methods in the light of ChIP-seq experiments in general and to what degree the methodology can be applied to other experimental data sets.

Chapter 2

Quantification of ChIP-seq signals

2.1 Introduction

In the last couple of years ChIP-seq and ChIP-chip have greatly enhanced our ability to elucidate the complex processes that control gene regulation and to identify the roles that different transcription factors play within these processes. Although the identification of functional binding events is one of the most challenging task in the analysis of genome-wide binding events, it might be very helpful to first investigate more closely the mechanisms by which transcription factors bind DNA with different affinities. The binding affinity of a transcription factor to DNA can be decomposed into sequence independent as well as sequence-specific components. Sequence independent components are the TF concentration within the nucleus and the sequence independent affinity of the TF to DNA that is mediated by protein contacts to the phosphate backbone of the DNA. Protein contacts with the nucleotide bases confer the sequence specificity of the interaction.

A number of models exist to predict transcription factor binding sites (TFBS) that basically score sequences based on their similarity to known consensus motifs or position specific weight matrices [10, 12] that have been compiled from *in vitro* experiments such as SELEX or gel shift experiments or sets of *in vivo* binding sites from literature. These approaches are extremely useful to rank sequences with respect to their affinity towards one specific transcription factor, but it remains difficult to compare TFs with respect to one sequence without the knowledge of TF concentrations and their unspecific binding affinities. Even though empirical genome-wide motif distributions can be used to transform binding affinities to p-values, which can be compared across different TFs [35], this approach is blind to changes of TF abundances across cell types.

The genome-wide prediction of TFBS is further complicated by the lack of information about the impact of DNA accessibility due to cell type specific chromatin-modifications, and the synergistic effect between TFs and cofactors. TFs may recognize distinct motifs depending on whether they bind as monomers or as heterogeneous complexes.

Since high-throughput ChIP assays facilitate the mapping of *in vivo* recognition motifs on a genome-wide scale, this data might be useful to distinguish different classes of binding events and to quantify the difference in binding affinities with respect to the motif composition of the corresponding sequences.

Motifs within ChIP-seq peaks exhibit a strong positional bias towards the summit (Figure 2.1A, [25]). This enrichment is due to the distribution of immunoprecipitated fragments around the true binding site and is often used to assess the quality of the data and peak calling method [24]. It

has to be noted that motifs for putative cofactors may show distinct enrichment at neighboring locations (Figure 2.1B). Widely used *de novo* motif discovery tools like MEME [36], Weeder [37], and Trawler [38] assume that binding motifs are equally likely to occur at all positions in each sequence. The problem of detecting motifs with positional bias with respect to a certain location has been mostly studied with approaches that focused on motifs upstream of the transcription start sites [39, 40]. However, many of these tools do not scale up to the great amount of sequences obtained by a single ChIP-seq experiment. More recently a method has been developed that was explicitly designed for the analysis of ChIP-seq data [41].

ChIP-seq signals are additionally ranked by their signal intensity, that may be defined as the number of reads, fold change relative to control, or p-value. Both signals, positional bias and signal intensity, can be visualized together with a predicted binding affinity on a three-dimensional space with the peak ranking on the x-axis, the position relative to the summit on the y-axis and the affinity with respect to a certain motif is encoded by the different colors. Figure 2.1C shows the TRAP affinity values [12] for the top-ranked motif for ChIP-seq data of the transcription factor Runx2 identified by the Amadeus software [40]. Other recent *de novo* motif finders explicitly use the ranking of the ChIP-seq peaks to identify motifs that are enriched among the strongest peaks [42].

In contrast to the aforementioned methods, our interest is not to use the ranking, to identify the best motifs, but rather to develop a model, that can be used to explain the signal obtained by the ChIP-seq experiment. A previous study compared binding affinities for position specific weight matrices with experimental binding strength as determined by ChIP-chip on yeast [12]. Most of the experimental data sets showed correlations between $0.3 < r < 0.8$. Modeling the ChIP-chip and ChIP-seq signal using motif sets can provide insights into the general mechanisms of protein-DNA interactions that involve binding of large multi-protein complexes as well as variability in chromatin accessibility.

In the following section we will investigate for the TF Runx2 as an example, to what degree the ChIP-seq signal can be explained by the DNA binding affinities of Runx2 and putative cofactors. The second part of this chapter analyzes data for the TF Stat1 and correlates different classes of binding events with methylation data for histone H3K4.

2.2 Methods

2.2.1 Alignment and peak-calling

We filtered 28.9 million raw Illlumina GAII reads for mean phred quality score above 30 and removed all but one copy of multiple reads that have the identical sequence in order to avoid read stacking artifacts. 13.5 million (82%) of the remaining non-redundant 16.3 million reads could be uniquely aligned with up to three mismatches to the chicken genome (WUGSC 1.1/galGal3) using the Bowtie aligner (version 0.12.5 with -v 3 -m 1 options). Of the 42 million raw reads for the input control, 25.6 (51%) million could be aligned uniquely to the chicken genome.

We used the program MACS [25] to identify enriched regions relative to the input controls. We ran the program with default parameters for the Runx2 ChIP-seq sample together with the input control and extracted all peaks with p-value < 0.001. MACS reports for each peak the number of reads, false discovery rate, fold change *vs.* control and the summit position. The summit denotes the position within the peak that shows the highest read coverage after 3' shifting [25]. This can

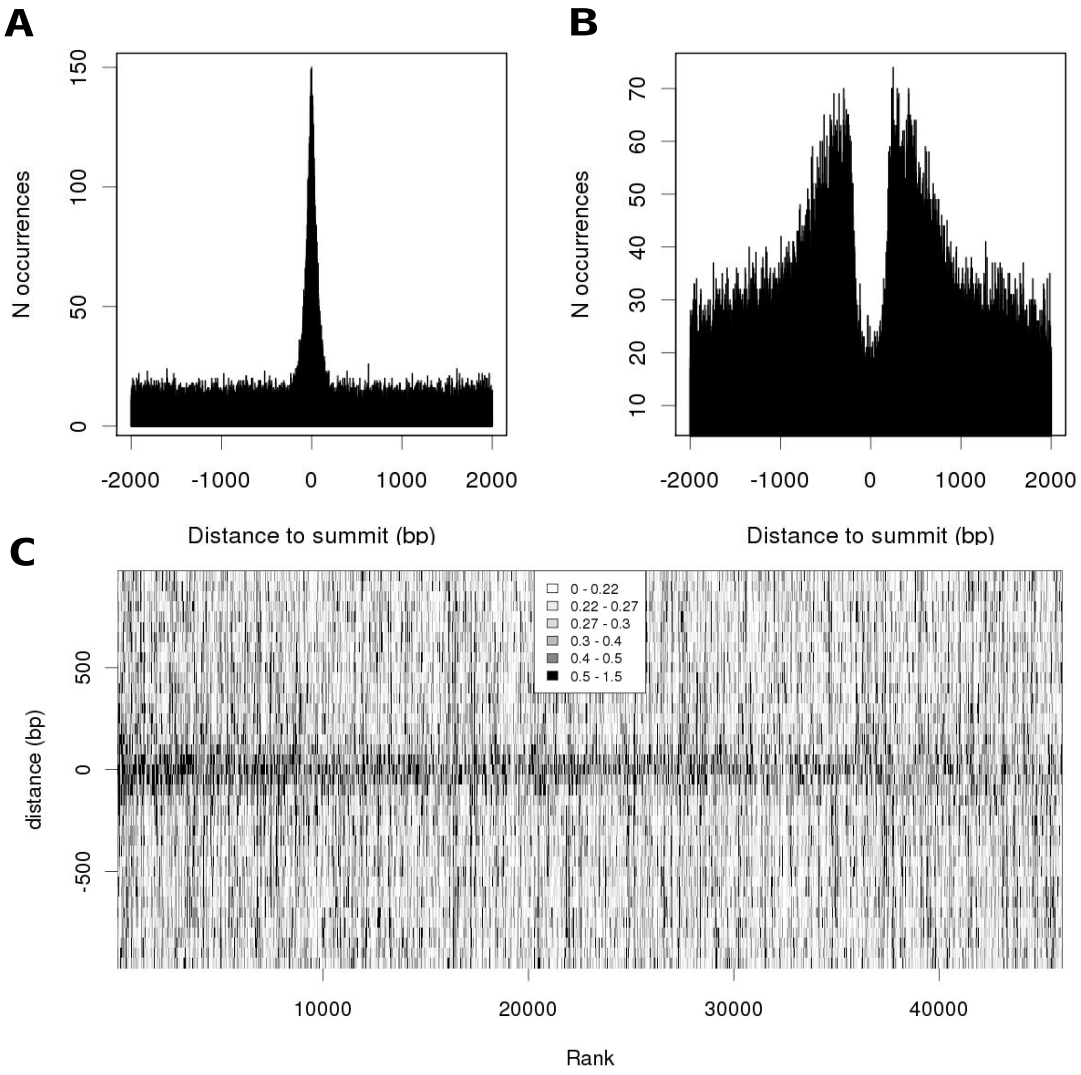


Figure 2.1: **A)** Positional bias of the top ranked motif identifed by Amadeus [40] within the Runx2 ChIP-seq peaks. **B)** Positional bias of a GC rich motif identifed by Amadeus in the sequences of Runx2 peaks. **C)** TRAP map of 45,984 sequences with Runx2 peaks. Peaks are ranked according to highest fold change between Runx2 and input sample. Using a sliding window of 100bp with 50 bp step width, binding affinities for the Runx2 matrix model was determined and binding affinities were computed using TRAP [12]. In addition to the positional enrichment of high affinites around the summit (position 0 on y-axis) the highest ranked peaks show a cluster of high affinity windows whereas the peaks with lowest ranks show a depletion of high affinity windows.

be interpreted as the most likely binding site. All further analyses use the summit coordinates and the fold change values.

Data for Stat1 binding [21] was downloaded from Gene Expression Omnibus (GSE15353) and processed in the same way. H3K4me1 and H3K4me3 data [43] was downloaded and aligned to the human reference genome (hg19).

2.2.2 Motif analysis

From the total peak set, we sampled 3000 ChIP-seq summits as foreground set and chose 12,000 random genomic locations as background set. Not all ChIP-seq locations were used because the software Amadeus has a limit on the number of sequences (32,000) that can be analyzed. We fixed sequence length to the exact location ± 500bp. We applied the Amadeus program [40] to find motifs that are either enriched in the foreground set relative to the background and to detect motifs that show a positional bias relative to the background. For the overrepresentation Amadeus divides the data into three bins according to GC-content and uses a hypergeometric test to compute p-values. For the positional bias, it divides each sequence into 10 windows and uses a χ^2-test for p-value calculation.

2.2.3 Prediction of binding affinities with TRAP

We used the program TRAP [12] to predict binding affinities for sequences with ChIP-seq peaks. The count matrices for TRAP were created from the overrepresented k-mers that were reported by the Amadeus software. The matrices were converted to TRAP format using a unified GC content of 0.35 and λ-value were optimized by varying λ in the range between 0.1 and 20 and choosing the best value as the one that yielded the highest Pearson correlation between the $\log_2$ fold change and the affinity of the ±100 bp region around the summit. We also tested larger or smaller sequence windows around the summits, but they showed lower correlation coefficients. We used all affinities for the individual matrices and fitted a linear regression model in order to generate a combined affinity prediction (Figure 2.2 I).

2.2.4 Predicting affinities from k-mer counts

In order to directly model the observed signal in the ChIP-seq experiment we model the total signal as a linear combination of the contributions from single k-mers μ.

In a preprocessing step, we exhaustively counted all k-mers with length $2 \leq k \leq 7$ and up to three consecutive 'N' characters between any single pair of consecutive positions in the k-mer. The 'N' characters match any type of nucleotide and model motif positions with low information content. To account for potentially variable motif distributions around the summit (Figure 2.1 A and B), we considered all occurences of a k-mer μ with maximal distance $d = \{50, 150, 450\}$ bp to the position of the summit as individual predictors. Thus for each k-mer, we save the number of occurences in a window defined by a maximum of d bp around the summit. We filtered out any combination of k-mer and d pairs with Pearson's correlation coefficient $-0.04 < r < 0.04$. In addition to the k-mers we used the number of peaks in 10, 20, 50 kb distance as additional candidate predictor. The total set of exhaustively computed k-mers and number of neighboring peaks is denoted by M, of which a subset M' denotes the predictor set that is used by the linear regression model. This transforms to a linear regression model in the form

$$I_p(M', \beta) = \beta + \sum_{\mu \in M'} \alpha_\mu N_{\mu,p} \tag{2.1}$$

I_p denotes the logarithm base two of the ratio between the number of reads in the ChIP-seq and input sample. $N_{\mu,p}$ equals the number of occurences of k-mer μ in the peak, whereby M' is the subset of predictors that is used by the model. This set is iteratively extended to maximize correlation with the observed signal. α_μ is the weight of the predictor μ. The model parameter β represents the baseline signal when no motif is present in the peak region. This is analogous to the approach used by Bussemaker *et al.* who used a linear regression model to identify motifs in upstream regions that are predictive for differences in gene expression levels across two cell types [44]. Optimal model parameters can be calculated by minimizing the square error between model prediction and observed value for all n peaks

$$E(M', \beta) = \sum_{i=1}^{n} (I^i_{\text{observed}} - I^i_p(M', \beta))^2 \tag{2.2}$$

From the complete table of k-mer distance pairs and number of neighboring peaks for all peaks, we greedily selected the predictor for which the predicted value yielded the best improvement in correlation coefficient between the model predictions and the observed ChIP-seq signal for all peaks. This procedure was repeated N=50 times.

2.3 Results

2.3.1 Overrepresented motifs show weak correlation with ChIP-seq signal

For identification of Runx2 binding sites, chicken micromass (chMM) cell cultures from embryonic limb buds were infected with an avian specific RCAS virus carrying a copy of the chicken Runx2 gene with an attached triple-flag tag that is recognized by a specific antibody. On day nine after infection chromatin immunoprecipitation was performed as in [45] and sequenced on a Illumina GA II. We called 45,984 peaks using MACS [25] with a tolerant p-value cutoff of 0.001 in order to detect strongly bound as well as weakly bound regions. We used the Amadeus software to search for overrepresented motifs as well as motifs showing a positional bias with respect to random genomic sequences of same length and GC-content [40]. The top ranked motif (Figure 2.2 A) matches the known consensus motif for Runx2 [46], which is identical to a recently identified SELEX motif for Runx3 [9] $\frac{C}{T}G\frac{C}{T}GGT\frac{C}{T}$ (reverse complement: $\frac{A}{G}ACC\frac{A}{G}C\frac{A}{G}$). For all eight identified motifs (Figure 2.2 A-H), we used the reported k-mers to built position specific weight matrices and computed TRAP affinity values with optimized lambda values for each peak. Figure 2.2 shows the correlations of the predicted affinities with the $\log_2$ fold change between the ChIP-seq reads relative to a control sample of input DNA. The correlation coefficients range from r=0.363 for the top ranked matrix (Figure 2.2 A) to r=0.148 in the worst case (Figure 2.2 F). Using a linear combination of all eight affinities values only results in a slight improvement to r=0.378 (Figure 2.2 I).

2.3.2 k-mer counts improve the predictability of ChIP-seq signal

The above described method represents a two step strategy that first detects overrepresented motifs within the ChIP-seq peaks and optimizes the affinity predictions for these motifs to match the

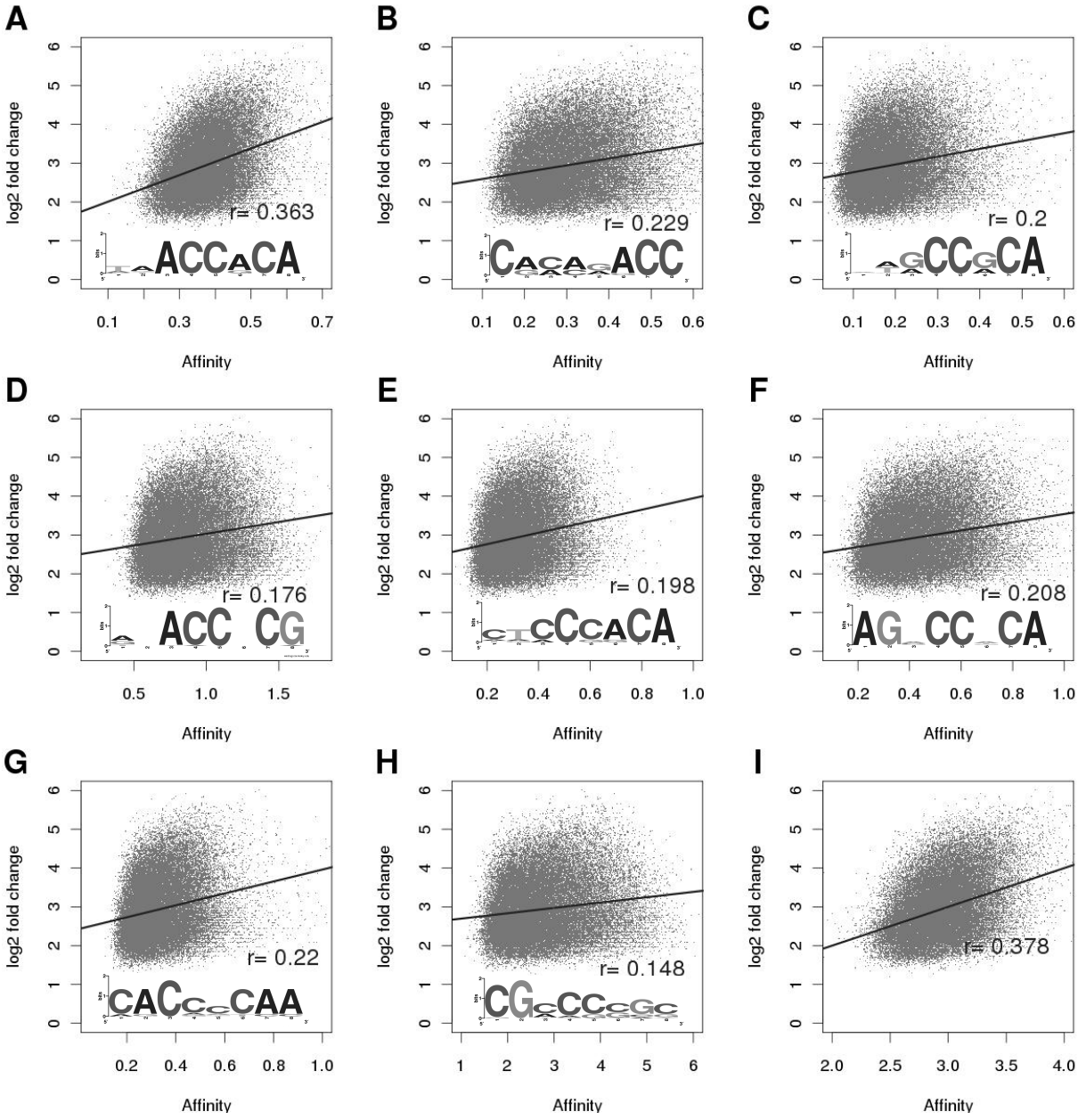


Figure 2.2: **A-H)** Correlations between TRAP affinities and $\log_2$ ChIP-seq signal for eight motifs identified by Amadeus [40]. Beneath the scatter plots sequence logos for the eight motifs are shown (Weblogo). **I)** Correlation of binding affinites for all motifs, predicted by a linear regression model using all eight motifs.

observed signal. In order to combine motif detection and prediction of binding affinities for the ChIP-seq peaks in one method, we used a linear regression model based on k-mer counts in variable sequence windows around the summit to directly correlate the observed counts to the obtained signal. Our method iteratively searches k-mers that maximally improve the correlation coefficients between the values that are predicted by the linear model and the ChIP-seq signal. By using only five k-mers, the correlation coefficient (r=0.392) is already better than for the combined TRAP matrices (Figure 2.2 I). After 50 iterations the correlation coefficient converges to a value of r=0.512 (Figure 2.3 A,B). We repeated the analysis by splitting the Runx2 peaks into two thirds training data, which we used to build the model and one third validation data, that we later used to assess the quality of prediction on an unknown data set. This procedure showed the same correlation coefficient for the validation set as for the training set. This suggests that the model generalizes to independent Runx2 ChIP-seq data.

The similiarities for 46 non-redundant k-mers with $k \geq 4$ are displayed in Figure 2.4. As in the case for the Amadeus motifs, most of the motifs show a strong match with the known Runx2 consensus motif [46].

Initially we used k-mers with $2 < k \leq 7$, but we found that 6-mers and 7-mers are almost never picked up by our approach among the top-ranked motifs. This is probably due to the low expected value of 6-mers and 7-mers in the small region around the summit (100bp, 300bp, 900bp).

In some genomic regions ChIP-seq peaks show a strong tendency towards clustering. Chromatin interactions not only between regulatory regions and promoters but also across regulatory regions have been observed by 3C assays [47]. We argued that interactions between regulatory regions may influence the signal observed in a ChIP-seq experiment and included also the number of peaks in up to 100kb distance as potential predictors of ChIP-seq signal. Interestingly the number of peaks in up to 25kb distance shows a Pearson correlation of $r = 0.11$ with the ChIP-seq signal and is picked up as the 11th best predictor.

To further investigate why even for larger number of iterations the correlations do not substantially increase above r=0.51, we considered the possibility that fitting always to the complete data set would disregard motifs for colocalizing factors. Such colocalizing factors might increase the Runx2 binding probability by stabilizing the protein-DNA interaction or decrease the binding probability by competitive occupancy of neighboring binding sites. We therefore separated the peaks into three classes, that are defined by the difference between predicted and observed value; a class of high fold change peaks where this difference is more than one standard deviation higher than the average difference, and analogously a low fold change class, the peaks with absolute difference smaller than one standard deviation constitute the 'average fold change' class. We ran the linear model separately on each of these classes with 50 iterations, which resulted in correlation coefficents r=0.81 for the high fold change class, r=0.78 for the low fold change class, and r=0.74 for the average fold change class. Interestingly, the top twenty predictors, which are enough to predict the signal of the three peak classes separately with $r > 0.7$, have 15 predictors in common for all three classes (Figure 2.3 C). Using only the 15 common predictors for each class separately accurately models the observed signal with $r > 0.68$ (Figure 2.3 D). Comparison of the weights that were assigned to the predictors showed that there exist two motifs that contribute negatively, but the majority of k-mers has a positive effect on the predicted value (Figure 2.3 E). However, the greatest difference is observed for the β parameter which models the baseline signal when no motif is present in the peak region. This indicates that a large fraction of the observed signal is not related to the sequence, at least in the range of motif space that we analyzed. It cannot be excluded that

complex motif combinations could account for the strong differences in the β parameter. However searches for motif pairs using the Amadeus software detected overrepresented pairs that occur at most in up to 200 hundred peaks. Thus it is unlikely that motif combinations would account for the roughly 7000 high and low fold change classes.

We tested how the three peak classes that were defined by the difference between the observed ChIP-seq fold change and the predicted affinity, are reflected on the read count level. We found that the high fold change class shows the highest median number of reads and the low fold change class the lowest median number of reads (Figure 2.3 E). This observation could be expected since the fold change, read numbers, and p-values show strong correlations ($r > 0.7$). These results suggest that the ChIP-seq signal is a combination of sequence specific signal and a baseline signal which cannot be predicted from the k-mer composition of the underlying sequence. In addition, this baseline level is not reflected in the signal obtained by sequencing input DNA, leading to the high correlation between fold change and absolute read counts.

2.3.3 Promoters with high fold change peaks show stronger upregulation

We hypothesized that the variability in the sequence-independent baseline signal is due to the chromatin state at the binding region. It has been shown that chromatin states may affect the crosslinking and fragmentation steps in the chromatin immunoprecipitation [48, 49]. Thus the possibility exists that the three defined classes correspond to different degrees of DNA accessibility, whereby the high fold change class would be more strongly associated with open chromatin than the two other classes.

We measured gene expression levels in the cell cultures with Affymetrix microarrays and compared the cell cultures infected by Runx2 carrying RCAS virus to cells that have been infected by an empty RCAS virus. Gene expression was measured following 3, 6, and 9 days after infection (see *Methods* section 4.2.2). To investigate whether the peak class definition has any impact on gene expression, we defined bound genes on the basis of Runx2 binding in the vicinity of a TSS ($\pm$2kb). Each peak class defined a different set of bound genes and we tested what fraction of bound genes exhibit expression fold change higher or lower than a given value. Since Runx2 acts primarily as activator [46] with one potential mechanism, that is the recruitment of coactivator p300 that leads to Histone acetylation and subsequent target gene expression [50], we expected to see an enrichment of differential expression for high fold change peaks only in the case of upregulation. Figure 2.5 shows the fraction of Runx2 bound genes with differential expression. Previous studies indicated that Runx2 binding basically stays invariable across the timecourse (data not shown), thus we also looked at gene expression at previous time points because secondary effects could blur the expected pattern. Although only a small fraction of bound genes show expression fold change > 2, for all timepoints genes with high binding fold change promoter peaks showed the highest enrichment in upregulated genes (Figure 2.5 A-C). This is not the case for downregulation (Figure 2.5 D-E). For day 9, 61 (6%) of 1065 genes with high binding fold change peaks exhibit an expression fold change > 2, this is a 1.5 fold enrichment relative to the low binding fold change class ($P = 0.06$, Fisher's exact-test) and 1.7 enrichment relative to average ($P = 0.001$). For day 6, 36 (3%) genes with high binding fold change peaks are 2.3 fold enriched in > 2-fold upregulated genes relative to average ($P < 10^{-3}$) and 2.1 fold relative to low binding fold change ($P = 0.01$). Amongst these 36 genes are Runx2 itself and three known Runx2 targets: osteopontin (Spp1), bone sialoprotein (Bsp), and Panx3 and 12 other genes with known role in bone and limb development. Thus the binding strength as measured by ChIP-seq shows an effect on the expression level of the target genes, but

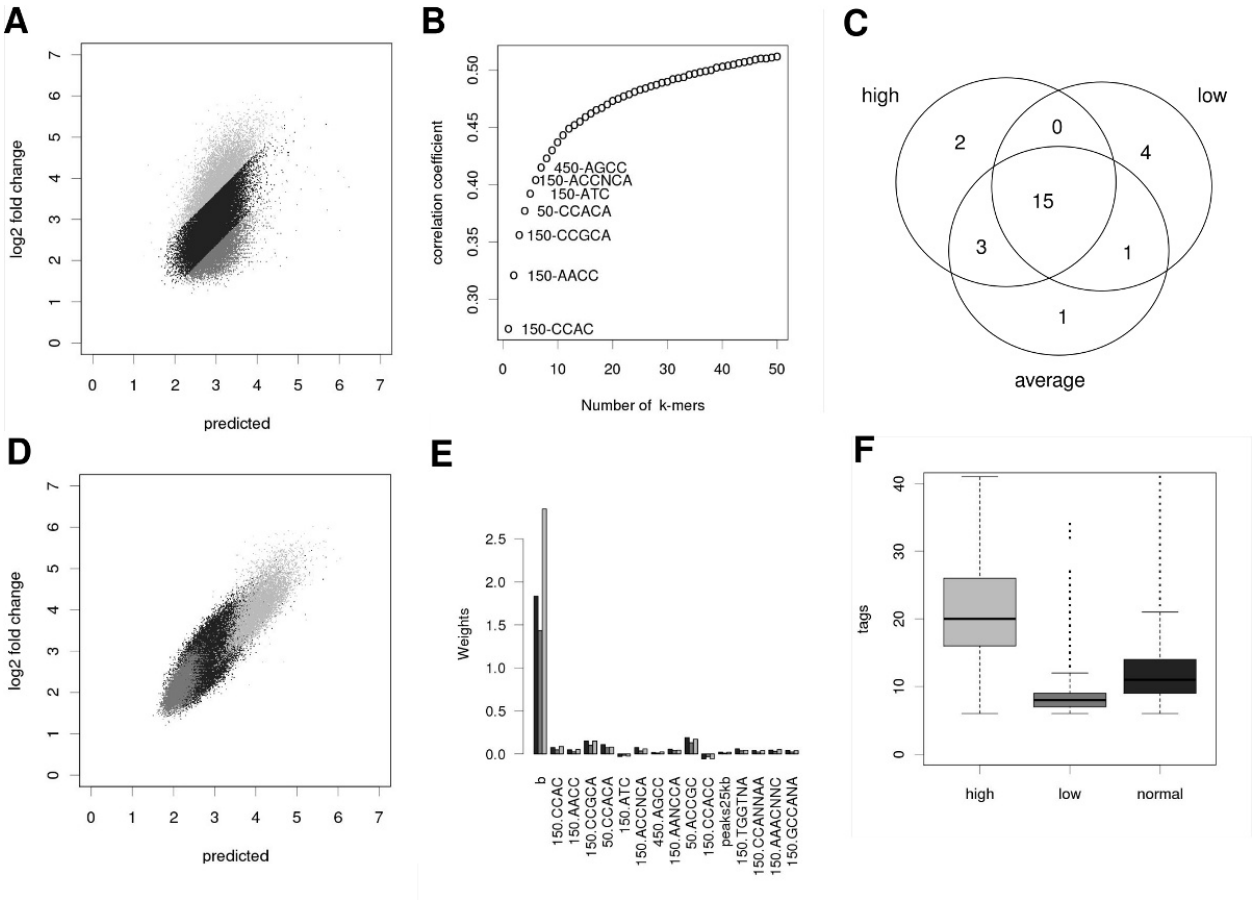


Figure 2.3: **A)** Correlation between linear regression model predictions from 50 k-mers and $\log_2$ fold change in Runx2 ChIP-seq signal. Based on the difference between prediction and observed value, three different peak classes were defined. **B)** Improvement in correlation coefficient for increasing numbers of k-mers.**C)** Overlap between top 20 k-mers for each of the three peak classes after separate model fitting. **D)** For each of the three peak classes, a linear model was fitted using the 15 common predictors. Correlations between the predictions the Runx2 ChIP-seq signal are shown for all three classes. **E)** Weights for the individual predictors in the three models. The greatest difference is observed for the β parameter which model a baseline. **F)** Peaks of the high fold change class show the highest median read number.

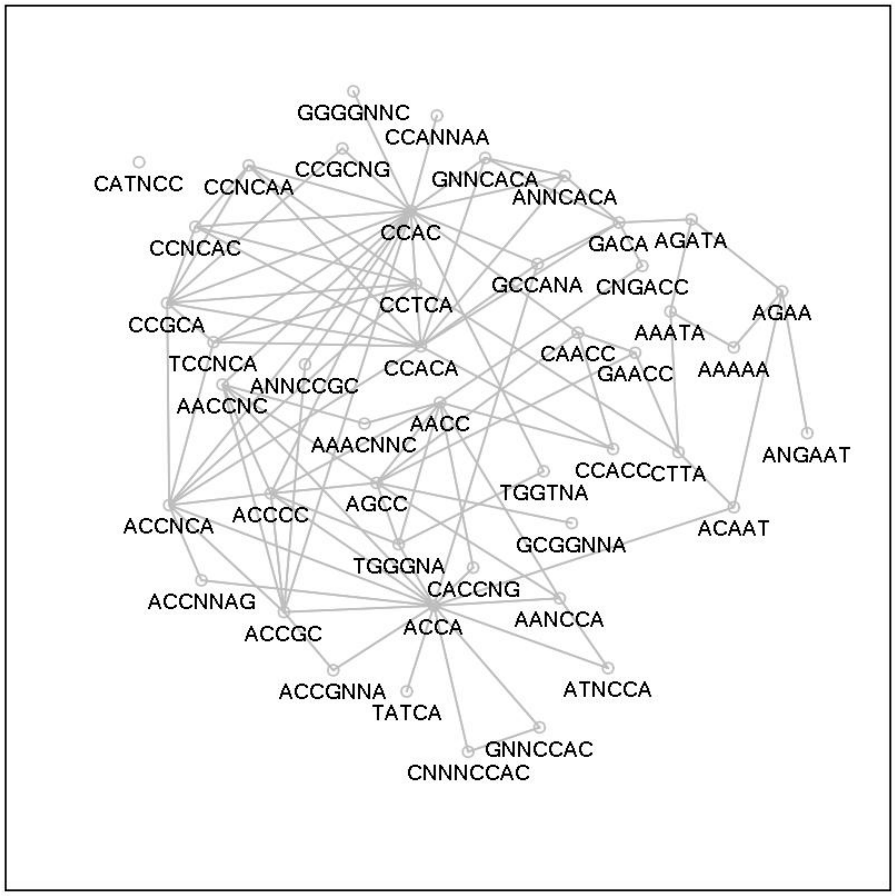


Figure 2.4: Similarities for 46 k-mers ($k \geq 4$) that are used to predict Runx2 ChIP-seq fold change for the complete peak set. An edge drawn between two k-mers if the Hamming distance $d \leq 1$. Most of the k-mers show a strong similarity to the known Runx2 consensus motif $\frac{A}{G}ACC\frac{A}{G}C\frac{A}{G}$ [46], however a number of k-mers may correspond to binding motifs for putative cofactors such as Sox proteins (ACAAT), GATA factors (AGATA), or others.

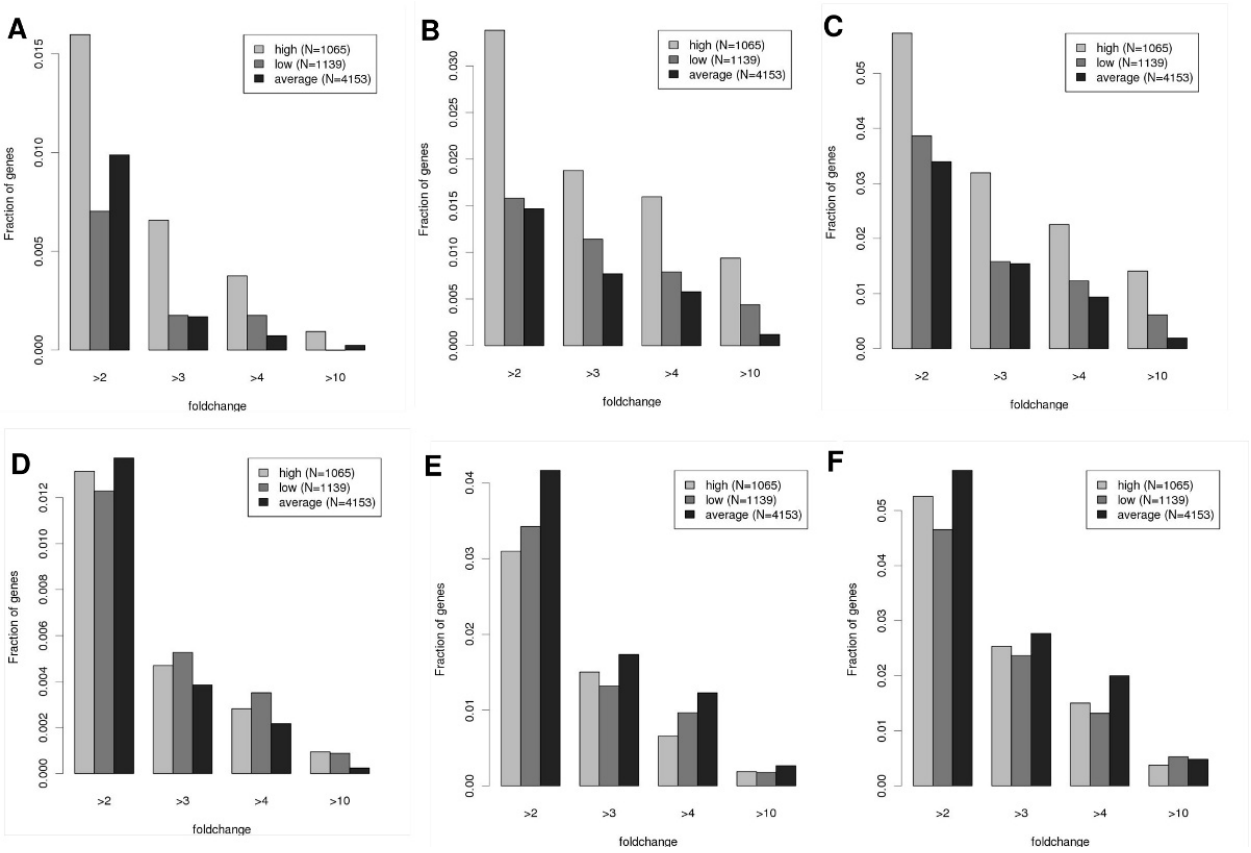


Figure 2.5: Fraction of genes that show differential expression above a certain threshold (y-axis) are shown for each of the three classes of binding fold changes, peaks with high, average, and low fold changes in ChIP-seq signal. **A-C)** Upregulated genes for 3, 6, and 9 days. **D-F)** Downregulated genes for 3, 6, and 9 days. For downregulated genes, the indicated expression fold change corresponds to a reduction.

it is yet unclear whether this effect is due to different chromatin states across the binding events. To further investigate this question we analyzed a published data set of Stat1 binding in HeLa S3 cells [21] and compared these data set with histone modification profiles [43].

2.3.4 Low fold change Stat1 peaks show increased ChIP-seq signal in unstimulated cells

Signal transducer and activator of transcription (STAT) proteins denote a family of transcription factors that are activated by extracellular signaling by growth factors and chemokines [51]. One member of this family, Stat1, is activated upon viral infection and inflammation and induces anti-proliferative and pro-apoptotic events. In response to extracellular signaling by Interferons (IFN), Stat1 is phosphorylated by receptor associated Janus tyrosine kinases, dimerizes, translocates into the nucleus, and binds specific DNA-response elements in the promoters of target genes to activate their transcription. Stat1 was one of the first transcription factors for which genome-wide binding

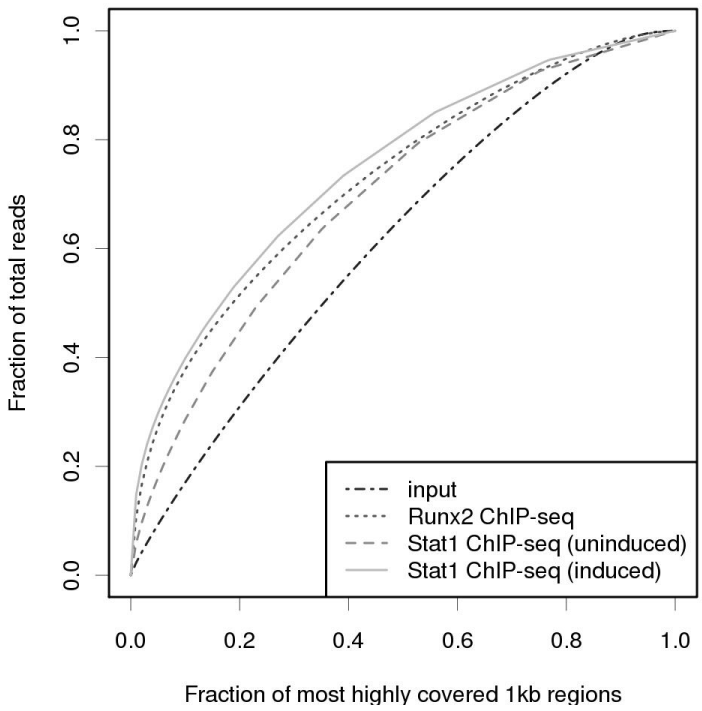


Figure 2.6: Cumulative distribution of read counts in genomic windows. The numbers of aligned reads in non-overlapping 1kb windows were counted and windows were ordered by decreasing read counts. Runx2 and Stat1 ChIP-seq experiments show the strongest enrichment in windows with high read numbers, followed by the Stat1 ChIP-seq in uninduced cells and the input sample

profiles were detected by ChIP-seq [21]. Chromatin from IFN-γ stimulated and unstimulated HeLa S3 was immunoprecipitated using Stat1 antibodies and sequenced. We reprocessed this data set and called 46,416 peaks that were enriched in the IFN-γ treated cells relative to the untreated cells. The control data from unstimulated HeLa cells represents a different type of control than the input DNA. This can be seen in the distribution of aligned read counts across the genome (Figure 2.6). Although all distributions show a clear trend towards enrichment of certain genomic sequences, the input sample exhibits the strongest similarity to a random uniform distribution. ChIP-seq experiments for Runx2 and Stat1 in stimulated HeLa S3 cells show a strong enrichment in highly covered genomic windows. The ChIP-seq experiments for Stat1 in unstimulated cells shows an intermediate level of enrichment between the input sample and the two other ChIP-seq samples.

We used the linear regression model for k-mer counts to predict the $\log_2$ binding fold changes. After 50 iterations, the correlation coefficient reached $r = 0.497$ (Figure 2.8 A), which is in the range of the correlation obtained for Runx2. When we split the peaks into two thirds training data and one third validation, application on the validation data resulted in a slight decrease of $r = 0.480$.

Sequence similarities for 44 non-redundant k-mers are visualized in Figure 2.7. After dividing the peaks into three classes and refitting the model, all predictions showed correlation coefficients $r > 0.68$ with 12 of the top 20 predictors common between all three classes (Figure 2.8 C). These 12 k-mers were sufficient to produce the degree of correlation shown on Figure 2.8 D. Like in the case of Runx2 the strongest difference in the models was observed in the β parameter (Figure 2.8 E) indicating a difference baseline signal levels. In contrast to the analysis of the Runx2 data set, the low fold change peak class did not correspond to the smallest number of reads in peaks, but was on a comparable level as the class of peaks with average fold change (Figure 2.8 E). We speculated that the low fold change in this class was due to signals that were already observed in the unstimulated cell line, which represents a different control sample than the sequenced input DNA in the case of Runx2. To further investigate this hypothesis we compared the three peak classes with respect to the read distribution in the unstimulated samples. For each distance to the summit, we calculated the observed frequency of reads in the untreated sample starting or ending at the given distance. The Stat1 peaks in the low fold change class showed the expected enrichment in ChIP-seq signal in the untreated sample (Figure 2.9 A). The high and average fold change peaks showed similar levels of ChIP-seq signal in the untreated sample. Interestingly this level was even lower than for the low fold change peaks. For the treated samples, the read distributions reproduce the previous results showing a strong enrichment for the high fold change class and comparable levels for the low and average fold change class (Figure 2.9 B).

2.3.5 Low fold change Stat1 peaks show increased levels of H3K4 methylation

Robertson *et al.* used antibodies against histone H3 trimethylation H3K4me3 and monomethylation H3K4me3 to investigate the relationship of these histone marks and transcription factor binding [43]. H3K4me3 is a mark that is predominantly found in promoters and is associated with active transcription. H3K4me1 is primarily located at intergenic sites and is considered as a mark for enhancer regions [52].

Robertson *et al.* focused on H3K4 methylation at distal and proximal binding sites. They defined roughly 70,300 Stat1 binding events after IFN-γ stimulation, of which they classified 75% as distal and 25% as proximal. In both classes, they found that roughly 80-90% of binding events were associated with at least one type of H3K4 methylation. When contrasting distal *vs.* proximal

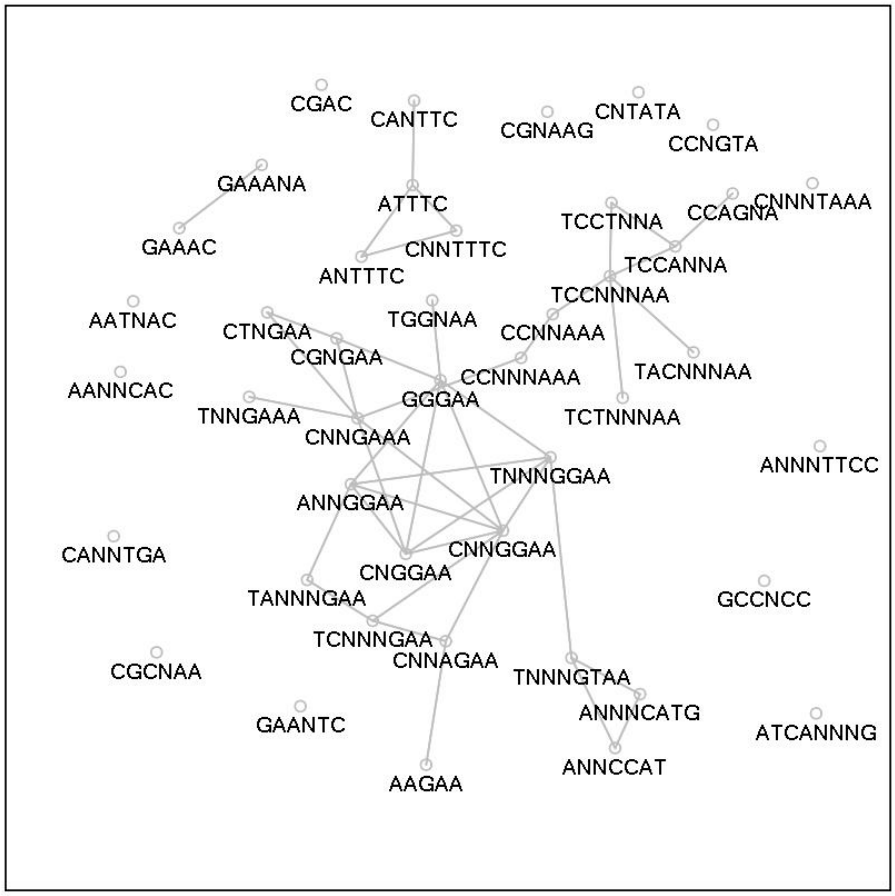


Figure 2.7: Similarities for 44 k-mers ($k \geq 4$) that are used to predict Stat1 ChIP-seq fold change for the complete peak set. An edge is drawn between two k-mers if the Hamming distance $d \leq 1$.

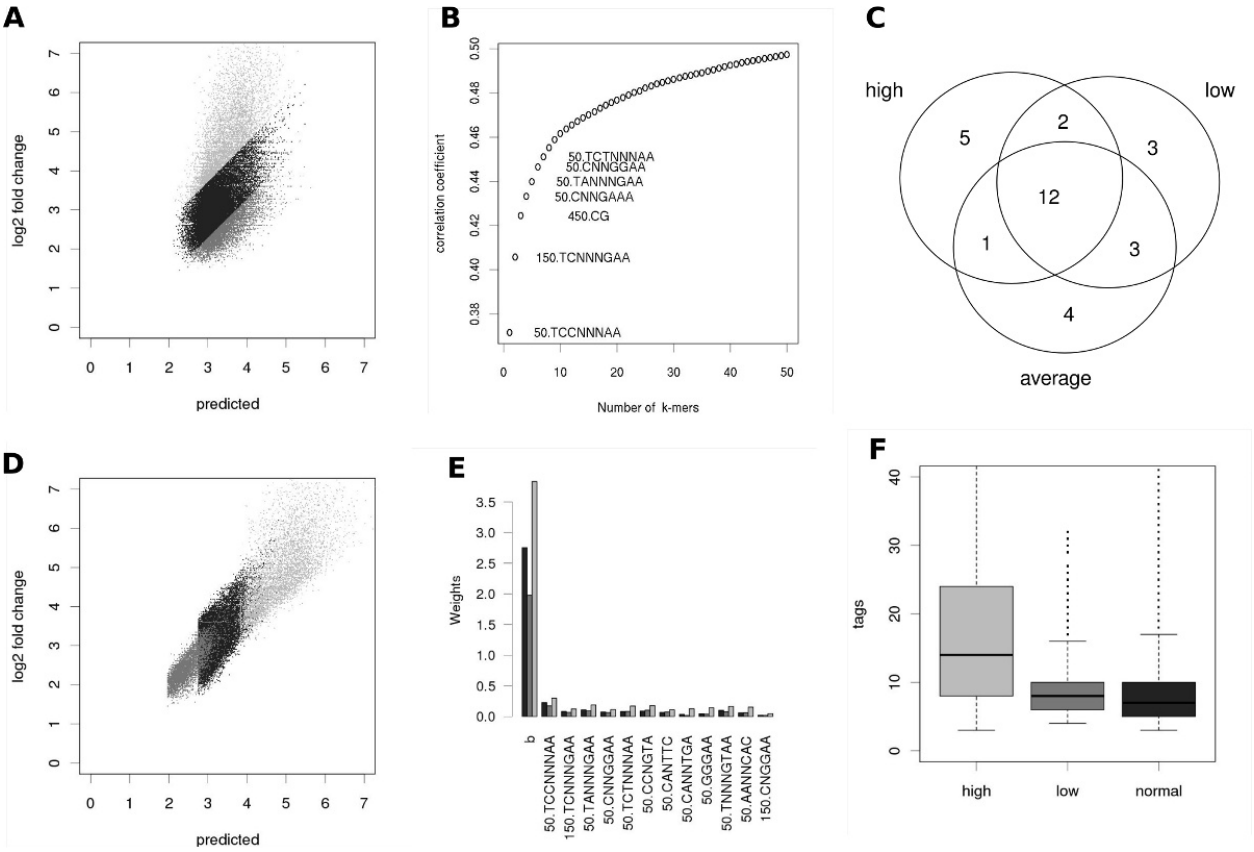


Figure 2.8: **A)** Correlation between linear regression model predictions from 50 k-mers and $\log_2$ fold change in Stat1 ChIP-seq signal. Based on the difference between prediction and observed value, three different peak classes were defined. **B)** Improvement in correlation coefficient for increasing numbers of k-mers. **C)** Overlap between top 20 k-mers for each of the three peak classes after separate model fitting. **D)** For each of the three peak classes, a linear model was fitted using the 12 common predictors. Correlations between the predictions the Stat1 ChIP-seq signal are shown for all three classes. **E)** Weights for the individual predictors in the three models. The greatest difference is observed for the β baseline parameter. **F)** Peaks of the high fold change class show the highest median read number.

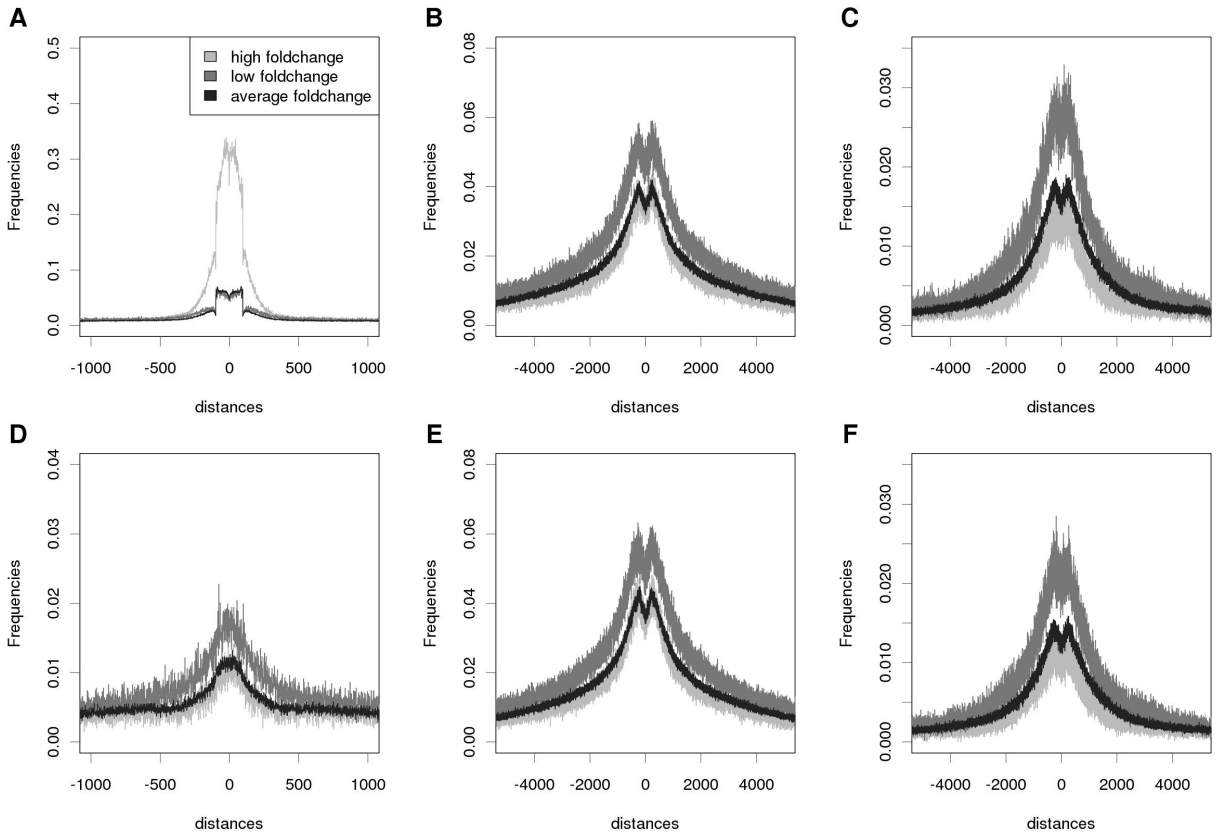


Figure 2.9: **A)** Read distance frequencies toward summits in Stat1 ChIP-seq data for stimulated HeLa S3 cells. Based on the peak classification into high, low and average fold change peaks, three distance profiles are drawn. As expected, high fold change peaks show the highest read distance frequency around their summits. **B)** Read distance frequencies around Stat1 summits for H3K4me1 ChIP-seq in stimulated cells. Here, low foldchange peaks show the strongest signal. **C)** Read distance frequencies around Stat1 summits for H3K4me3 ChIP-seq in stimulated cells. The lower panel shows read distance frequencies around Stat1 summits for Stat1 ChIP-seq reads **(D)**, H3K4me1 ChIP-seq reads **(E)**, and H3K4me3 ChIP-seq reads **(F)** in unstimulated HeLa S3 cells.

binding events, they found that distal sites showed enrichments in regions that were associated only with H3K4me1 and proximal sites showed enrichments for regions that are only associated with H3K4me3 but also for regions that show both epigenetic marks [43]. In total 25% of H3K4me1 regions were associated with Stat1 binding after stimulation.

If the high baseline level that had been observed in the low fold change Stat1 peaks was due to the chromatin state, these regions should be also associated with open chromatin marks such as H3K4me1 and H3K4me3. By looking at the read distribution of H3 methylation marks in unstimulated and stimulated cells, we found that the low fold change peaks showed an overall higher level of H3K4me1 as well as H3K4me3 signal in the unstimulated (Figure 2.9 B,C), as well as stimulated cells (Figure 2.9 E,F). Stat1 peaks that are classified as high or medium fold change peaks with respect to the uninduced Stat1 ChIP-seq data showed a median of 7 and 8 H3K4me3 reads in the $\pm$100bp region around the Stat1 summit. In contrast Stat1 peaks that were classified as low fold change peaks showed a median H3K4me3 read number of 11 reads indicating a higher H3K4me3 methylation signal at these sites ($P < 10^{-16}$, Wilcoxon test). A similar enrichment of of H3K4me1 can be observed for low fold change Stat1 peaks ($P < 10^{-16}$). This does not change between induced and uninduced state. The similarities of H3K4 methylation profiles in stimulated and unstimulated cells are in agreement with results from Robertson *et al.* who observed a high concordance between H3K4 methylation at Stat1 binding sites in both cell types [43]. This indicates that at least a fraction of the signal in ChIP-seq experiments is due to the open chromatin state of the DNA.

2.4 Discussion

In this chapter we have shown that motifs that are highly overrepresented with respect to their occurrences as well as their location in the ChIP-seq peaks, only weakly correlate with the observed ChIP-seq signal as measured as a fold change in reads between ChIP-sample and control. We have developed a linear regression model in order to predict ChIP-seq signal from word counts in the vicinity of the summits. The model greedily incorporates more words in order to maximize the correlation between observed ChIP-seq signal and predicted affinity. After $N = 50$ iterations this model yields a moderate improvement (from $r = 0.38$ to 0.52) in the correlation with the observed ChIP-seq signal when compared to the use of the matrices that were found by overrepresentation alone. We hypothesized that the observed signal is not only dependent on the sequence content but also on the chromatin structure. It has been shown that chromatin states may affect the crosslinking and fragmentation steps in the chromatin immunoprecipitation [48, 49]. Further analyses of the ChIP-seq data for Runx2 from chicken micromass cell cultures and for Stat1 binding with histone modifications of lysine K4 in histone H3 have revealed a number of findings that support this hypothesis. First, the information about the number of neighboring peaks in up to 50kb distance improves the predictions for both data sets. Second, after fitting new models on three classes of peaks, that were defined by the difference between observed *vs.* predicted fold change, among the top twenty predictors, 15 and 12 were common in all three classes for Runx2 and Stat1 respectively. Using only the common predictors raised the correlation coefficients to ~ 0.7 for each of the classes. For all three classes, the fitted models were highly similar except for a difference in baseline signal. Although it cannot be excluded that more detailed motif analysis that would also incorporate synergistic effects between the co-occurrence of a number of words could resolve this baseline signal, there seems to be a substantial part of the signal, which is sequence-independent. Finally

we found that an enrichment in read counts for unstimulated HeLa cells is responsible for the low fold change of a fraction of Stat1 peaks in stimulated HeLa cells, such peaks also show higher level of open chromatin marks H3K4me1 and H3K4me3. All these findings suggest, that there is an impact of the chromatin structure on the observed signal. But clearly, more data is needed to further support this hypothesis.

It is important to note that the Runx2 and Stat1 ChIP-seq used different control experiments. Whereas for Stat1, chromatin from IFN-γ stimulated and unstimulated cells was immunoprecipitated, in the case of Runx2 input DNA was sequenced as a control. The sequenced input DNA shows only a small number of ~ 500 locally enriched regions whereas in the unstimulated sample several thousand of peaks could be called [21, 43].

We speculated that for Runx2 binding events that show a higher than expected fold change, the ChIP-seq signal is composed of the sequence dependent predicted binding affinity and an impact of the chromatin accessibility. Given a potential knowledge about the chromatin structure, that we extracted from the ChIP-seq signal, we hypothesized that Runx2 peaks that are located in rather "open chromatin" regions are more functional than binding in "closed chromatin" regions.

Using microarray expression data for Runx2 overexpressing chicken cell cultures, we found that upregulated genes show indeed a significant overrepresentation of peaks with high ChIP-seq fold change in their promoter sequence. No overrepresentation of high fold change peaks was detected in promoters of downregulated genes, which is in line with the known role of Runx2, that predominantly acts as a transcriptional activator [46]. We conclude from these findings that the excess in ChIP-seq signal above the level that can be predicted from the sequence alone, is a useful measure for the functionality of the peaks and might be used to prioritize functional binding events with high binding probability from non-functional binding events with low binding probability.

Chapter 3

Global comparison of DNA binding profiles

3.1 Introduction

In the previous chapter, we analyzed to what degree transcription factor-DNA binding events depend on the sequence content of the DNA. We showed that ChIP-seq signal is influenced also by chromatin structures as measured in histone modifications or presence of neighboring binding events. This chapter investigates the variation in DNA-binding with respect to changes in the amino acid sequence of the transcription factor. Changes in the protein sequences, especially in the DNA-binding domain of transcription factors may affect the specific recognition motif as well as the nonspecific DNA-binding affinity. Moreover the capability to interact with cofactors can be modified.

We will analyze ChIP-seq data for the homeodomain transcription factor Hoxd13. Hoxd13, the most 5' gene of the Hoxd cluster, is an important regulator of limb patterning and growth [53]. A diverse spectrum of limb morphopathies are associated with Hoxd13 mutations. Three distinct classes of mutations have so far been described in HOXD13: polyalanine tract expansions, truncations, and specific amino-acid substitutions. Each class of mutation has a different molecular pathophysiology with corresponding differences in the phenotypic outcome [54, 55].

We will focus on two mutations R298Q and Q317K, in the homeodomain of Hoxd13 that have been detected in patients with distinct malformations of hands and feet. In order to functionally characterize these mutations, we will compare their genome-wide binding profiles to the profile of Hoxd13 wildtype (wt). The glutamine-lysine substitution at position 317 occurs naturally in a few other homeodomain transcription factors that include the members of the Obox family. However the factor with the greatest protein similarity to Hoxd13 and with the lysine at the equivalent position is Pitx1, a factor that is specifically expressed in the hindlimbs but not in the forelimbs. One important question, that could be answered from the comparison of binding profiles would be, whether the Hoxd13 Q317K mutations changes the recognition motif into one similat to that of Pitx1. We performed a ChIP-seq experiment with Pitx1 in order to see, whether the Pitx1-like mutation in the protein is also associated with a more similar binding behavior.

The regulation by Hox proteins has been shown to be strongly dependent on cofactor interactions. The most prominent Hox cofactors are Pbx1 and Meis1 [56, 57], which are also homeodomain proteins. Hox proteins of paralogous group 13 have also been shown to specifically interact with

Transcription factor	N_{all}	$N_{after\ qualitiy}$	$N_{non\ redundant}$	$N_{uniquely\ aligned}$	N_{peaks}	d (bp)
Hoxd13	24.1	21.5	18.9	14.9	32,902	231
R298Q	46.9	31.5	27.8	23.8	42,835	222
Q317K	60.7	45.2	38.5	33.4	66,020	213
Pitx1	42.3	32.6	28.0	22.9	57,218	226
Smad5 (Hoxd13)	52.2	43.4	37.1	29.4	1615	200
Smad5 (Gdf5)	62.3	43.3	37.1	27.6	262	200
input	42.0	31.6	29.9	25.3	-	-

Table 3.1: ChIP-seq analysis overview. Number of reads $\times 10^6$, called peaks and estimated fragment size d

Smad proteins which are mediators of bone morphogenetic protein (BMP) signaling. In the second part of the chapter, the influence of the amino acid substitutions on possible Hoxd13 cofactor interactions is analyzed and observations based on motif analysis are correlated with cofactor ChIP-seq experiments. Mutations in the Bmp receptor ligand GDF5 show similar phenotype (brachydactyly type A2) as the R298Q substitution in Hoxd13 [58]. In addition Hoxd13 has been shown to regulate the Bmp4 promoter. Thus, Hoxd13 is connected to Bmp signaling on multiple levels and analysis of ChIP-seq data may allow to gain further insights into the functionality of these interactions.

3.2 Methods

3.2.1 Alignment and Peak Calling

For all ChIP-seq and input samples we filtered raw Illlumina GAII reads for mean phred quality score above 30 and removed all but one copy of multiple reads that have the identical sequence in order to avoid read stacking artifacts. We aligned the remaining non-redundant reads with up to three mismatches to the chicken genome (WUGSC 1.1/galGal3) using the Bowtie aligner (version 0.12.5 with -v 3 -m 1 options).

We used the program MACS 1.4.0beta to identify enriched regions relative to the input controls [25]. We ran the program with -mfold=6,30 and –tsize=36, –gsize=1000000000 parameters for all ChIP-seq samples together with the sequences from the input DNA sample as control and extracted all peaks with p-value $< 10^{-5}$. Table 3.1 gives an overview of the alignment and peak calling results.

3.2.2 Identification of differentially bound regions

To define sets of differentially bound regions for Hoxd13 *vs.* R298Q and Hoxd13 *vs.* Q317K, we ran the MACS program by substituting the input control with the Hoxd13 mutation sample. This yielded 11,795 differentially bound peaks for Hoxd13 *vs.* R298Q and 21,830 Hoxd13 *vs.* Q317K peaks. 25,896 regions were defined as negative peaks in the comparison Hoxd13 *vs.* R298Q. The negative peak set denotes the peak set that is called after sample swap. Thus, these 25,896 peaks are more strongly bound by the R298Q mutatnt than by the wildtype Hoxd13. 57,915 negative peaks were called for Q317K *vs.* Hoxd13.

3.2.3 Significance of peak overlaps

We tested for significant overlap of peaks for two ChIP-seq experiments by counting the number of peak summits from the first experiment that are found in $<$ 1kb distance to a peak summit from the second experiment. For each peak in the second experiment, we chose a random genomic coordinate not allowing for overlaps between the randomly selected genomic locations. We determined empirical p-values by counting the number of 1000 random selections, in which the peak summits from the first showed an equal or greater colocalization with a random set.

3.2.4 UniPROBE motif analysis

We downloaded position specific weight matrices (PWM) for Hoxd13 and Pitx1 from the UniPROBE database [59]. These matrices were derived from 8-mer enrichment scores that were obtained by measuring protein binding to immobilized double stranded DNA on a microarray [8]. To score ChIP-seq sequences with respect to these matrices, we used the program MATCH [10] which computes similarity scores for the whole matrix as well as for a core part. The score for a sequence is the sum of all frequencies for all positions multiplied with an information vector, which is used to penalize mismatches at highly conserved positions more strongly than mismatches at less conserved positions [10]. This score is then divided by the best achievable score and cutoffs for the whole matrix and the core part define whether a sequence is scored as a hit or not. We defined the core part as the longest stretch of consecutive positions with frequencies at the most frequent base above 0.9, and set core threshold to 0.9 and whole matrix cutoff to 0.75. We evaluated the ChIP-seq sequences by sampling 10,000 peaks and running MATCH on the peaks that were extended by 1000 bp into both directions. From the result, we computed histograms of the distances of matrix hits to the peak midpoints.

3.2.5 *de novo* motif discovery

For all peak sets, that were defined by comparison to the input control we sampled 3000 ChIP-seq peaks as foreground set and chose 12,000 random genomic locations as background set. We fixed sequence length to the peak midpoint $\pm$ 500bp and applied the Amadeus program [40] to find motifs that are either enriched in the foreground set relative to the background and to detect motifs that show a positional bias relative to the background.

For detection of overrepresented motifs in differentially bound regions, we used a foreground set from the comparative MACS run and used 12,000 peaks from the negative set as background sequences.

3.3 Results

3.3.1 Hoxd13 and Pitx1 motifs are bound *in vivo*

We used PWMs for Hoxd13 and Pitx1 (Figure 3.2 A and B) from the UniPROBE database [59] to check, whether these motifs are enriched near the peak midpoints of the Hoxd13 and Pitx1 ChIP-seq data. We also wanted to see, whether the point mutations impair with the ability to bind the Hoxd13 motif. Figure 3.1 shows histograms for predicted matrix hits for Hoxd13 and Pitx1 matrices. Hoxd13 ChIP-seq data as well as Pitx1 data show an enrichment of hits for the *in vitro* binding matrix models near the peak midpoint. Both mutants show a clustering of Hoxd13

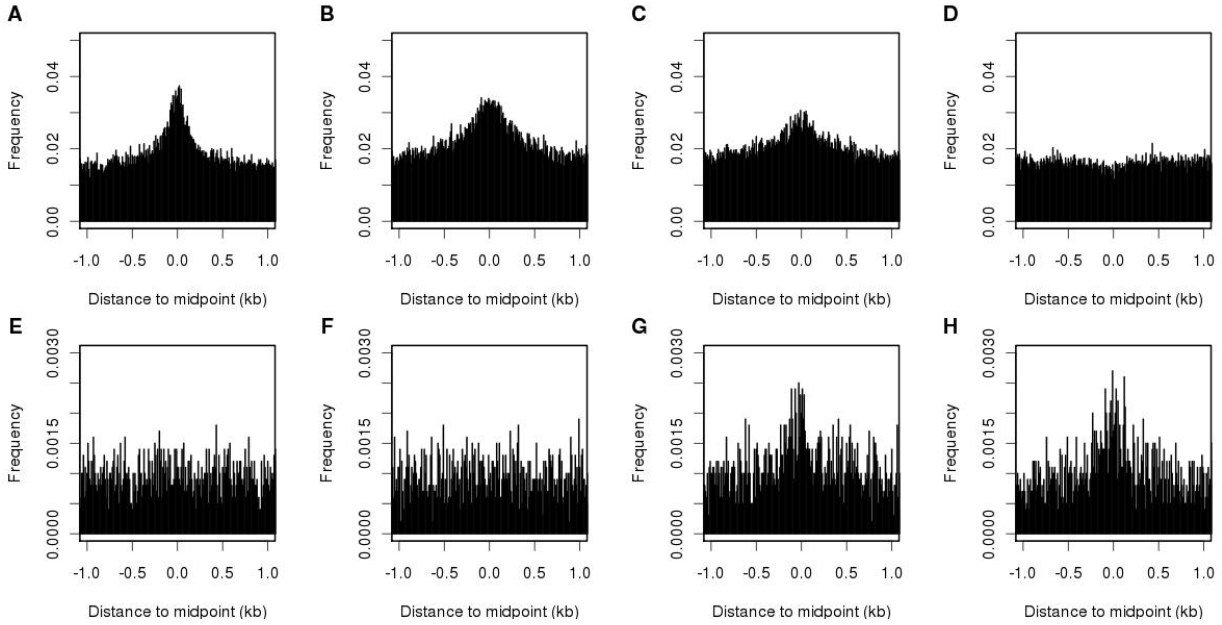


Figure 3.1: Histograms of matrix hits for Hoxd13 and Pitx1 matrices. For all distances in non-overlapping 10bp windows, we counted the number of hits for the Hoxd13 matrix from UniPROBE [59] in peaks from Hoxd13 ChIP-seq **(A)**, R298Q ChIP-seq **(B)**, Q317K ChIP-seq **(C)**, and Pitx1 ChIP-seq **(D)**. Numbers of hits were normalized by the number of peaks. Histrograms for Pitx1 hits are shown for Hoxd13 ChIP-seq **(E)**, R298Q ChIP-seq **(F)**, Q317K ChIP-seq **(G)**, and Pitx1 ChIP-seq **(H)**.

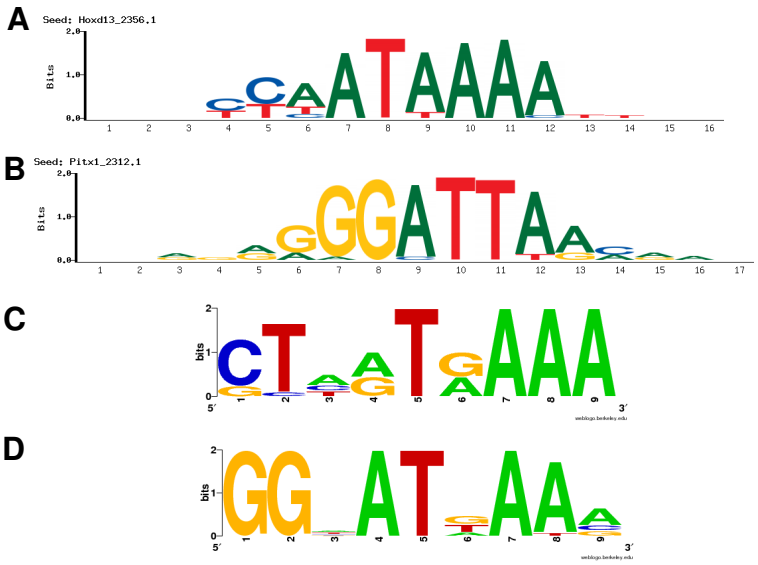


Figure 3.2: **(A)** and **(B)** Hoxd13 and Pitx1 motif from UniPROBE database. **(C)** R298Q ChIP-seq motif, identified by the Amadeus software, **(D)** Q317K ChIP-seq motif. Motif logos for (C) and (D) were generated by WebLogo [60].

hits near the peak center. In addition to predicted Hoxd13 sites, the Q317K mutant shows also an enrichment of the Pitx1 motif near the peak centers. This indicates that the Pitx1-like single amino acid exchange confers the ability to bind a Pitx1 motif.

In order to confirm these results, we performed electrophoretic mobility gel shift assays (EMSA) with purified Hoxd13, Q317K, and R298Q homeodomains, that we expressed in *E. coli*. We used Cy3-labeled and unlabeled oligos carrying either a Hoxd13 or a Pitx1 motif. Hoxd13 and R298Q but not Q317K homeodomains show a band with Hoxd13 oligos, which is weakened if unlabeled competitor DNA is added. For Pitx1 oligos we found that neither Hoxd13 nor R298Q bind these oligos, but Q317K mutant does. Thus the experiments confirm the binding of Hoxd13 and R298Q to the *in vitro* motif (Figure 3.1 A and B) but not the weaker binding of Q317K (Figure 3.1 C). However the expected binding behavior of the Q317K mutant which binds the Pitx1 motif (Figure 3.1 E-G) is nicely reflected by the gel shift experiments.

3.3.2 *de novo* motif discovery identifies affected positions in the Hoxd13 recognition motif

Both Hoxd13 mutants show a weaker enrichment of the canonical Hoxd13 motif at their peak midpoints (Figure 3.1 A-C). To investigate the effect of the mutations on the Hoxd13 recognition site, we ran the *de novo* motif finder Amadeus [40] on the R298Q and Q317K ChIP-seq peaks. The primary motif for the R298Q mutant shows a strong similarity to the Hoxd13 motif. All positions of the Hoxd13 motif match their equivalent positions in the R298Q motif, although at some positions specificity is lost, or a different nucleotide is preferentially recognized (Figure 3.2 C).

The Q317K mutant shows a stronger match to the Pitx1 motif, as the Cytosine before the AT-rich homeodomain core is exchanged by a Guanine like in the Pitx1 recognition sequence. This

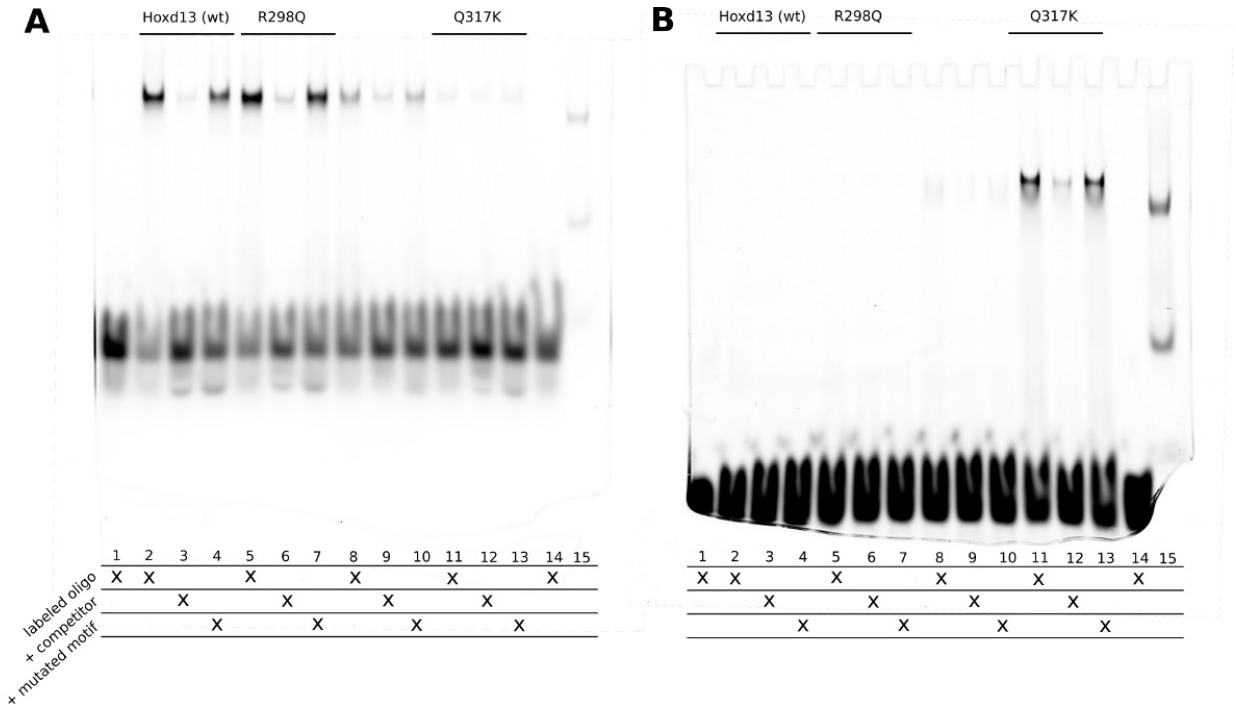


Figure 3.3: **(A)** EMSA experiment with Hoxd13 oligo. For Hoxd13 wildtype, R298Q, and Q317K homeodomains, we tested, whether the oligo containing the Hoxd13 motif is bound. Hoxd13 (wt) and R298Q show a band (lane 2 and 5), which disappears if unlabeled competitor oligos are used (lane 3 and 6). If the Hoxd13 motif in the competitor is mutated, the competitor is not bound anymore and no weakening of the Hoxd13 and R298Q bands is observed (lane 4 and 7). **(B)** EMSA experiment with Pitx1 oligo. Only Q317K mutant shows a band for the Pitx1 oligo and is sensitive to Pitx1 containing competitor. For (A) and (B), lane 1 and 14 have no added protein, lane 8-10 are experiments for a third mutation in the homeodomain, which is not subject of this study. Lane 15 shows a size marker.

stronger degeneration of the Hoxd13 recognition motif for the Q317K mutant is in line with the decreasing trend in Hoxd13 binding affinity from the wildtype ChIP-seq peaks to the Q317K peaks (Figure 3.2 D).

3.3.3 GC-rich cofactor motifs are enriched in differentially bound regions

Mutations in DNA-binding domains do not necessarily only affect the protein-DNA interactions. The homeodomain has been shown to mediate also protein-protein interactions. In the case of the homeodomain protein PBX1, the interaction with other Hox proteins is mediated by contacts of a short Hox hexapeptide motif with the PBX1 homeodomain [61]. We therefore compared differentially bound regions between Hoxd13 wildtype and mutant ChIP-seq profiles in order to identify putative cofactor motifs that may be affected by the single amino acid exchanges. We ran MACS using the Hoxd13 wildtype sample as the foreground set and the mutant samples as controls. The identified peaks represent differentially bound regions that show Hoxd13 wildtype binding but no binding for the mutant. The simultaneously computed negative peak set, which is normally used to define the FDR, represents the regions that are bound by the mutant but not by the wildtype protein. We used the Amadeus software to detect overrepresented motifs in these differentially bound regions. Compared to the R298Q mutant, the Hoxd13 wildtype data shows a broad enrichment of a GC-rich motif (Figure 3.4). For the comparison with the Q317K mutant, we identified three motifs among which one is the Hoxd13 wildtype motif itself. Again we observed a broad enrichment of a GC-rich motif but also a sharp enrichment at the peak center of a relative unspecific CGT motif that resembles UniPROBE motifs for Hoxa13, Hoxb13, and Hoxc13. In addition the sequence GTCGTAAA is part of a motif that has been found as the recognition site for the Hoxa9-Meis1 heterodimer [56]. Also Hoxd13 forms heterodimers with the Meis1. Thus it may be that this motif is a signature of the Hoxd13-Meis1 heterodimer [56].

3.3.4 Smad5 binds DNA indirectly via Hoxd13

It has been shown previously that the Hoxd13 dependent regulation of Bmp4 is mediated by a GC-box motif [4]. The authors speculated that this regulation involves interaction with Sp1. Although they could show a functional cooperation between Sp1 and Hoxd13 in luciferase assays, they could not show a physical interaction between Hoxd13 and Sp1. Thus it remained unclear which protein acts physically as the direct cofactor of Hoxd13.

Hox proteins from paralogs 9, 11, 12, and 13 form heterodimers with the homeodomain protein Meis1 [56], however it has not been reported that Hoxa9-Meis1 complex binds GC-rich motifs like the ones we identified in Figure 3.4. Williams *et al.* showed by yeast two-hybrid experiments that Hox proteins of the paralogous class 13 specifically interact with receptor-regulated Smad (R-Smad) proteins 1,2 and 5 [62]. *de novo* motif detection on ChIP-chip data for Smad1, Smad4, and Smad5 in mouse embryonic stem cells showed that GC-rich motifs are enriched in Smad1 and Smad5 bound sites [27]. Thus it is reasonable to wonder whether the motifs, that are detected in Hoxd13 bound regions represent Smad binding motifs. Typically R-Smads are located at the Bmp receptors, only upon ligand binding, they are phosphorylated, form complexes with Smad4 (coSmad) and translocate to the nucleus. In order to perform a ChIP-seq experiment for activated Smad5, we performed a double infection experiment in chicken micromass cell cultures with RCAS A and RCAS B viruses. One virus carried Hoxd13, which is an activator of Bmp4 and thus induces Bmp signaling and the other virus carried Smad5 with a flag tag. We immunoprecipitated DNA

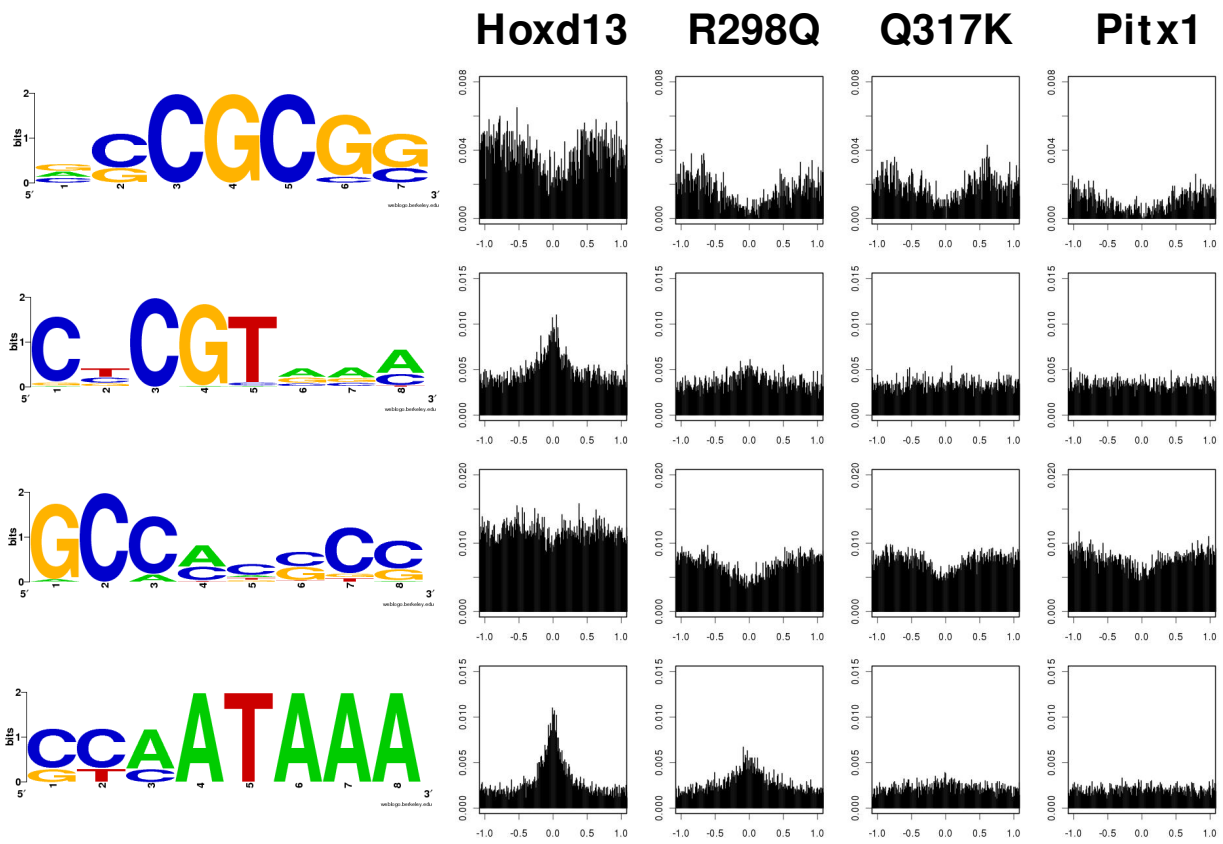


Figure 3.4: Motif overrepresentation analysis of differentially bound regions in ChIP-seq experiments for Hoxd13 and the two mutants identified four motifs. The first motif was identified using the positive and negative peak set from Hoxd13 *vs.* R298Q peak calling run as foreground and background set. The histograms show frequencies of word matches to the corresponding k-mers, that were identified by the Amadeus software [40]. The three other motifs are enriched in regions that are bound by Hoxd13 but not by Q317K. The last motif is almost identical to the Hoxd13 motif from UniPROBE data base, indicating a loss of binding affinity for the Q317K mutant. Motif logos were created by the WebLogo software [60]

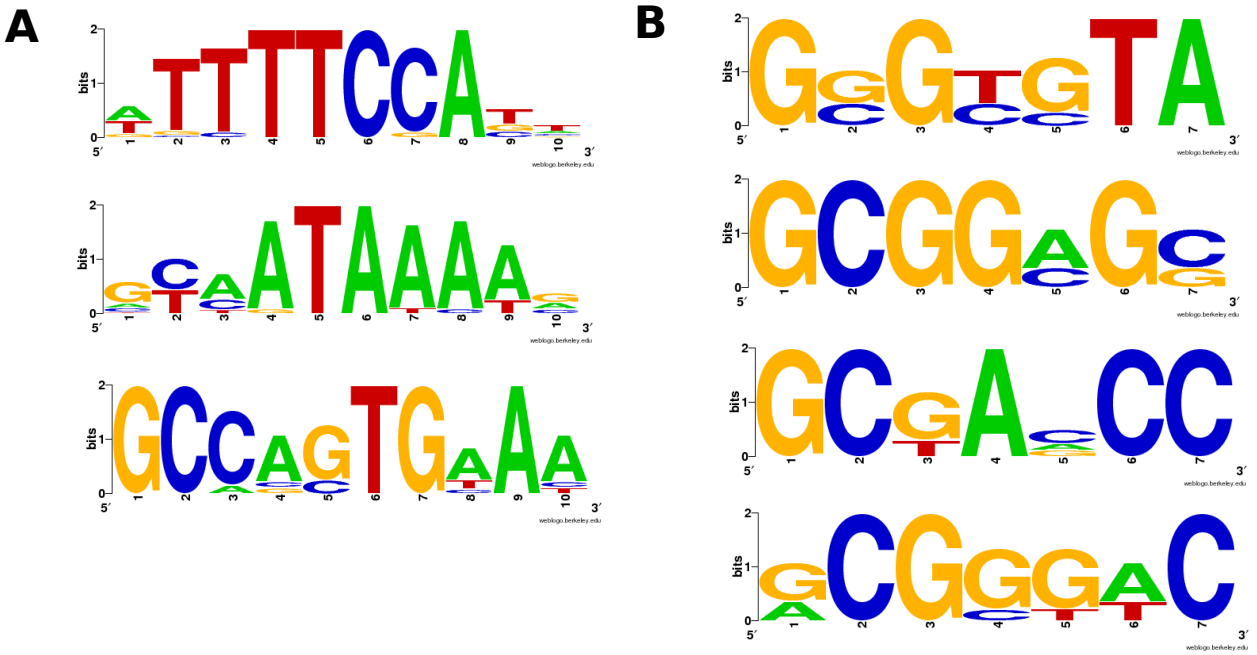


Figure 3.5: **(A)** Overrepresented motifs in Smad5 ChIP-seq peaks from cells that were coinfected with Hoxd13. The second motif strongly resembles the Hoxd13 motif indicating that Smad5 indirectly binds via Hoxd13 **(B)** Overrepresented motifs in Smad5 ChIP-seq peaks from Gdf5 overexpressing cells. Motif logos were generated by the WebLogo software [60]

from these cell cultures on day 9 after infection using a flag-tag antibody, followed by Illumina GAII sequencing.

We identified 1615 peaks that are bound by Smad5 in Hoxd13 overexpressing chicken micromass cell cultures. We carried out a *de novo* motif detection on these sequences, which identified three significantly overrepresented motifs of which the second motif strongly resembles the Hoxd13 motif. This indicates that the immunoprecipitated chromatin contains sequences, that are indirectly bound by Smad5 via Hoxd13 (Figure 3.3.5 A). This confirms the interaction between Smad5 and Hoxd13 that has been observed by Williams *et al.*.

3.3.5 R298Q ChIP-seq peaks are depleted of Smad5 binding

We next investigated the impact of the Hoxd13 mutations on the interaction with Smad5 by assessing the degree of overlap between Hoxd13 and Smad5 ChIP-seq experiments. We performed a ChIP-seq experiment for flag-tagged Smad5 in chicken micromass that were coinfected with Gdf5 to activate Bmp signaling. Using standard peak-calling, we identified 262 Smad5 peaks. We identified four overrepresented motifs (Figure B) that show a high similarity with Smad5 motifs identified from ChIP-chip experiments in mouse embryonic stem cells [27], given that there was a low number of Smad5 bound regions and there were differences in species, tissue, developmental stage, antibody and experimental protocol.

75 Hoxd13 ChIP-seq peaks show an overlap with the Smad5 peaks in Gdf5 expressing cells. This represents a four fold higher overlap than would be expected from a random distribution of peaks ($P < 0.001$). For Q317K we found 107 peaks that overlap Smad5 bound regions, which represents a 2.5 fold enrichment over random expectation ($P < 0.001$). However for R298Q mutant we identified only 15 overlaps, which is a 1.7 fold depletion of Smad5 binding overlap ($P = 0.01$).

The differences in cooperative binding with Smad5 becomes even more evident if we consider only those peaks that are identified as differentially bound between Hoxd13/mutant samples. Regions that are bound by Hoxd13 but not by Q317K and vice versa show a significant four fold, repectively two fold enrichment in Smad5 peak overlap ($P < 0.001$). However, Hoxd13 regions that are not bound by R298Q show a more than ten fold higher overlap with Smad5 regions ($P < 0.001$) and regions that are bound by R298Q but not by Hoxd13 show a 2.5 fold reduction relative to the expected overlap ($P < 0.001$). This suggestes that the interaction with Smad5 and potentially other Smad proteins (Smad1 and Smad2) is strongly affected by the R298Q mutation in the homeodomain.

3.4 Discussion

In this chapter we have compared genome-wide binding profiles for Hoxd13 wildtype and Hoxd13 mutants with single missense mutations. We combined identification of differentially bound regions with motif analysis on subclasses of peaks in order to detect changes in the Hoxd13 recognition motif as well as affected cofactor interactions.

Using motif analysis that involves identification of hits for *in vitro* motifs from the UniPROBE database, it was possible to show that Hoxd13 and R298Q ChIP-seq data show stronger enrichments of the *in vitro* Hoxd13 motifs than the Q317K ChIP-seq data. Similarly the *de novo* motif for R298Q had a much stronger match to the *in vitro* Hoxd13 motif than the *de novo* motif for Q317K. In contrast the Q317K mutant shows a higher similarity to the Pitx1 motif and also exhibits elevated

number of hits for the *in vitro* Pitx1 motif. These findings were confirmed by EMSA experiments.

For the identification of differentially bound regions we applied the MACS peak caller by replacing the control sample with the ChIP-seq data sets for the Hoxd13 mutations. This is analogous to comparing the observed peak counts in genomic windows that were defined by the fragment length estimation integrated in MACS, using a local Poisson distribution [31]. Alternative approaches exist that are based on count statistics of digital gene expression experiments [63], however without an independent validation set such as a number of ChIP-qPCR experiments, a comparison of the methods would not be conclusive.

Identification of overrepresented motifs in regions that are bound by Hoxd13 wildtype but not by the R298Q yielded a GC-rich motif that shows a non-centered clustering around the midpoints of the Hoxd13 wildtype peaks. GC-rich motifs are recognized by a large number of transcription factors involving E2F transcription factors, zincfinger proteins and Smads, thus it is difficult to conclude from the motif alone, what protein is one of the cofactors that bind these sequences. However previous studies have shown the importance of GC-rich motifs in Hoxd13 dependent transcriptional regulation of the Bmp4 gene [4]. Smad proteins that are mediator of Bmp signaling have been shown to interact specifically with Hox proteins of the paralogous Hox13 family [62]. In order to validate the decreased affinity between Smads and the R298Q mutant, a co-immunoprecipitation exeriment could be performed. However co-immunoprecipitation experiments require the use of Smad-specific antibodies and in addition it is difficult to quantify the difference in binding affinitiy if the interaction is not totally disrupted. We thus chose to use an alternative approach that indirectly measures the impact of the single nucleotide substitution on the Hoxd13-Smad5 interaction by performing ChIP-seq experiments in coinfected chicken micromass cells and comparing the genome-wide binding profiles. In cells that were coinfected by Hoxd13 and a flag-tagged Smad5, we could detect the Hoxd13 motif as being bound by Smad5, which is a confirmation of the results obtained by Williams *et al.* based on yeast two-hybrid experiments [62]. In cells that were coinfected with flag-tagged Smad5 and Gdf5 as activator of Bmp signaling we showed that the peaks for Hoxd13 wildtype and Q317K mutant show a significant colocalization with Smad5. However R298Q peaks are significantly depleted in Smad5 binding. This is an example of how the comparison of genome-wide binding profiles and assessment of significant colocalizations can validate the results that were obtained by the *de novo* motif analysis. The only unclarity in the analysis is given for the differences in the GC-rich motifs that were identified by the different experiments. It may well be, that depending on the abundances of transcription factors, varying fractions of Hoxd13-Smad5 pairs can bind DNA cooperatively, thus we do not know whether what we detect are rather Smad5 binding alone or really binding of a Hoxd13-Smad5 complex. This could be one explanation for the dissimilarity between the GC-rich motifs. Another explanation could be the low overall number of peaks in the Smad5 ChIP-seq experiments which leads to a not well defined motif and the fact that Hoxd13 also interact with Smad1 and Smad2 [62] and that these potentially interact with other proteins such as Smad4 and Sp1. Thus more experiments and further analysis would be needed to investigate the combined interaction between Smads and Hoxd13 in greater detail.

Chapter 4

Combining binding patterns for identification of regulatory modules

4.1 Introduction

In the previous chapter I have used motif analyses on differentially bound regions between two ChIP-seq experiments, to compare the global binding pattern of transcription factors with respect to protein-DNA binding but also cofactor association. I also used colocalization of putative cofactor ChIP-seq peaks as an argument to confirm the predicted alterations in cofactor associations. In this chapter I will focus more on the functional aspects of combinatorial gene regulation.

Combinatorial binding of transcription factors has been suggested as a means for modulating regulatory outcomes of transcription factor binding. For example the interactions of Hoxa10 with Pbx1 has been shown to inhibit the expression of target genes by recruitment of histone deacetylases to their promoter regions [50]. For Hoxd13 it has been shown that the interaction with a GC-box binding protein is necessary to drive the expression of Bmp4 and that loss of predicted Hoxd13 binding sites had no effect on this activation [4]. Therefore the identification of transcription factor interactions has drawn much attention and a number of experimental and computational approaches exist to screen for physical as well as functional interactions [64, 65]. In addition to the effect of combinatorial regulation at the proximal promoters, recent studies that applied ChIP-based approaches, identified a large portion of binding events in intergenic regions [66]. Intergenic regions comprise *cis*-regulatory modules such as enhancers and silencers that coordinate tissue-specific expression of their target genes by bringing distant regulatory regions into spatial proximity with the promoters of their target genes [32]. It has been hypothesized that the tissue and developmental stage specific activity of *cis*-regulatory modules is dependent on the *trans*-state, that is the protein levels of the transcription factors in the given cell type. Chen and Xu *et al.* performed ChIP-seq for 13 sequence specific transcription factors and two cofactors in mouse embryonic stem cells and identified two clusters of transcription factor combinations that showed significant overlap of binding events. The first cluster comprised the factors Nanog, Sox2, Oct4, Smad1, and STAT3, the second cluster contained n-Myc, c-Myc, E2f1, and Zfx. Both clusters were enriched in the vicinity of embryonic stem cell specific genes [67].

In this study we investigate the combinatorial binding of six transcription factors in chicken micromass (chMM) cell cultures, that serve as a model for chondrocyte differentiation. We performed ChIP-seq for Hoxd13, Runx2, Msx2, Dlx5, Twist2, and Osr1. Our focus is on the runt-related tran-

scription factor *Runx2*, which is the key regulator of osteoblast differentiation. Runx2 knockout in mice leads to a complete lack of bone due to a defect in osteoblast, osteoclast and chondrocyte differentiation [68, 69]. Although it has been reported that Runx2 together with Ikzf1 directly inhibit the expression of the zinc-finger transcription factor odd-skipped related 1 (Osr1) [13], Runx2 predominantly acts as transcriptional activator [46] and regulates the expression of important osteoblast differentiation markers such as osteocalcin [70] and osteopontin [71]. The homeodomain protein MSX2 is able to bind the RUNX2 protein and to repress its activating effect on the osteocalcin promoter, however this repressive effect can at least partially be alleviated by DLX5 via binding MSX2 [72]. Msx2 and Dlx5 interaction play also a role in the regulation of Runx2 itself since Msx2 suppresses the activating effect of Dlx5 binding [73]. Runx2 and Dlx5 expression patterns show a strong overlap and the expression level of Dlx5 is decreased in the Runx2 knockout mice [2]. Together with the finding that Dlx5 itself is an activator of Runx2 [74], this suggests that Runx2 and Dlx5 form a mutually activating double positive feedback loop [75] and synergistically drive osteoblast differentiation with Msx2 as antogonist acting on the level of Runx2 expression and target gene regulation. A second antagonist of osteoblast differentiation is the basic helix-loop-helix protein Twist2 which can bind the Runt domain of RUNX2 and in the case of Bone-sialoprotein (Bsp) it has been shown that it alleviates the activating effect of Runx2 [76, 77]. Runx2 can be activated by a number of homeodomain proteins [78, 74], recently it was shown that the expression level of Runx2 lies under direct control of another homeodomain factor, Hoxd13 which is expressed in the most distal parts of the developing limbs [79]. Further experimental evidence suggest that Hoxd13 also activates Msx2 (Villavicencio-Lorini, unpublished). Figure 4.1 summarizes regulatory interactions between the six transcription factors. The activating and inhibitory edges indicate known regulatory interactions in the gene regulatory network centered at Runx2. Interactions from literature [76, 77, 74, 78, 79] were extendend by qPCR and microarray expression data from chicken cell cultures that overexpress either Runx2 or Hoxd13 (unpublished data from Villavicencio-Lorini and this study). Already this small subset of key players in osteooblast and chondrocyte differentiation demonstrates the high degree of mutual regulation and protein interaction within the underlying regulatory network, e.g. while Hoxd13 directly activates Runx2 [79], activity of Runx2 in turn results in the downregulation of Hoxd13.

In this study, we generated ChIP-seq data for all of the above mentioned factors and identify recurring patterns for combinatorial regulation of target genes. We developed a method that ranks classes of putative *cis*-regulatory modules with respect to their association with differentially expressed genes. Our method is based on the comparison of enrichment scores of gene set enrichment analyses [80, 81] and takes into account ascertainment biases that are due to assignment of binding events to the nearest genes [82]. Our analyses suggest that *cis*-regulatory modules, defined by colocalized binding events of specific transcription factor combinations show a stronger enrichment in differentially expressed genes, than binding events of single factors. When excluding binding events, that do not show a substantial functional association, our method can be used to dissect the initial low-stringency peak sets of roughly 45-68,000 binding events into smaller subsets of a few hundred target genes that are associated with one particular class of *cis*-regulatory module and that can be characterized in further analyses. Thus our method represents a powerful filter for classification of ChIP-seq data and for investigation of combined regulatory interactions.

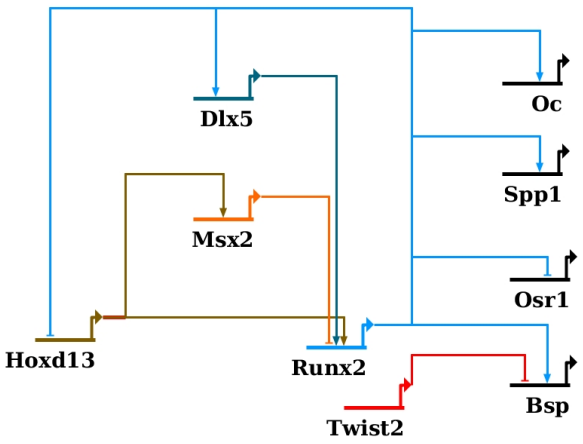


Figure 4.1: A subset of the gene regulatory network involving Runx2. The indicated regulatory interactions denote direct regulation or co-regulation from literature. The upregulation of Msx2 by Hoxd13 and the downregulation of Hoxd13 by Runx2 is based on microarray expression data in chicken micromass cell cultures (this study) and qPCR experiments (Villavicencio-Lorini, unpublished).

Transcription factor	**N $_{\text{all}}$**	**N $_{\text{after qualitiy}}$**	**N $_{\text{non redundant}}$**	**N $_{\text{uniquely aligned}}$**	**N $_{\text{peaks}}$**	**d (bp)**
Hoxd13	61.2	33.8	25.8	20.7	68,134	236
Runx2	28.9	26.0	16.3	13.5	45,983	246
Dlx5	34.4	29.3	24.1	20.7	68,108	238
Msx2	46.3	32.7	27.5	24,2	62,363	251
Twist2	62.8	34.6	26.3	23,9	59,333	243
Osr1	31.9	27.3	17.9	14,7	47,740	246
input	42.0	31.6	29.9	25.3	-	-

Table 4.1: ChIP-seq analysis overview. Number of reads $\times 10^6$, called peaks and estimated fragment size d

4.2 Methods

4.2.1 Alignment and peak calling

For all ChIP-seq and input samples we filtered raw Illlumina GAII reads for mean phred quality score above 30 and removed all but one copy of multiple reads that have the identical sequence in order to avoid read stacking artifacts. We aligned the remaining non-redundant reads with up to three mismatches to the chicken genome (WUGSC 1.1/galGal3) using the Bowtie aligner (version 0.12.5 with -v 3 -m 1 options).

We used the program MACS [25] to identify enriched regions relative to the input controls. We ran the program with default parameters for the Runx2 ChIP-seq sample together with the input control and extracted all peaks with p-value < 0.001. MACS reports for each peak the coordinates od the whole peak region, the coordinates of the summit (the position with highest read coverage), the number of reads, false discovery rate, and the fold change*vs.* control. We used the peak summit coordinates for further analysis. Table 4.1 gives an overview of the alignment and peak calling results. We tested for significant colocalization of peaks for two ChIP-seq experiments by counting the number of summits from the first experiment that are found in $<$ 1kb distance to a summit from the second experiment. For each peak in the second experiment, we chose a random genomic coordinate not allowing overlaps between the randomly selected locations. We determined empirical p-values by counting the number of 100 random selections, in which the summits from the first showed an equal or greater colocalization with a random set.

4.2.2 Analysis of gene expression data

mRNA levels for chMM cell cultures were measured on Affymetrix chicken GeneChip expression arrays. We normalized intensity values using GC-robust multiarray average (GCRMA) normalization and computed gene expression fold changes for every gene as the log_2 fold change relative to the empty RCAS virus transfected cells. If multiple probe sets were assigned to a gene, we used the median value as the representative fold change.

4.2.3 Gene set enrichment analysis (GSEA)

Gene set enrichment analysis (GSEA) [80, 81] is a method to quantify the association of a ranked list of genes to some kind of biological or phenotypic annotation. In our case the ranked list is defined by the ranking of genes, ordered by their expression fold change, and the biological association is the information whether the promoter of a gene is bound or not. Traditionally gene expression profiling assays are used to identify lists of significantly differentially expressed genes. Usually multiple replicate experiments are done to estimate the gene-specific expression variance, which is needed to calculate the significance of a given fold change. In a second step the significantly differentially expressed genes are analyzed for a common biological association, that may be represented by an enrichment in biological functions or pathways. Typically Fisher's exact-test is used to detect significant overrepresentation of a biological function [83, 84]. In addition to the need of replicate experiments for p-value calculation, these methods do not take the ranking of genes in the list of expression fold changes into account. Instead the significance of differential expression defines a cutoff which is used to divide the complete gene list into two halves. GSEA was developed to detect common biological associations for a ranked list of genes without the need for specifying a cutoff.

In GSEA the ranked list is processed from top to bottom and an enrichment score (ES) is calculated by adding positive weights for genes that are in the target set and subtracting a negative weight otherwise. Empirical p-values are calculated by randomly permuting the target gene set. ES for different gene sets are usually not comparable if the numbers of target genes, i.e. the numbers of positive labels vary across the sets. Instead associations are ranked by P-values which have to be computed using a lot of random permuations.

For ChIP-seq data one could apply GSEA to answer the question whether genes with proximal binding events are overrepresented in a list of differentially expressed genes that are ranked by their expression fold change relative to a control. In our case the control sample is defined by the expression values in cells that were transfected with an empty RCAS virus. For each gene, the promoter region of a fixed size has to be defined by a certain criterion and all genes with ChIP-seq peaks in their promoter regions are labeled as being bound. Starting with an initial score $ES(0) = 0$, enrichment scores are then computed for each position i in the list L of up and downregulated genes by summing up the positive and negative weights

$$ES(i) = ES(i-1) + \begin{cases} 1 & \text{if gene } i \text{ is bound} \\ -\frac{N_{bound}}{N_{all} - N_{bound}} & \text{otherwise,} \end{cases} \tag{4.1}$$

whereby N_{bound} represents the number of genes with promoter peaks and N_{all} the number of all genes in list L. This forces the enrichment score to start and end at a value of 0.

4.2.4 ChIP-seq enrichment analysis (CSEA)

For a given number k of ChIP-seq experiments with $n_1, n_2 \dots n_k$ peaks per experiment we can classify the peaks into classes of putative *cis*-regulatory modules based on the combination of transcription factors that exhibit colocalized binding.

Given a peak of a certain transcription factor i, we test whether the transcription factor j shows a colocalized binding event as defined by a distance between peaks of i and j which is smaller than a given threshold $\theta = 1$ kb. This procedure results in 2^k classes of binding events. To compare these classes with respect to their functional relevance as measured in their association with differential expression, we can simply compare the p-values of the gene set enrichment analysis. However, since the p-values are computed empirically by permuting of labels, the number of simulated labelings grows exponentially with the precision of the p-value. In addition, p-values have to be corrected for multiple testing errors and thus even more permutations are needed to detect a significant associations. The comparison of enrichment scores seems to be more intuitive and requires less computation. However since the numbers of instances within each putative *cis*-regulatory module class differs, we need some adjustment to take into account the differing numbers of binding events.

The enrichment analysis of ChIP-seq data is complicated by the fact, that the majority of binding events occur in intergenic regions [66]. Figure 4.3 B shows that for all experiments $< 20\%$ of peaks are located within 5kb upstream distance to a TSS. In order to incorporate also intergenic binding events into the analysis it seems reasonable to assign a binding event to the nearest gene (the gene with the nearest TSS), but doing so would lead to an ascertainment bias due to the non-random distribution of gene deserts across individual loci [82]. Gene Ontology (GO) enrichment of locations that were distributed randomly across the genome and are assigned to their nearest gene would result in a strong overrepresentation of GO terms such as 'multicellular organismal development', because genes that are annotated with this term show a strong tendency to be flanked by large intergenic gene deserts [85, 86].

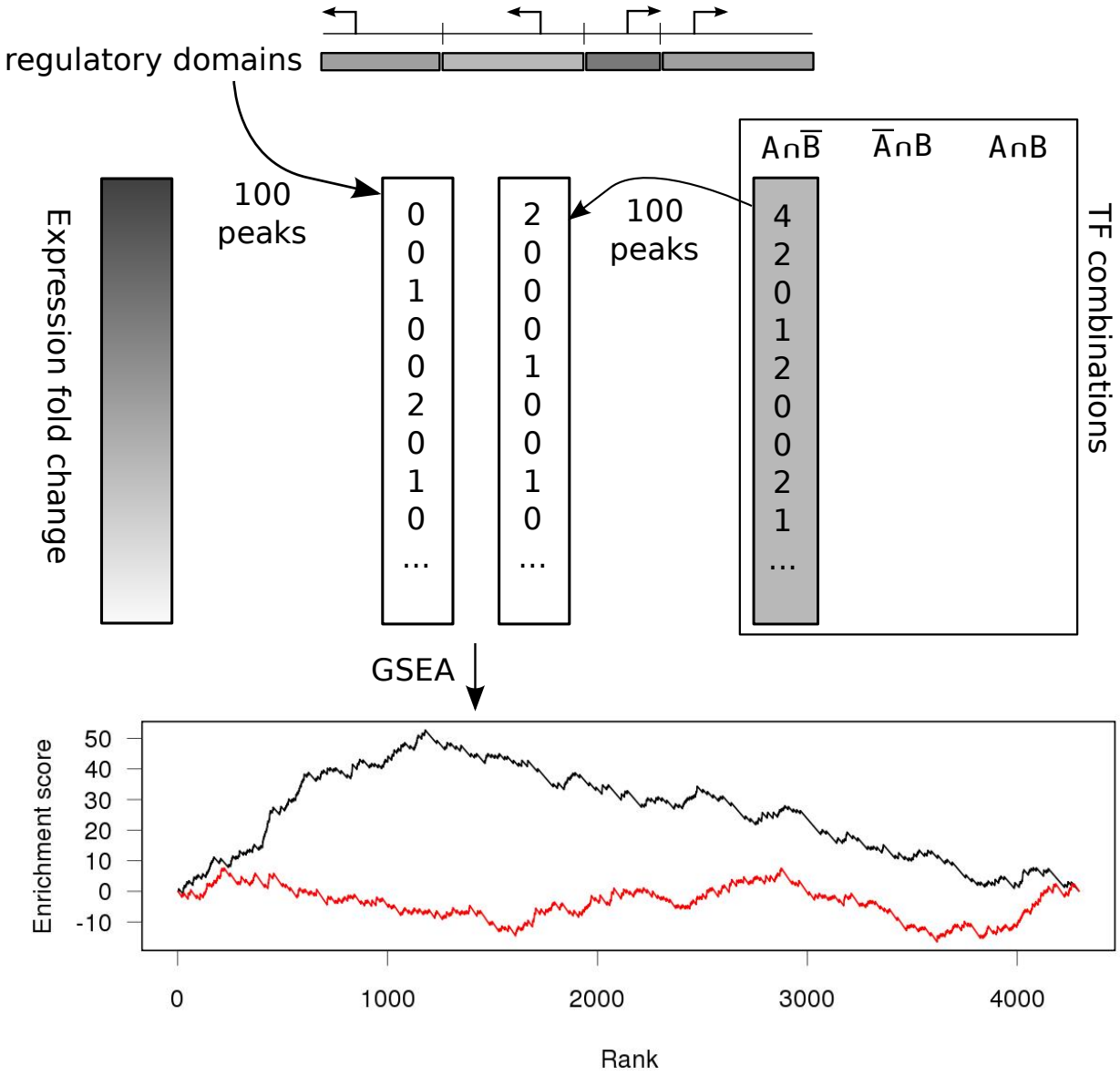


Figure 4.2: ChIP-seq peaks are divided into classes that are defined by the combination of transcription factors that show colocalized binding in multiple experiments. Based on the distribution of TSSs, regulatory domains for each gene are defined and 100 peaks are randomly chosen with probability proportional to the size of the regulatory domains. An equal number of peaks is sampled from the distribution of ChIP-seq peaks for a given TF combination and enrichment scores are computed using GSEA. This is repeated 100 times and p-values are computed using a Wilcoxon ranksum test.

We developed a variant of GSEA that we termed ChIP-seq enrichment analysis (CSEA). CSEA takes the ascertainment bias due to genomic distribution of ChIP-seq peaks into account and it can be used to rank classes of putative *cis*-regulatory modules with respect to their overrepresentation in a ranked list of differentially expressed genes. The general methodology of CSEA is outlined in Fig. 4.2.

For each TSS we define a regulatory domain that is defined as the sum of the genomic distances to the nearest upstream and downstream TSS of which each is divided by two. This is equivalent to assigning a peak that is defined by the summit position, to the nearest TSS. In cases of genes with multiple transcription start sites, the regulatory domain for the gene is simply defined as the sum over all regulatory domains of each TSS. Given the size distribution of gene regulatory domains, we sample 100 times 100 random peaks and use the associated gene sets as a background model. For each *cis*-regulatory module class with more than 100 peaks, we sample up to 100 times 100 peaks with replacement from the observed peak distribution. If 100 times 100 peaks exceeds the number of peaks in this class (n), only $\lfloor n/100 \rfloor$ times 100 peaks are sampled. We then define all genes with at least one peak as being 'bound' and compute enrichment scores for each sampled gene set. Always choosing a fixed number of peaks keeps the number of associated genes on a comparable level so that we can compare enrichment scores across different *cis*-regulatory module classes.

Typically empirical p-values are calculated as the fraction of random samplings that exhibit equal or higher enrichment scores as the observed target gene set [81]. Since we split the ChIP-seq data into subsets of $n = 100$ peaks, we obtain two distributions of enrichment scores. To calculate p-values for the difference between the distribution of enrichment scores for the random sets and the enrichment scores for the ChIP-seq peaks, a Wilcoxon ranksum test is applied. Wilcoxon p-values are corrected for multiple testing by using FDR correction and a significance level of 0.05.

4.3 Results

4.3.1 Overlap of ChIP-seq profiles indicates large-scale colocalization

In order to compare the raw ChIP-seq profiles, we calculated pairwise Pearson correlation coefficents of read numbers in non-overlapping 1kb windows (Fig. 4.3 A). ChIP-seq profiles between Runx2, Dlx5, Msx2, Twist2, and Osr1 show highly significant pairwise correleation coefficients with $r \geq 0.33$ which is higher than any pairwise comparison of any of these factors and Hoxd13. This is in agreement with the known protein-protein interactions between Runx2, Dlx5, Twist2, and Msx2 [72, 76], no interaction between any of these factors and Hoxd13 is known. The strong correlation between Osr1 and these factors might suggest that Osr1 might be located in the same multiprotein complex as at least some of these factors.

After peak-calling with MACS [15], we tested how many peaks colocalize for each pairwise comparison (Fig. 4.3 C). We defined colocalization between two peaks, if the summits are located less than 1kb away from each other.

Again pairwise comparisons involving Hoxd13 showed a lower average peak overlap than for any other comparison. With the exception of Hoxd13-Runx2 colocalization all pairwise comparisons indicated that more peaks tend to colocalize than would have been expected from a random distribution of peaks ($P < 0.01$). Instead Hoxd13 and Runx2 showed significantly less colocalization than expected ($P < 0.01$).

Figure 4.3 B shows the distribution of peaks relative to gene structures. 60-80% fall into

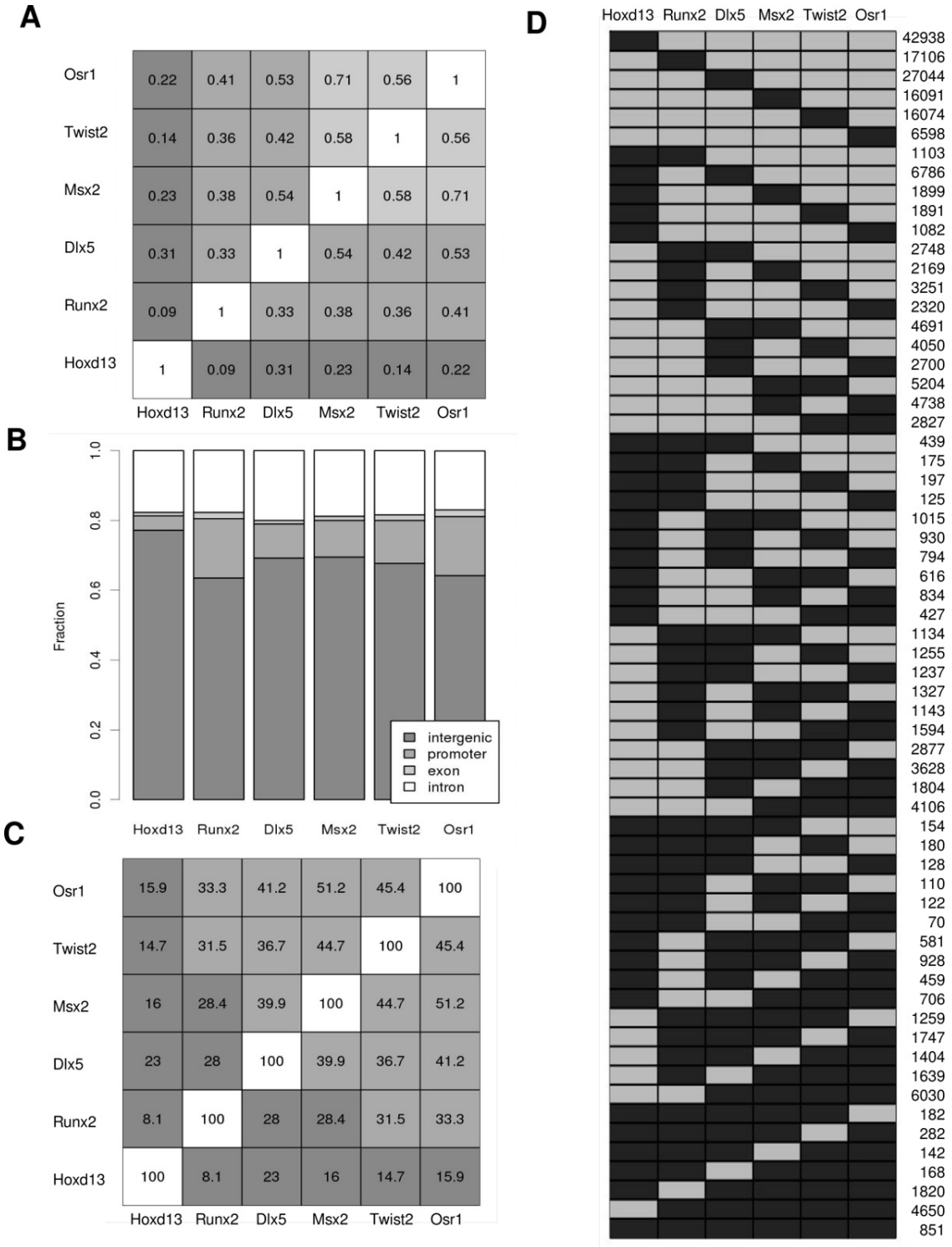


Figure 4.3: **(A)** Pearson correlation between read counts in non-overlapping 1kb windows. **(B)** Distribution of peaks relative to gene structure, promoter sequences were defined as 5kb upstream to the TSS. **(C)** Average pairwise peak overlap between ChIP-seq experiments. **(D)** Colocalization of binding events. For each combination of transcription factors, as indicated by dark squares, the number of colocalized bindinge events is shown, that are defined as summits with $<$ 1kb distance.

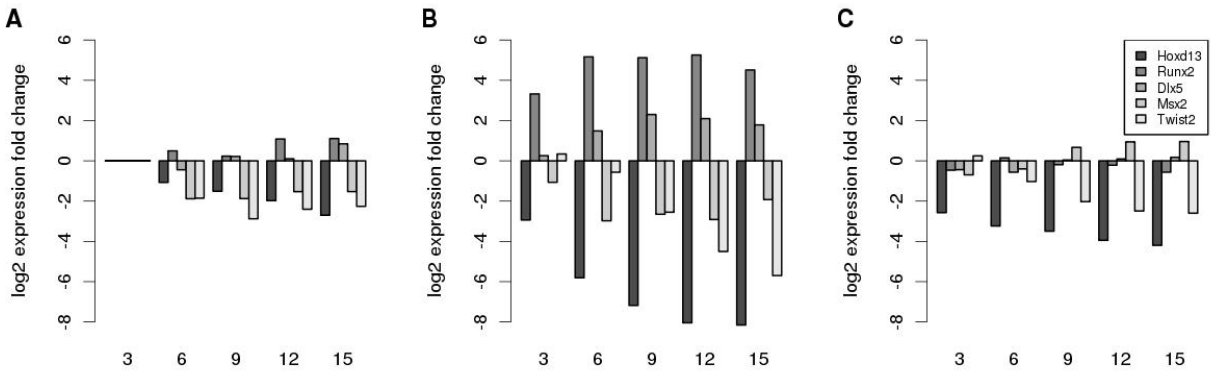


Figure 4.4: Expression fold changes for Hoxd13, Runx2, Dlx5, Msx2, and Twist2 in chMM cell cultures over a timecourse from 3 to 15 days after infection with the RCAS virus. The microarray does not include a probe set for Osr1, therefore no expression data for Osr1 is shown. Fold changes are relative to expression levels in empty RCAS virus infected cells at day 3 after infection. **A)** Expression in empty RCAS infected cells. **B)** Expression in Runx2 RCAS virus infected cells. **C)** Expression in Hoxd13 RCAS virus infected cells. A decrease of Hoxd13 expression is observed due to the fact that the mouse *Hoxd13* was transfected and that the chicken Hoxd13 probe sequence on the array differs from the mouse sequence.

intergenic regions and roughly 10-20% of Runx2, Dlx5, Msx2, Twist2, Osr1 peaks fall into promoter regions as defined by the 5kb upstream region of the TSS.

Figure 4.3 D shows the number of peaks that correspond to one class of *cis*-regulatory module. For a definition of colocalized binding events for more than two experiments, we defined the summit from the first experiment as the reference coordinate and tested all peaks from other experiments with respect to this reference. Previous studies have shown protein-protein interactions between Msx2 and Runx2, Dlx5 and Msx2, and Twist2 and Runx2 [72, 76], indeed there is a substantial peak overlap between these factors. Interestingly Osr1 also shows an extensive overlap with these factors, suggesting a potential interaction with the other factors. In total 4650 peaks are found, that show binding of Runx2, Dlx5, Msx2, Twist2, and Osr1.

4.3.2 Distinct classes of *cis*-regulatory modules show variable effects on expression of target genes

The six ChIP-seq data sets were obtained from different chMM cell cultures each overexpressing the factor of interest. chMM cell cultures transfected with these factors show a diverse spectrum of morphological changes that include different proportion of chondrocytes as indicated by alcian blue staining. Therefore all cell cultures represent distinct mixtures of cells overexpressing a different transcription factor. Thus we cannot conclude from colocalized binding events to combinatorial interaction of transcription factors. However the ChIP-seq data can be interpreted as evidence that the transcription factors may bind the same genomic location in independent assays. Thus we treated the ChIP-seq peaks in the same way as one could do with computational predictions. We classified the peaks into 2^6 classes of putative *cis*-regulatory modules that correspond to combina-

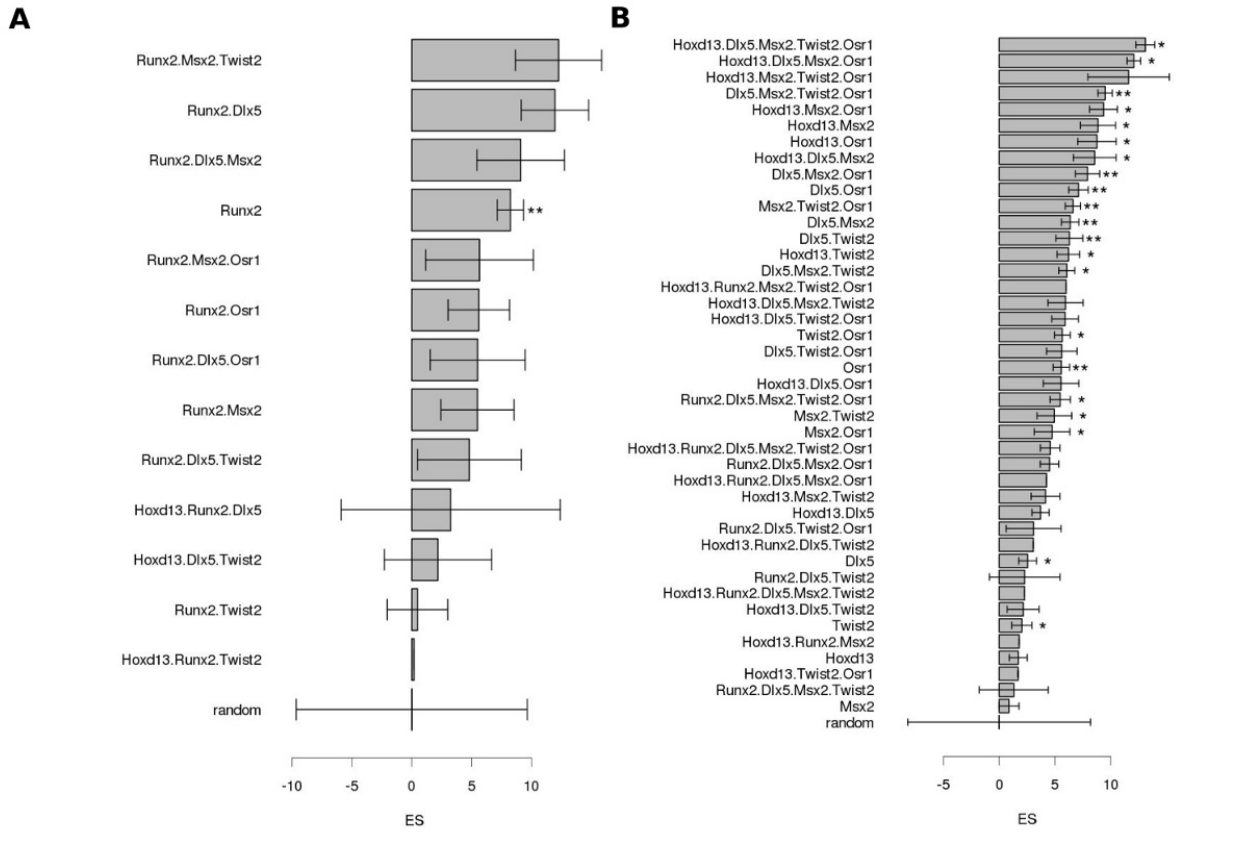


Figure 4.5: Regulatory module enrichment in differentially expressed genes. **(A)** For each *cis*-regulatory module class we calculated enrichment scores with respect to the upregulation of their putative target genes in Runx2 overexpressing chMM. The bars indicate mean values ± standard error of 100 peak sets per class. Enrichment scores were rescaled by subtracting the median enrichment scores for random peaks. '*' and '**' indicate significant enrichments with respect to random assignment of target genes ($P < 0.05$ and $P < 0.001$, Wilcoxon ranksum test with FDR correction). **(B)** Results for downregulated genes in Runx2 overexpressing cells. A number of *cis*-regulatory module classes that do not show binding of Runx2 exhibit the highest enrichment scores indicating that the observed downregulation may be a secondary effect of Runx2 downregulation.

tions of transcription factors that have the capability to bind these modules. We investigated the functional impact of each module class by measuring the degree of overrepresentation of predicted target genes in lists of differentially expressed genes using an approach that is based on gene set enrichment analysis that extends to the application on ChIP-seq data from multiple experiments, termed ChIP-seq enrichment analysis (CSEA). Figure 4.3.3 shows the results for CSEA of the *cis*-regulatory module classes that show the strongest association with upregulated and downregulated genes in Runx2 overexpressing chMM cell cultures. We observed a higher number of putative *cis*-regulatory module classes to be associated with downregulation rather than upregulation. This effect is likely due to differences in sizes of gene regulatory domains for upregulated and downregulated genes. 4293 upregulated genes (expression fold change > 0) have a mean regulatory domain size of 60kb, whereas 4157 downregulated genes (fold change < 0) have a mean domain size of 69kb ($P < 10^{-9}$, Wilcoxon rank sum test). However, the complete set of regulatory domains of upregulated genes spans a region that is 29Mb larger as the regulatory domain size of downregulated genes (286Mb *vs.* 257Mb). This is due to the higher number of upregulated genes (4293 *vs.* 4157) and leads to a higher chance that a randomly picked genomic location will be assigned to an upregulated gene. This leads to a higher background level in the CSEA for upregulated genes.

Relative to empty RCAS virus transfected cells, the gene expression data for Runx2 overexpressing chMM shows a dramatic reduction of Hoxd13 expression to 2% and a 40% reduction of Msx2 level. Dlx5 shows a roughly two-fold increase. Twist2 levels were only increased by 25%, no data is available for Osr1 since no probe set was present on the Affymetrix array. Expression fold changes for all five transcription factors are visualized in Figure 4.4. The enrichment scores that are computed by our method can be used to rank *cis*-regulatory module classes with respect to their functional association but it remains difficult to decide whether the effect on expression is due to a loss of inhibition or due to a gain of activation. However in the case of Runx2, it is reasonable to interpret an overexpression of predicted target genes as effect of direct regulation, since Runx2 is predominantly acting as an activator [46]. This is supported by the finding that *cis*-regulatory modules that exhibit Runx2 binding alone are significantly associated with upregulated genes but not with downregulated genes (Fig. 4.3.3 A and B).

Interestingly the top ten *cis*-regulatory modules for upregulated genes all show binding of Runx2. However with the exception of Runx2 binding alone, all these associations are not significant. This might be due to the low number of peaks for each of these classes or due to the fact, that the expression resulting from Runx2 binding alone overshadows the signals that might be detected by the analysis of the other combinations. However the presence of Runx2 in the *cis*-regulatory module classes with highest enrichment scores strongly suggests a role of combinatorial regulation involving Runx2 in target gene activation. It is interesting, that the ranking of these classes agrees well with the analysis carried out by Shirakabe *et al.* who investigated the role of protein-protein interactions between Runx2, Msx2, and Dlx5 with respect to the Runx2 dependent activation of the osteocalcin promoter [72]. They showed that Msx2 binding to Runx2 impairs with the ability of Runx2 to activate the osteocalcin promoter. Based on the genome-wide data, our analysis shows a simililar trend, that the CSEA enrichment scores for Runx2-Msx2-sites are lower than for Runx2 alone. Shirakabe *et al.* found also that Dlx5 can alleviate the negative effect of Msx2. In our analysis, the Runx2-Dlx5-Msx2-sites show an enrichment score that is even higher than for Runx2 alone. With respect to the inhibitory effect of Twist2 on Runx2, our results show a lower CSEA enrichment scores for sites with colocalized Runx2-Twist2 binding than for Runx2 binding alone [76]. Although these observations may be biased by indirect effects such as the downregulation of

Msx2 in Runx2 overexpressing cells, it seems as if our analysis correlates with the limited knowledge on the Runx2-dependent combinatorial regulation. Interestingly the class with highest enrichment scores is Runx2-Msx2-Twist2, which is counterintuitive since Msx2 and Twist2 are reported as antagonists of Runx2. We speculate that these genes are normally inhibited by Msx2, but the overexpression of Runx2 results in a loss of inhibition due to downregulation of Msx2 itself and simultaneously Runx2 mediated upregulation, but to clarify this point more data would be needed.

We repeated the analysis using expression data for Hoxd13 overexpressing chMM cultures (Fig. 4.3.3). Overexpression of Hoxd13 causes a 25% reduction in Runx2, 6% reduction in Dlx5, a 5.8 fold increase in Msx2, and a 80% increase in Twist2 expression levels with respect to the same timepoint in empty RCAS virus transfected cells. Interestingly the top 19 *cis*-regulatory module classes that are associated with upregulation are all bound by Hoxd13. Similarly, the top fourteen *cis*-regulatory module classes that are associated with downregulated genes are bound by Runx2. The *cis*-regulatory module class that shows the strongest significant association with downregulation is again Runx2-Msx2-Twist2. Since Hoxd13 induces Msx2 expresssion this finding is further evidence that these genes are actively suppressed by Msx2.

The overall enrichment of Runx2 bound modules near downregulated genes is likely an indirect effect, resulting from the downregulation of Runx2 that can be observed in Hoxd13 overexpressing cells (Figure 4.4 C). This is also indirect evidence that combinatorial regulation involving Runx2 is required for activation of these target genes. In summary these two findings suggest a complementary role between Hoxd13 and Runx2.

4.3.3 Roles of Hoxd13 and Runx2 in limb development

To elucidate potential roles of Hoxd13 and Runx2 in limb development, we performed Gene Ontology (GO) term overrepresentation analyses on putative target gene sets for both factors. Intuitively good candidate target genes would be those, that are bound by the factor and are differentially expressed. Thus we defined a so called 'leading edge target gene subsets' of genes that are associated with Hoxd13 and Runx2 binding alone. The leading edge subset was initially defined as the set of all genes with positive labels in the upper part of the ranked list above the position of the maximal GSEA enrichment score [81]. To avoid false positives in the leading edge subset that are due to random fluctuations near the position, where the GSEA enrichment score reaches its maximum, we used a 95% cutoff on the maximum GSEA enrichment score. From the ranked list of upregulated genes in Runx2 and Hoxd13 overexpressing cells we selected all genes in the upper part of the lists, that are associated with binding by either of these factors alone. Table 4.2 shows overrepresented biological processes for 863 potentially Hoxd13 regulated genes. However for the analogously defined set of 708 potential Runx2 targets, no significant enrichment could be detected. This could be due to the overexpression in the chicken micromass cell cultures, which would lead to additional binding events. The resulting regulation may thus represent an experimental byproduct rather than true biology. Under the assumption, that even in overexpressing cells, cofactor levels are influenced to a smaller degree, the limitation of cofactor abundances would suggest, that regions with colocalized peaks should be enriched in true biological binding events. This is consistent with conservation analysis which shows that binding sites that do not show colocalized binding events are less conserved than sites that are bound by multiple factors (Figure 4.7).

We therefore defined new target gene sets based on the indirect evidence of target gene downregulation in Runx2 and Hoxd13 overexpressing cells. For Runx2 we chose the top ten *cis*-regulatory module classes, that are significantly associated with downregulated genes in Hoxd13 overexpress-

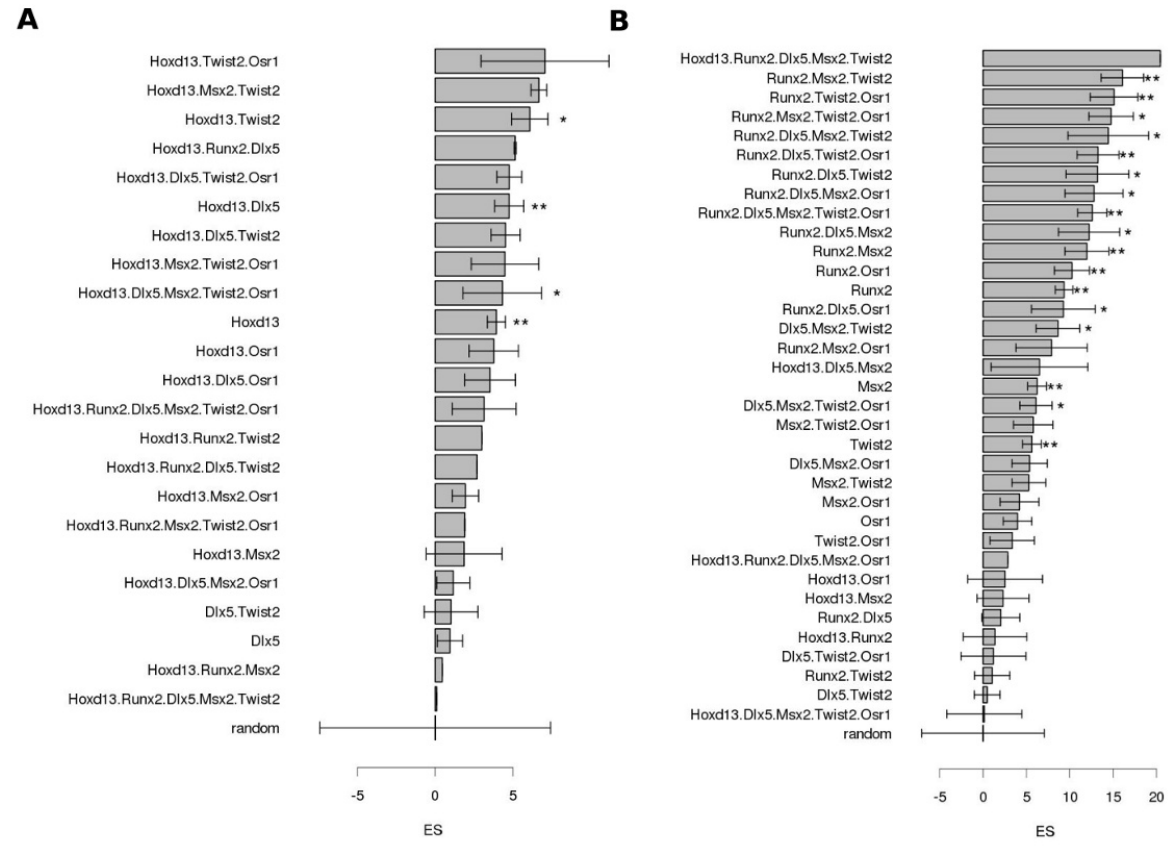


Figure 4.6: ChIP-seq enrichment analysis for Hoxd13 transfected chicken chMM cell cultures. **(A)** Enrichment in upregulated genes from chMM overexpressing Hoxd13. **(B)** Enrichment in downregulated genes from chMM overexpressing Hoxd13.

ing cells (Figure B) excluding Hoxd13 containing classes which might indicate direct inhibition. Similarly we chose among the top 10 classes that are associated with downregulation in Runx2 overexpressing cells (Figure B) all significant Hoxd13 but not Runx2 containing classes. We then defined a 95% GSEA enrichment score gene set using the downregulated genes in Runx2 and Hoxd13 overexpressing cells. This method yielded a set of 368 putative Hoxd13 target genes and 513 putative Runx2 target genes. Gene Ontology analysis identified a number of biological processes that are overrepresented in the two sets (Table 4.3 and 4.4). Hoxd13 shows a strong association with negative regulation of cell proliferation and apoptosis, whereas Runx2 is additionally associated with processes such as cartilage development, and skeletal system development. Interestingly the most significant overrepresentation was detected for sterol metabolic process. Defficiencies in steroid metabolisms are known to affect several processes such as bone mineralization, resorption, and it has been shown that defects in steroid metabolism inhibit growth plate closure. Recently it was shown by Teplyuk *et al.*, that Runx2 regulates a number of genes that are involved in sterol/steroid metabolism [88].

The fact that both factors are associated with negative regulation of transcription might explain for Runx2, that repression of target genes such as Hoxd13 is achieved by activation of repressors

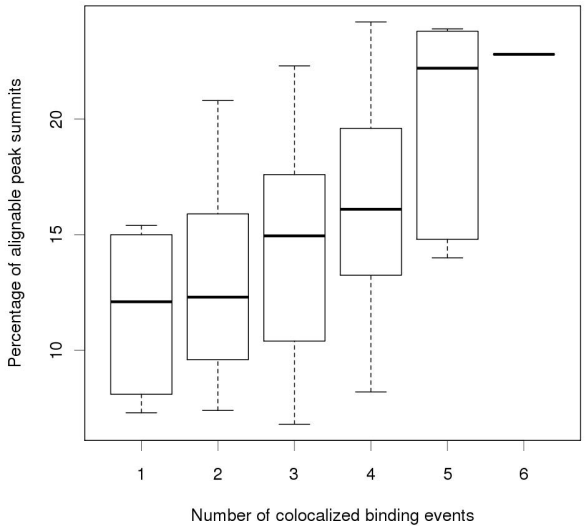


Figure 4.7: Conservation analysis for ChIP-seq summits that are associated with colocalized binding events. Whole genome chicken-human alignments from UCSC encompass 90 Mb [87], which corresponds to $\sim 3\%$ of the human and 9% of the chicken genome. We tested what percentage of summits for each *cis*-regulatory module class is alignable between human and chicken. *cis*-regulatory module classes that are associated with colocalized binding events show a stronger conservation than single factor binding events.

rather than by direct inhibition. The same could be true for Hoxd13 since for both factors we observe a lack of binding events near downregulated genes, which could suggest that both factors predominantly act as transcriptional activators.

GO-term	Description	Enrichment	*P*
GO:0007507	heart development	3.48	0.00519
GO:0051094	positive regulation of developmental process	3.21	0.01640
GO:0040007	growth	3.80	0.01591
GO:0007276	gamete generation	3.74	0.01429
GO:0045597	positive regulation of cell differentiation	3.38	0.01959
GO:0019953	sexual reproduction	3.30	0.02207
GO:0048738	cardiac muscle tissue development	5.78	0.02413
GO:0032504	multicellular organism reproduction	3.05	0.02402
GO:0048609	reproductive process in a multicellular organism	3.05	0.02402
GO:0007167	enzyme linked receptor protein signaling pathway	2.62	0.02404
GO:0014706	striated muscle tissue development	3.86	0.03602
GO:0035051	cardiac cell differentiation	7.31	0.03538
GO:0060537	muscle tissue development	3.71	0.04258
GO:0003013	circulatory system process	4.00	0.04262
GO:0008015	blood circulation	4.00	0.04262
GO:0001568	blood vessel development	2.91	0.04131
GO:0008285	negative regulation of cell proliferation	3.02	0.04243
GO:0010941	regulation of cell death	2.11	0.04318
GO:0007517	muscle organ development	3.34	0.04206
GO:0035295	tube development	2.71	0.04501
GO:0006928	cell motion	2.44	0.04387
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	4.18	0.04237
GO:0001944	vasculature development	2.78	0.04414
GO:0042127	regulation of cell proliferation	2.03	0.04553
GO:0022610	biological adhesion	1.93	0.05
GO:0007155	cell adhesion	1.93	0.05

Table 4.2: Gene Ontology enrichment analysis of 863 genes that are associated with Hoxd13 binding alone. To test for overrepresented terms in GO biological process we used the DAVID enrichment analysis and filtered for FDR corrected p-values < 0.05 [84].

GO-term	Description	Enrichment	*P*
GO:0008285	negative regulation of cell proliferation	6.54	0.00002
GO:0042127	regulation of cell proliferation	3.67	0.00003
GO:0006355	regulation of transcription, DNA-dependent	2.38	0.00002
GO:0051252	regulation of RNA metabolic process	2.34	0.00003
GO:0045449	regulation of transcription	2.07	0.00005
GO:0045893	positive regulation of transcription, DNA-dependent	3.65	0.00291
GO:0051254	positive regulation of RNA metabolic process	3.60	0.00293
GO:0010628	positive regulation of gene expression	3.24	0.00298
GO:0051094	positive regulation of developmental process	4.45	0.00413
GO:0006357	regulation of transcription from RNA polymerase II promoter	3.13	0.00381
GO:0016202	regulation of striated muscle tissue development	10.79	0.00360
GO:0048634	regulation of muscle development	10.79	0.00360
GO:0045941	positive regulation of transcription	3.16	0.00484
GO:0010604	positive regulation of macromolecule metabolic process	2.72	0.00448
GO:0043067	regulation of programmed cell death	3.02	0.00427
GO:0010941	regulation of cell death	3.01	0.00424
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	3.61	0.00484
GO:0045597	positive regulation of cell differentiation	4.72	0.00716
GO:0010557	positive regulation of macromolecule biosynthetic process	2.86	0.00681
GO:0040008	regulation of growth	4.28	0.00652
GO:0042981	regulation of apoptosis	2.92	0.00790
GO:0031328	positive regulation of cellular biosynthetic process	2.78	0.00864
GO:0009891	positive regulation of biosynthetic process	2.76	0.00872
GO:0045935	positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.83	0.01027
GO:0035295	tube development	3.88	0.01273
GO:0051173	positive regulation of nitrogen compound metabolic process	2.75	0.01313
GO:0001558	regulation of cell growth	6.16	0.01304
GO:0051270	regulation of cell motion	4.84	0.02245
GO:0022610	biological adhesion	2.48	0.02629
GO:0007155	cell adhesion	2.48	0.02629
GO:0030334	regulation of cell migration	5.12	0.03723

Table 4.3: Gene Ontology enrichment analysis of 368 genes that are associated with colocalized binding events involving Hoxd13, Dlx5, Msx2, Twist2, and Osr1. To test for overrepresented terms in GO biological process we used the DAVID enrichment analysis and filtered for FDR corrected p-values < 0.05 [84].

GO-term	Description	Enrichment	*P*
GO:0016125	sterol metabolic process	9.02	0.01944
GO:0051216	cartilage development	6.23	0.01294
GO:0010629	negative regulation of gene expression	3.25	0.01278
GO:0008610	lipid biosynthetic process	3.85	0.01709
GO:0008203	cholesterol metabolic process	8.72	0.02305
GO:0016481	negative regulation of transcription	3.24	0.02172
GO:0001501	skeletal system development	3.13	0.02730
GO:0031327	negative regulation of cellular biosynthetic process	2.86	0.02466
GO:0009890	negative regulation of biosynthetic process	2.80	0.02796
GO:0045892	negative regulation of transcription, DNA-dependent	3.54	0.03228
GO:0010558	negative regulation of macromolecule biosynthetic process	2.79	0.03785
GO:0051253	negative regulation of RNA metabolic process	3.42	0.03705
GO:0045934	negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.80	0.04888

Table 4.4: Gene Ontology enrichment analysis of 513 genes that are associated with colocalized binding events involving Runx2, Dlx5, Msx2, Twist2, and Osr1. To test for overrepresented terms in GO biological process we used the DAVID enrichment analysis and filtered for FDR corrected p-values < 0.05 [84].

4.4 Discussion

In this chapter we have analyzed ChIP-seq data for six transcription factors with respect to their potentially combinatorial binding and regulation. We defined a number of putative *cis*-regulatory module classes by a specific colocalization pattern of these six factors. Although our data does not allow to claim, that colocalized binding in micromass cell cultures each of which overexpresses a different transcription factor, is an indicator of combinatorial or cofactor binding, the ChIP-seq data can be used as a reasonable predictor of genomic regions that play a role in this kind of synergistic gene regulation.

Traditional gene set enrichment analysis (GSEA) can be used to detect enrichments of bound genes among lists of differentially expressed genes. However, the null model assumes equal probabilities for genes to be bound. This is legitimate if only a region of a fixed size around the TSS is used to define whether a gene is bound or not. Figure 4.2 B shows that less than 20% of binding events fall into the 5kb upstream region of the TSS. To overcome the limitation of the small percentage of peaks that contributes to the analysis, we have developed a method that allows to integrate the complete set of binding events. The method, called ChIP-seq enrichment analysis (CSEA), measures the association of a class of putative *cis*-regulatory modules with differential expression of predicted target genes, that are identified based on genomic distance to the modules. CSEA takes into account the ascertainment bias, that arises from the non-random distribution of gene deserts, which leads to a greater chance for some genes, that randomly picked genomic locations are assigned to these genes [82]. In contrast to GSEA, our method evaluates subsets of peaks of a fixed number ($n = 100$) for each *cis*-regulatory module class and the null model. This results in an enrichment score distribution for each module class and the null model. Thus, p-values are not calculated empirically as the ratio of random sets that exhibits equal or higher enrichment score than an observed value, but are rather calculated by comparison of the enrichment score distributions of the null model to the distribution, obtained by sampling an equivalent number of *cis*-regulatory modules from the experimental ChIP-seq data. Due to the fixed number of sampled peaks, the enrichment scores are comparable across different *cis*-regulatory module classes. This allows to rank the individual classes with respect to their association with differential expression.

Application of this method on expression data for Runx2 and Hoxd13 overexpressing cells, identified *cis*-regulatory module classes that are directly associated with upregulation of putative target genes, but also revealed candidate *cis*-regulatory module classes that are linked to downregulation of target genes in cells that show decreased expression of the respective transcription factor. Gene Ontology analysis on sets of predicted target genes showed significant overrepresentaion of Hoxd13 target genes in antiproliferative and proapoptotic processes. Simililarly for Runx2 the target gene sets showed increased levels in processes such as sterol metabolism, cartilage and skeletal development. One problem in the analysis could be that experimental byproducts, such as differentially regulated genes due to the overexpression of transcription factors might bias the lists of differentially expressed genes in a way that *cis*-regulatory modules are not recognized as significant since the expression changes are rather subtle in comparison to the results of virus induced overexpression. For the case of Runx2, the overexpression is even increased due to the fact that Runx2 positively autoregulates its own expression. As a consequence, results that show significant regulatory impact but a lack of coherent function in the target gene set, like in the case of Runx2 binding alone should be regarded with care.

In summary CSEA can be used to integrate binding data from multiple ChIP-based experiments in order to prioritize specific *cis*-regulatory module classes with respect to their functional associa-

tion. Thus it facilitates the investigation of combinatorial gene regulation and simultaneously filters out binding events with low functional relevance. As it has been discussed in the first chapter, this distinction between binding events constitutes a great challenge in the current analysis of ChIP-seq data and thus methods such as CSEA might proove helpful for integrating genome-wide binding and expression data into models, that allow to gain insights into the gene-regulatory networks that control vertebrate development.

Chapter 5

Assigning genome-wide binding events to target genes

5.1 Introduction

In the previous chapter, we have used colocalization of transcription factor binding in order to prioritize bound sequences according to their functional association. The majority of bound sequences are located in intergenic and intronic regions. This is similar to the observations that also highly conserved non-coding elements (HCNEs), that have been shown in many cases to encode tissue-specific enhancers [32, 89], are enriched in gene deserts and are located up to several hundred kilobases away from the nearest gene [86, 32]. Many of these intergenic and intronic regions represent evolutionarily conserved enhancers and silencers, which we will refer to as 'enhancers' in the following. Enhancers coordinate tissue and developmental stage-specific expression of their target genes by inducing changes in chromatin conformation in order to bring distant regulatory elements into spatial proximity of the transcription start sites (TSS) of their target genes. Extensive experimental and computational work has been carried out on the detection of enhancer regions [90, 91].

The advent of high throughput chromatin immunoprecipitation assays (ChIP-chip and ChIP-seq) has made genome-wide *in vivo* mapping of protein-DNA interactions possible. In agreement with the observation that evolutionarily conserved regulatory elements are located primarily in intergenic regions, less than 10% of transcription factors have greater than 50% of their binding sites within 2.5kb of a transcription start site [92]. Recently, Visel *et al.* [20] employed ChIP-seq to identify several thousand genomic loci in mouse embryonic tissues which were bound by the enhancer-associated p300 protein. p300 is a transcriptional coactivator [93] that is recruited by other DNA binding proteins in a tissue and cell-type specific manner to form an enhanceosome complex with regulatory activity [67]. About 87% of the p300 bound loci regions showed tissue-specific enhancer activity [20].

Global correlations with expression data [20] and strong biases for HCNEs to occur in the vicinity of transcription factors and developmental genes [32] support the assumption that enhancers regulate nearby genes. To date, computational and experimental approaches for enhancer detection have employed proximity-based cutoffs on genomic distance or nearest gene assignments to associate putative enhancer regions to their target genes [91, 94, 95, 96]. Although the genes located closest to the enhancers are reasonable candidates for the target genes, this is not a general rule. For

instance, a *Pax6* enhancer is located in an intron of a neighboring gene [97, 98]. Interactions between enhancers and their target genes can span large genomic distances. For instance, an enhancer of the sonic hedgehog (Shh) gene is located one megabase upstream of the Shh gene [99]. For these reasons, enhancer targets cannot be reliably predicted by simple computational rules based on genomic proximity.

Besides genomic distance, conserved synteny is the only feature that has been considered to possess predictive power for enhancer-target gene interactions [100, 101]. Conserved synteny generally describes a relative order of two or more genomic loci that is conserved in more than one species (Figure 5.1). This might reflect a certain pattern of co-evolution between regulatory region and target gene. The Vista enhancer browser [89] allows manual investigation of flanking genes, and some genome browsers like SynBrowse include information about conserved synteny [102], but no automated approaches exist that specifically predict the target genes of a number of predicted or known enhancers [66]. Consequently, existing approaches for enhancer detection [103, 104, 91] remain incomplete and fail to integrate important developmental target genes into larger regulatory modules and networks that control multicellular organismal development.

One impediment to progress in this area is the paucity of experimental enhancer-target gene interaction data. Commonly used *in vivo* assays for enhancer activity that use co-injection of enhancer and minimal promoter reporter genes [105, 32, 20] provide evidence about the tissue-specificity of the enhancer but do not indicate which genes are targets of the enhancer. On the other hand, chromosome conformation capture (3C) assays [106, 47] test for physical interactions between enhancer and promoter regions, and thus can be used to identify enhancer target genes. However no large-scale data set of enhancer-specific chromatin interactions is available with which to assess the quality of prediction methods.

To our knowledge, there has been no previous large-scale computational analysis of the prediction of enhancer targets. Ahituv et al. [100] mapped conserved blocks of synteny (CBSs) that were homologous among human/mouse/chicken or human/mouse/frog genomes and identified approximately 2000 CBSs $>$ 200 kb for each comparison. They postulated that such CBSs were enriched for long range regulatory interactions between enhancers and target genes because the prevalence and distribution of chromosomal aberrations leading to position effects showed a clear bias not only for mapping onto CBS but also for longer CBS size. Using a similar definition based on alignments between human and zebrafish genomes, Akalin *et al.* identified a set of genomic regulatory blocks (GRBs) located within conserved human/zebrafish-syntenic regions and predicted a set of 269 target genes of within the GRBs [107]. The authors postulated that HCNEs within the GRBs are enhancers and that transcription factor genes within the GRBs are their targets, but did not develop a method for predicting target genes on a genome-wide basis or of predicting target genes of a specific enhancer protein. In this work, we present a method to predict the target genes of potential enhancers identified as bound DNA sequences in ChIP-seq and ChIP-chip experiments. We evaluated our method using published data for p300 and Gli3 in embryonic mouse tissues [108, 20]. Our method uses an integrative approach based on random forest analysis of a combination of genomic proximity, conserved synteny as well as distance in protein-protein interaction (PPI) networks, and Gene Ontology (GO) similarities between regulator and putative target gene. Our algorithm showed a substantially better accuracy than predictions based on any single feature in isolation.

5.2 Methods

5.2.1 Genome data and alignments

We downloaded pairwise net alignments generated by blastz [109] for mouse (*Mus musculus*, `mm9`) against opossum (*Monodelphis domestica*, `monDom4`), chicken (*Gallus gallus*, `galGal3`), frog (*Xenopus tropicalis*, `xenTro2`), zebrafish (*Danio rerio*, `danRer5`), and fugu (*Takifugu rubripes*, `fr2`). We initially used data from human and dog in our analysis, however including these data sets did not improve the results [1], and therefore these two genomes were not used for further analysis. In addition mouse RefSeq annotations for 22,468 genes were downloaded from the UCSC Genome Browser [87]. The phylogenetic distances between these species are shown in Figure 5.2.9, whereby the branch lengths reflect the average number of subsitutions per site as calculated by genome-wide blastz alignments [110].

5.2.2 p300 ChIP-seq data

Visel *et al.* [20] used chromatin immunoprecipitation with the enhancer-associated protein p300 followed by massively parallel sequencing to map the *in vivo* binding sites of p300 in mouse embryonic forebrain, midbrain, and limb tissue. We downloaded p300 ChIP-seq peaks and lists of upregulated genes that were identified by comparing forebrain and limb expression with E11.5 whole embryo gene expression as measured on Affymetrix GeneChip MouseGenome 430 2.0 arrays. The limb data [111] is based on E11.5 proximal hindlimb expression (GEO series GSE10516, samples GSM264689, GSM264690, and GSM264691) . The ChIP-seq also includes P300 bound sites from midbrain, but we did not use this data because no set of midbrain upregulated genes were defined by Visel *et al.* We focused on upregulated genes since Visel *et al.* [20] only observed ChIP-seq peak enrichments in the vicinity of genes that are significantly upregulated in the corresponding tissue, indicating that p300 acts as a coactivator rather than as a repressor. 2,453 ChIP-seq peaks were obtained for embryonic mouse forebrain tissue and 2,105 for limb. Additionally, 1,062 and 748 significantly upregulated probe sets were obtained that correspond to 555 upregulated genes with RefSeq IDs for forebrain and 347 for limb [20]. Affymetrix gene expression microarrays are not able to reliably distinguish between different transcripts of genes. Therefore, one representative transcript was chosen for each gene according to whether a transcript was in the set of differentially expressed probesets, or failing that, arbitrarily as the leftmost transcript on the Watson strand of the chromosome. This reduced the number of RefSeq IDs to 19,569.

5.2.3 Gli3 ChIP-chip data

We downloaded Supplementary data sets 1 and 2 from Vokes *et al.* [108]. These datasets contain 5,274 Gli binding regions and 753 responsive genes that were identified using pairwise and multiple sample comparison of expression levels (Affymetrix Mouse Exon 1.0 ST arrays) for overexpressed and mutated Gli3 vs. wildtype and anterior vs. posterior forelimbs [108].

5.2.4 Genomic distances between enhancer and target gene

For each gene in a genomic window centered at the enhancer, we calculated the genomic distance between enhancer and target gene as the minimal distance between the endpoints of the enhancer

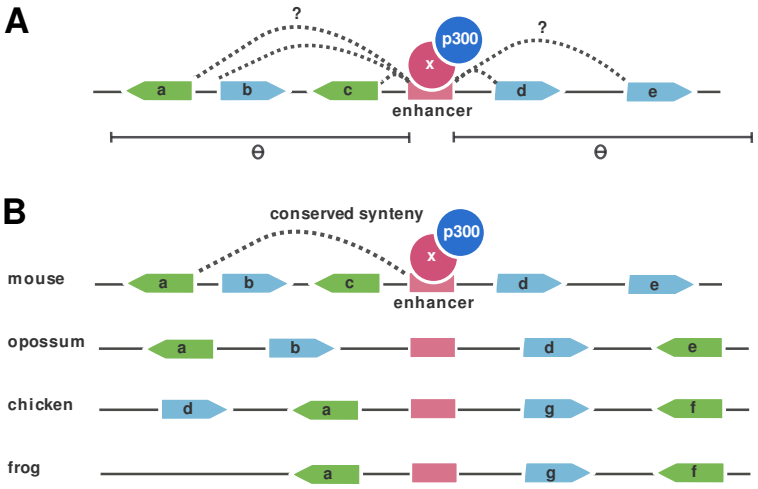


Figure 5.1: **Definition of conserved synteny. A)** p300 binds as a complex with potentially multiple other proteins (denoted by 'x') at enhancers, some of which can be shown to act in a tissue-specific fashion during mouse embryogenesis [20]. However, it is not known which genes these enhancers regulate. Within a genomic window defined by a maximal distance threshold Θ around an enhancer we consider all genes a, b, c, d, e as candidate target genes. **B)** The regulatory interaction between enhancer and target genes may be subject to evolutionary constraint and restricts the enhancer and target gene to be located in a certain genomic proximity. Such a constraint on colocalization is referred to as 'conserved synteny' and it was shown that putative enhancer regions are strongly enriched in highly conserved syntenic regions [100, 101, 1]. In this example, comparative genome analysis of the genomes of opossum, chicken, and frog reveals that although gene a was subject to genomic rearrangements since the split between mouse and chicken common ancestor reducing the genomic distance, gene a exhibits the highest degree of conserved synteny indicating that it is most likely the target gene.

region and the TSS of the candidate target genes. For the genomic-distance based predictions, the gene with the minimal distance was predicted to be the target gene.

5.2.5 Calculation of conserved synteny score (CSS)

We defined for each enhancer e a genomic interval in the reference species r by selecting all genes for which the genomic distance between enhancer and TSS of the gene g is less than a maximal distance threshold Θ (Figure 5.1). For each gene g in this region, we define a conserved synteny score (CSS) by testing in other species $s = 1, \ldots, k$ whether the distance $d_s(e, g)$ between aligned regions of enhancer and TSS is smaller than the maximal distance threshold Θ. The CSS is then calculated as the sum of phylogenetic distances $\phi(r, s)$ (Figure 5.2) between the reference r and species s. The phylogenetic distance function $\phi(r, s)$ is defined as the branch length between species r and s as measured in substitutions per site according to the species tree from Miller *et al.* [110].

$$\mathrm{CSS}(e, g) = \sum_{s=1\ldots k} \delta_s(e, g) \times \phi(r, s) \tag{5.1}$$

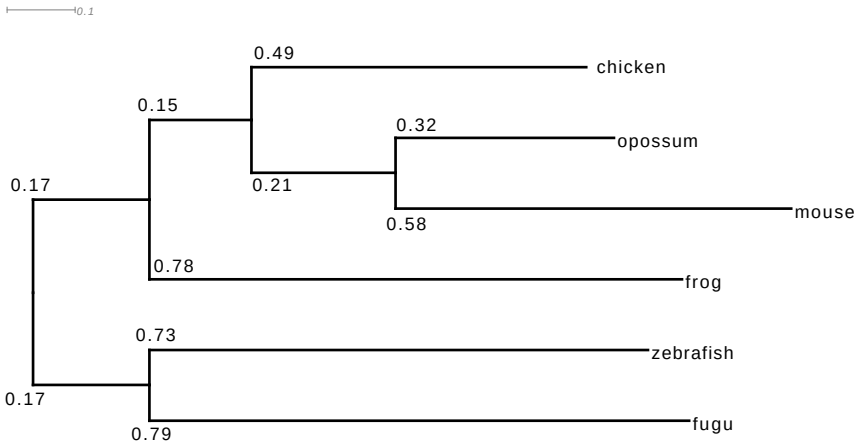


Figure 5.2: **Evolutionary Distances.** We used the phylogenetic tree from the 28-way vertebrate alignments [110] and extracted a subtree with the Dendroscope viewer [112]. The distances between these species represent the average number of substitutions per site as calculated by genome-wide blastz alignments [110]. We defined the phylogenetic distance $\phi(r, s)$ between different species as used in equation 1 as the branch distances measured in substitutions per site according to this tree. For instance $\phi(\text{mouse}, \text{oposssum}) = 0.9$.

The factor $\delta_s(e, g)$ ensures that the phylogenetic distance of a species pair is added only if the genomic distance between the orthologous gene and enhancer region is smaller than the maximal distance threshold Θ.

$$\delta_s(e, g) = \begin{cases} 1 & \text{if } d_s(e, g) < \Theta \text{ in species } s \\ 0 & \text{otherwise} \end{cases}$$

$\delta_s(e, g)$ was also taken to be zero if an orthologous gene and enhancer could not be identified in the other species. Since some genomes in our analysis are not finished and annotations are incomplete, we identified orthologous genes on the basis of the presence of aligned sequences around the promoter region as defined by the TSS $\pm$1kb. This assumes that the enhancer specifically interacts with the promoters of their target genes. This is supported by our recent finding that the occurrence of intergenic HCNEs correlates with conservation in promoter regions of nearby genes [113], which we interpret as evidence for similar evolutionary constraints acting on the enhancers as well as the promoter regions.

For the enhancer sequence, orthologous sequences were identified on the basis of aligned sequence in the other species. We note that rearrangements that disrupt collinearity are not penalized by our scoring scheme because according to our assumption, enhancers can retain their function even after chromosomal rearrangements that invert genes or change their order.

5.2.6 Gene Ontology similarity definition

We calculated for each GO term (t) in the ontology an information content value (IC) defined as $IC(t) = -\log\ p_t$, where p_t is the number of genes annotated by GO term t divided by the total number of annotated genes. The similarity between two terms can be calculated as the IC of their

most informative common ancestor ($MICA$) [114]. This can be used to calculate the similarity (sim) between one set of terms, to another set of terms, each of which belongs to a particular gene (g_i, g_j):

$$sim(g_i \rightarrow g_j) = \frac{\left[\sum_{t_1 \in g_i} \max_{t_2 \in g_j} \; IC(MICA(t_1, t_2))\right]}{|g_i|}. \tag{5.2}$$

Note, that $sim(g_i \rightarrow g_j)$ is not necessarily equal to $sim(g_j \rightarrow g_i)$. As previously described [115] we defined the similarity between two genes as the symmetric version of the formula above by calculating:

$$sim(g_i, g_j) = \frac{sim(g_i \rightarrow g_j) + sim(g_j \rightarrow g_i)}{2}. \tag{5.3}$$

5.2.7 Distance computation for protein-protein interaction networks

In order to define the similarity between two genes we created a network based on the data of the STRING 8.2 database [116], physical and functional interactions. The network consists of 138,156 interactions including 194 direct interactions between p300 and other proteins. In a previous study we have shown that global network similarity measures are better suited for defining functionally associated groups of genes [117]. We constructed a mouse-specific adjacency matrix, which was transformed into a column-normalized adjacency matrix (A). The random walk starts at a certain node corresponding to a gene g_i at timepoint t and randomly visits adjacent nodes. The random walk distance $p_{t+1}(g_i, g_j)$ is defined as the probability of the random walker being at node g_j at timepoint $t + 1$ given that the walker started at g_i. For a vector of starting probabilities p_0, the state probabilities p_{t+1} can be computed iteratively:

$$p_{t+1} = (1 - r)\mathbf{A} \times p_t + r \times p_0, \tag{5.4}$$

whereby r denotes the restart probability ($r = 0.7$). For $t \rightarrow \infty$ the state probabilities converge to a stationary distribution p_∞ that can be written as:

$$p_\infty = (\mathbf{I} - ((1 - r)\mathbf{A}))^{-1} \times r \times p_0. \tag{5.5}$$

The matrix $\mathbf{I}$ denotes the identity matrix and the starting probabilities p_0 were set to 1 for g_i and 0 for all other genes. For two genes g_i and g_j, we define a symmetric PPI distance score by taking the average of the probabilities to encounter g_j when starting at g_i and vice versa.

5.2.8 Binary and discriminative random forest classifier

We first developed a binary random forest classifier [118] for the problem of deciding whether a single gene is an enhancer target based on the four features genomic distance to an enhancer, CSS, PPI distance, and GO similarity ('binary RF'). The classifier learns to predict the class from the four features and to output the ratio of trees which voted for this class. In case of missing values, we assigned the median GO similarity or PPI distance value between p300 and all other genes for the respective feature. We used an implementation of Breiman's algorithm that uses random selection of features at each node to determine a split [119] (R package 'randomForest', version 4.5-34), to train a random forest of 1000 randomly generated decision trees. The final prediction was made by selecting among all genes in the interval the one with the highest probability (i.e., the highest number of trees voting for it).

The binary RF can yield only a yes/no decision as to whether a gene is an enhancer target or not, and is not designed to rank all the candidate genes in the interval. We therefore implemented a second classifier ('discriminative RF') that evaluates each gene pair g_i and g_j in the interval using feature values as well as pairwise rankings and then decides among the following outcomes:

1. g_i is the target gene
2. g_j is the target gene
3. neither g_i nor g_j is the target

This RF takes 12 input features, corresponding to 8 feature values for both genes (genomic distance, CSS, GO similarity, and PPI similarity for g_i and for g_j) and 4 features that assign g_i either to *winner* (W) or *looser* (L) or *tied* (equal, E) in the comparison with the respective feature of g_j. Since GO and PPI annotations are incomplete, we added two labels 'W?' and 'L?' to model the uncertainty that is associated with gene pairs for which one value is missing ('NA'). Then, for each of the four comparisons between genes g_i and g_j, a feature f is assigned to g_i:

$$f = \begin{cases} W & \text{if value}(g_i) > \text{value}(g_j) \\ L & \text{if value}(g_i) < \text{value}(g_j) \\ E & \text{if value}(g_i) = \text{value}(g_j) \\ W? & \text{if value}(g_i) \geq \text{median and value}(g_j) = \text{'NA'} \\ L? & \text{if value}(g_i) < \text{median and value}(g_j) = \text{'NA'} \end{cases}$$

A random forest of 1000 trees was trained using these 12 features. The output consisted of the probabilities for the three classes and the class with the majority vote. The final prediction was made by summing over all probabilities for target gene assignments in pairwise comparisons for all pairs in the interval and reporting the gene with the highest sum as the target gene. A schematic overview of both classifiers is shown in Figure 5.3.

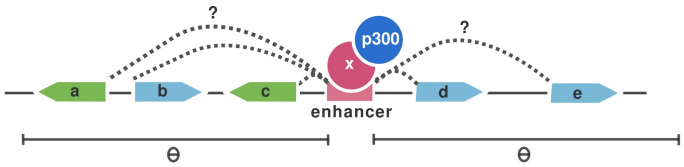
5.2.9 Statistical analysis

For evaluation of various values of the maximal distance parameter Θ on the forebrain and limb data, we used only the p300 enhancers with distance $< \Theta$ to a differentially upregulated gene. Depending on the distance parameter Θ, it may be that multiple differentially upregulated genes are located in a given genomic interval. In such cases, we counted the prediction as a 'correct prediction' if at least one of the upregulated genes was unambiguously predicted as a target gene by any of the prediction methods.

We calculated the precision of a method as $\frac{N_{\text{correct predictions}}}{N_{\text{predictions}}}$ and recall as $\frac{N_{\text{correct predictions}}}{N_{\text{enhancer}}}$. Precision indicates the probability that a prediction is correct and recall denotes the ratio of enhancers for which a correct prediction could be made. Precision and recall values are highly similar for most methods, only differing in cases where multiple genes are assigned the same maximal score by a method. These cases were counted as 'no prediction' in the precision and recall calculations.

For the training of the random forest classifiers we split the enhancer sets into 80% training samples and calculated the precision and recall values on the remaining 20% validation samples. This was done ten times, the values in Figure 5.7 represent the means of the different iterations. For both models we subsampled the training set so that each possible outcome occurred an equal amount of times.

A



B

Gene	d	CSS	PPI	GO	RF class (ratio)
a	152762	2.18	2.0	-15.8	1 (0.80)
b	49140	3.9	1.7	-19.6	1 (0.93)
c	31918	0.9	1.3	-15.8	1 (0.74)
d	61411	0	0.6	-12.7	1 (0.55)
e	235807	0	1.6	-14.8	0

Gene	a	b	c	d	e
Prediction score	0.8	0.93	0.74	0.55	0

C

Row	Genes g_i	Genes g_j	Gene g_i d	Gene g_i CSS	Gene g_i PPI	Gene g_i GO	Gene g_j d	Gene g_j CSS	Gene g_j PPI	Gene g_j GO	Comparison g_i vs g_j d	CSS	PPI	GO	RF class (ratio)
1	a	b	152762	2.18	2.0	-15.8	49140	3.9	1.7	-19.6	W	L	W	L	2 (0.36)
2	a	c	152762	2.18	2.0	-15.8	31918	0.9	1.3	-15.8	W	W	W	E	1 (0.63)
3	a	d	152762	2.18	2.0	-15.8	61411	0	0.6	-12.7	W	W	W	L	1 (0.43)
4	a	e	152762	2.18	2.0	-15.8	235807	0	1.6	-14.8	L	W	W	L	1 (0.77)
5	b	c	49140	3.9	1.7	-19.6	31918	0.9	1.3	-15.8	W	W	W	L	1 (0.63)
6	b	d	49140	3.9	1.7	-19.6	61411	0	0.6	-12.7	L	W	W	L	1 (0.55)
7	b	e	49140	3.9	1.7	-19.6	235807	0	1.6	-14.8	L	W	W	L	1 (0.88)
8	c	d	31918	0.9	1.3	-15.8	61411	0	0.6	-12.7	L	W	W	L	0
9	d	e	61411	0	0.6	-12.7	235807	0	1.6	-14.8	L	E	L	W	0

Gene	a	b	c	d	e
Prediction score	1.83	2.42	0	0	0

Figure 5.3: **(A)** Five genes a, b, c, d, e are located in a genomic window defined by a region of Θ kb around a putative enhancer. **(B)** The binary classifier uses the four features genomic distance (d), conserved synteny (CSS), proximity in PPI-network (PPI), and GO similarity (GO) to assign each gene to one of the two RF classes 'target gene' (1) and 'not-target gene' (0). The ratio of votes amongst 1000 randomized decision trees for a predicted target gene is taken as a prediction score and the gene with highest score (gene b) is finally reported as target. **(C)** The discriminative random forest classifier evaluates all gene pairs in the interval using the four feature values for both genes and four features specifying the comparison between the two genes. In order to use information about the comparison between g_i and g_j, this information has to be explicitly encoded as an additional feature. The classifier is trained to decide whether g_i (1) or g_j (2) is target or if neither of them is the target (0). The final prediction score is then computed as the sum of all voting ratios for all predicted target genes and the gene with highest score is reported as target. For instance, gene b is predicted to be the target in rows 1 and 5–7, and its prediction score is thus 0.36+0.63+0.55+0.88=2.42; likewise, the prediction score for gene a, which was the winner in rows 2–4, is 0.63+0.43+0.77=1.83. Genes c, d, and e did not win in any comparison, and therefore receive a score of zero. Gene b is chosen as the overall prediction because it has the highest overall score.

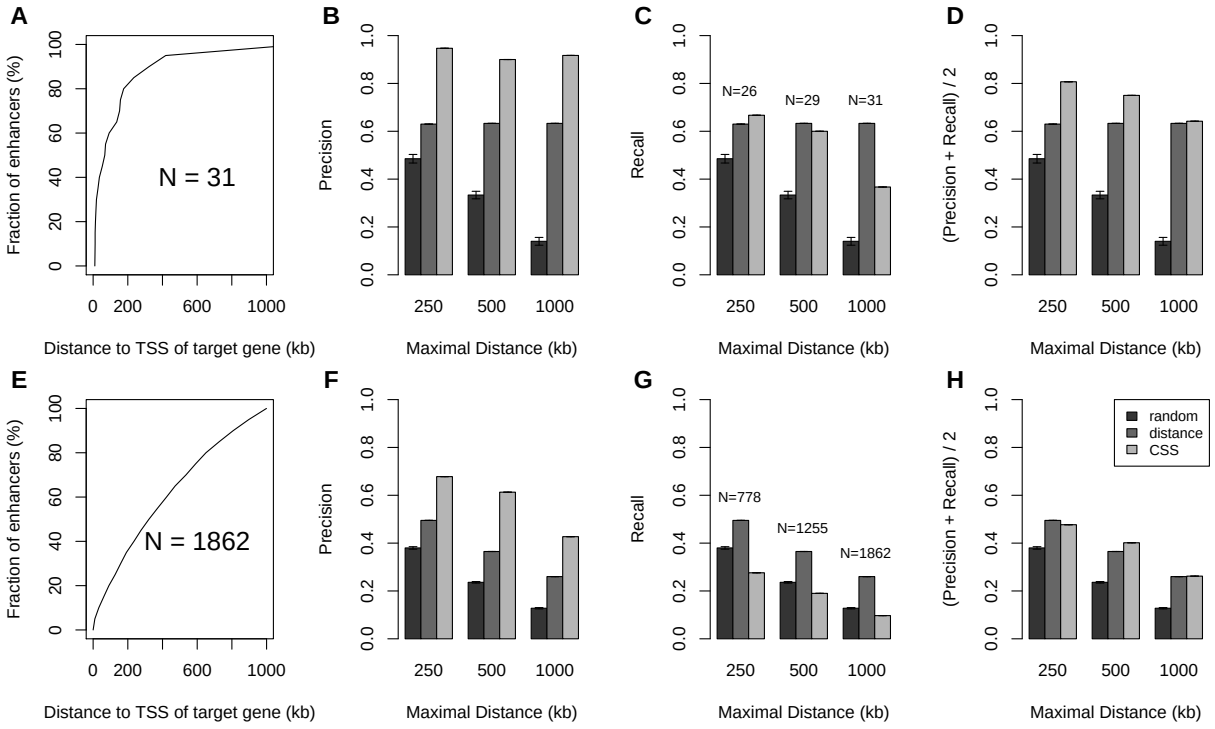


Figure 5.4: A) Distance distribution between enhancers and target genes for 31 regulatory interactions from the literature. B-D) Evaluation of precision (B), recall (C), and average precision and recall (D) for predictions based on conserved synteny, genomic distance, or random predictions of a gene in the genomic interval defined by a maximal distance threshold $\Theta = \{250, 500, 1000\}$ kb around the enhancer. E) Distance distribution between p300 enhancers and putative target genes [20]. F-H) Evaluation of precision (F), recall (G), and average precision and recall (H) on different sets of p300 enhancers. Although literature data and ChIP-seq data show substantially different distributions of enhancer target gene distances, conserved synteny shows the highest precision values for both data sets.

5.3 Results

5.3.1 Conserved synteny predictions of enhancer targets have low recall

Previously identified candidate enhancer regions have been shown to be enriched in the vicinity of transcription factors and developmental genes [120, 32] and to maintain conserved synteny [100, 101]. However, it is not clear to what degree conserved synteny or genomic proximity can be used to predict target genes. We therefore initially compared the performance of predictions based on genomic proximity (i.e., the nearest gene is taken to be the target of an enhancer), conserved synteny, and randomly choosing one of the genes in a window around the enhancer. The conserved synteny score (CSS) was calculated on the basis of a conserved association between the enhancer and the promoter regions of putative target genes in related genomes (Figure 5.1). We also evaluated ortholog predictions based on protein sequence similarity, but this approach showed

slightly lower precision and recall values than using the genomic alignments of promoter sequences [1]. The conserved syntenies were weighted by the evolutionary distances between mouse and the target species. For each of the three methods, we evaluated the quality of predictions by calculating precision and recall for various maximal distance thresholds Θ that define a genomic window centered around the enhancer region. Several studies have outlined that HCNEs and thus putative enhancers span genomic regions of several hundreds of kilobases around their target genes [120, 32], we therefore chose to assess the performance of predictions for $\Theta \in \{250, 500, 1000\}$kb. We chose an arbitrary maximal cutoff of $\Theta = 1000$kb because the great majority of experimentally validated enhancer/target gene pairs are separated by less than this amount (c.f. Fig. 5.4A).

At present, there is no database of enhancer targets, and information in the literature is sparse. We therefore compiled a set of 31 known enhancer-target gene interactions from the literature in order to estimate the quality of predictions that are based on genomic distance or conseved synteny. We included all interactions from either human or mouse that were identified either by observations of phenotypes due to genomic rearrangements [99], similar activation and expression pattern of enhancer and target gene [121], and 3C experiments [47]. We assumed that enhancer activities are conserved in human and mouse and mapped human enhancers to the homologous sequences using blastz alignments [109] (see Table 5.1 for a list of the 31 experimentally validated enhancer-target gene interactions). Synteny-based predictions showed a precision around 90% in contrast to about 61% for genomic distance (Figure 5.4 A-D). However recall values for synteny-based predictions only reach a level of 69% for $\Theta = 250$kb, 62% for $\Theta = 500$kb, and only 37% for $\Theta = 1000$kb, probably because it is not usually possible to assign an enhancer unambiguously to a single target gene on the basis of synteny alone. In all comparisons, predictions based on CSS and genomic distance perform substantially better than random.

Binding by the transcriptional coactivator p300 is thought to be a marker for enhancer activity. For instance, in one series of lacZ reporter gene assays in transgenic mice, 75 of 86 (87%) p300 ChIP-seq peaks showed enhancer activity [20].

In the following we will refer to the p300 ChIP-seq peaks as 'p300 enhancers'. It should be noted that a p300 ChIP-seq peak does not necessarily represent a biologically relevant enhancer, which is a limitation of our approach. For our evaluation, we extracted 1,862 enhancers from a set of about 4,500 p300 enhancers identified by Visel *et al.* [20] under the assumption that upregulated genes located in the same genomic region as p300 enhancers represent the target genes. Although the target genes of p300 enhancers are likely to also be differentially expressed, we note that this assumption may not be correct in all cases because the upregulation can be due to secondary effects. As with the gold-standard targets from the literature, we compared predictions for various maximal distance thresholds $\Theta \in \{250, 500, 1000\}$kb that define a genomic window centered around the p300 enhancer, and assessed the performance of synteny-based and genomic proximity-based predictions relative to random guessing. Figure 5.4 (E–H) shows precision and recall values for the predictions based on each of the two features and random guessing for the merged p300 enhancers from limb and forebrain. In agreement with our observations on the known target gene interactions, conserved synteny alone exhibits a higher precision compared to the use of distance alone. However conserved synteny could only unambiguously assign a minority of enhancers to their target genes leading to a recall of less than 20% for $\Theta = 1000$kb.

The difference in the quality of predictions for the two sets is likely to be related at least partially to the different distribution of distances between enhancer and target gene (Figure 2A and 2E). The results do suggest that using conserved synteny or genomic distance alone is not able to generate

accurate target gene predictions for the p300 enhancers.

Enhancer ID	mm9 coordinate	d_g to TSS (bp)	RefSeq	Gene	Reference
Kammandel_a	chr2:105505190-105505488	11113	NM_013627	Pax6	[122]
Kammandel_b	chr2:105504812-105505044	11557	NM_013627	Pax6	[122]
Kleinjan2004CE1	chr2:105526748-105527377	10147	NM_013627	Pax6	[123]
Kleinjan2004CE2	chr2:105528098-105528997	11497	NM_013627	Pax6	[123]
Kleinjan2004CE3	chr2:105529990-105530599	13389	NM_013627	Pax6	[123]
Kleinjan2006DRR	chr2:105609000-105645000	92399	NM_013627	Pax6	[98]
Uchikawa2003N1	chr3:34564105-34564402	15179	NM_011443	Sox2	[124]
Uchikawa2003N3	chr3:34529134-34529727	19199	NM_011443	Sox2	[124]
Uchikawa2003N4	chr3:34577673-34578139	28747	NM_011443	Sox2	[124]
Uchikawa2003N5	chr3:34558703-34559053	9777	NM_011443	Sox2	[124]
Kurokawa2004FM	chr14:49358192-49357895	75348	NM_144841	Otx2	[125]
Kurokawa2004FM2	chr14:49160485-49160364	122062	NM_144841	Otx2	[125]
Kimura2004F5	chr14:49071554-49106377	176170	NM_144841	Otx2	[126]
Kimura2004F4	chr14:49115970-49124899	157648	NM_144841	Otx2	[126]
Kimura2004F3	chr14:49124899-49125099	157448	NM_144841	Otx2	[126]
Kimura2004F11	chr14:49432656-49435079	150109	NM_144841	Otx2	[126]
Kimura2004F12	chr14:49435140-49474248	152593	NM_144841	Otx2	[126]
Dhaene2009SRO133	chr9:98602576-98608624	247455	NM_012020	Foxl2	[47]
Lettice2003shh	chr5:29270070-29741185	476429	NM_009170	Shh	[99]
Benko2009F2	chr11:111447329-111447587	1195936	NM_011448	Sox9	[127]
Rahimov2008	chr1:194967401-194967796	11509	NM_016851	Irf6	[128]
Miller2008	chr5:38243091-38243226	27267	NM_010835	Msx1	[129]
Bourdeau2004a	chr12:16771712-16771726	35966	NM_015764	Greb1	[130]
Bourdeau2004b	chr12:16767336-16767350	40342	NM_015764	Greb1	[130]
Bejerano2006	chr13:117422451-117422651	322555	NM_021459	Isl1	[121]
Uemura2005CREST1	chr13:116877737-116877973	221923	NM_021459	Isl1	[131]
Uemura2005CREST2	chr13:116738227-116738952	360944	NM_021459	Isl1	[131]
Magee2006	chr17:28557009-28557672	65361	NM_010220	Fkbp5	[132]
DeVal2004	chr13:83711257-83711407	68064	NM_025282	Mef2c	[133]
Sumiyama2003	chr11:94998597-94999132	17149	NM_010055	Dlx3	[134]
Forghani2001SCE1	chr18:82711659-82715634	67145	NM_001025245	Mbp	[135]

Table 5.1: **Collection of known enhancer target gene interactions from literature.** DNA sequences from the publications were searched in the genomes of either mouse (mm9) or human (hg18) using the BLAT application at the UCSC website [136, 137] and in case of human subsequently mapped to the mouse coordinate using pairwise blastz alignments for mouse and human downloaded from the UCSC genome browser [109, 137]. We assigned each enhancer an ID and computed the genomic distance d_g to the to the transcription start site (TSS) of the described target gene.

5.3.2 GO similiarity and proximity in PPI networks may be used to improve prediction of enhancer target genes

The above mentioned results demonstrated that genomic distance and conserved synteny are of limited utility in predicting the target genes of p300 enhancers. Although CSS has reasonably high precision values, it often fails to unambiguously assign an enhancer to a target gene because multiple genes in the interval exhibit the same degree of conserved synteny. This accounts for

20-50% of p300 enhancers and thus represents a major limitation in the use of conserved synteny. For instance, seven p300 enhancers for limb tissues are located in the *Sox9* locus and might account for the upregulation of *Sox9*, observed by Visel *et al.* [20]. An analysis based on genomic proximity alone would not identify Sox9 as the target, and an analysis based on conserved synteny would identify up to five additional genes in the vicinity as potential targets. In some cases, transcription factors regulate genes with which they also physically interact, e.g. Runx2 and Dlx5 [77, 74]. We therefore hypothesized that p300 and its targets are located more proximal to each other in PPI networks than p300 and non-target genes. We additionally hypothesized that functional similarity between p300 and targets is greater than between p300 and non-targets. p300 has a number of GO annotations related to organ development, regulation of transcription factor activity, response to stimuli including calcium, transcription cofactor activity, and others [1]. GO analysis of the limb and forebrain upregulated genes from Visel *et al.* [20] revealed that both sets are significantly enriched in terms such as 'developmental process' and 'transcription factor activity'. These observations reflect the known role of p300 in development [138, 139].

If we take the upregulation of Sox9 in the above mentioned experiment as evidence that Sox9 is the target gene of the enhancers, then the observation that Sox9 is a direct protein interaction partner of p300, and that it shares a number of GO annotations with p300 could be used to identify Sox9, and not one of the other five genes showing conserved synteny, as the correct target gene. This observation motivates our approach (Fig. 5.5).

We therefore tested whether GO similarity and PPI distance can be used to resolve the ambiguity in cases where CSS fails to unambiguously assign an enhancer to a target gene. Figure 5.6 shows that in case of ties in CSS, target genes show higher GO similarity and are closer to p300 in PPI networks than non-target genes. This observation motivated us to develop an integrative approach that combines all four features in a random forest classifier.

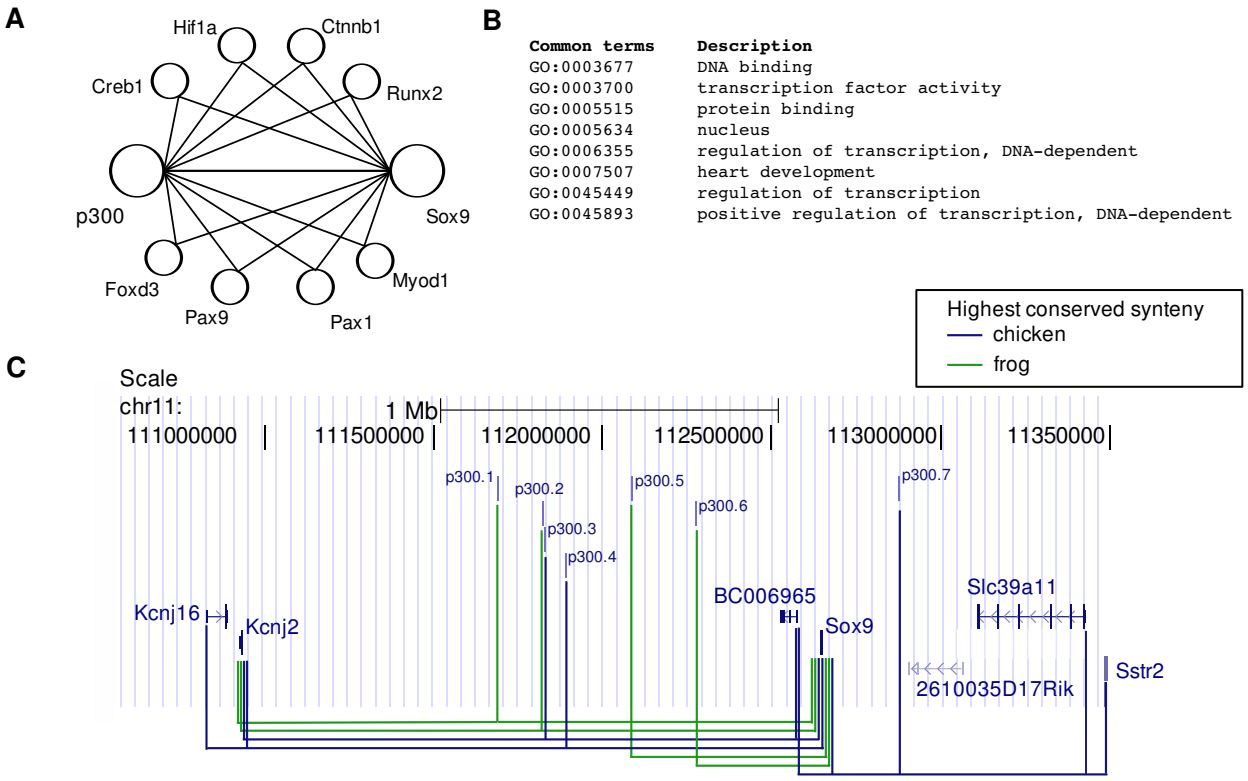


Figure 5.5: We hypothesize that the enhancer binding protein and its target genes show a tendency for shared functions such as 'transcription factor activity' and are located in the vicinity of one another in the protein-protein interactome. This is illustrated in the example of the potential regulation of *Sox9* by p300. A) Protein-protein interactions involving p300 and SOX9. p300 and SOX9 directly interact with one another [140], and also share a number of known or predicted intermediary interaction partners in the protein interaction network [116]. B) SOX9 has 21 GO annotations, and p300 has 35 GO annotations. The 8 shared annotations are shown. C) UCSC Genome Browser view on the *Sox9* locus. Seven p300 enhancers from mouse limb tissue [20] show the highest degree of conserved synteny with the *Sox9* promoter region however only the enhancers p300.5 and p300.6 can unambiguously be assigned to the *Sox9* promoter. For the remaining enhancers multiple genes including *Sox9* exhibit the same degree of conserved synteny. However, high GO similarity between p300 and *Sox9* as well as their proximity in the PPI network suggest that the target gene of these enhancers is *Sox9*.

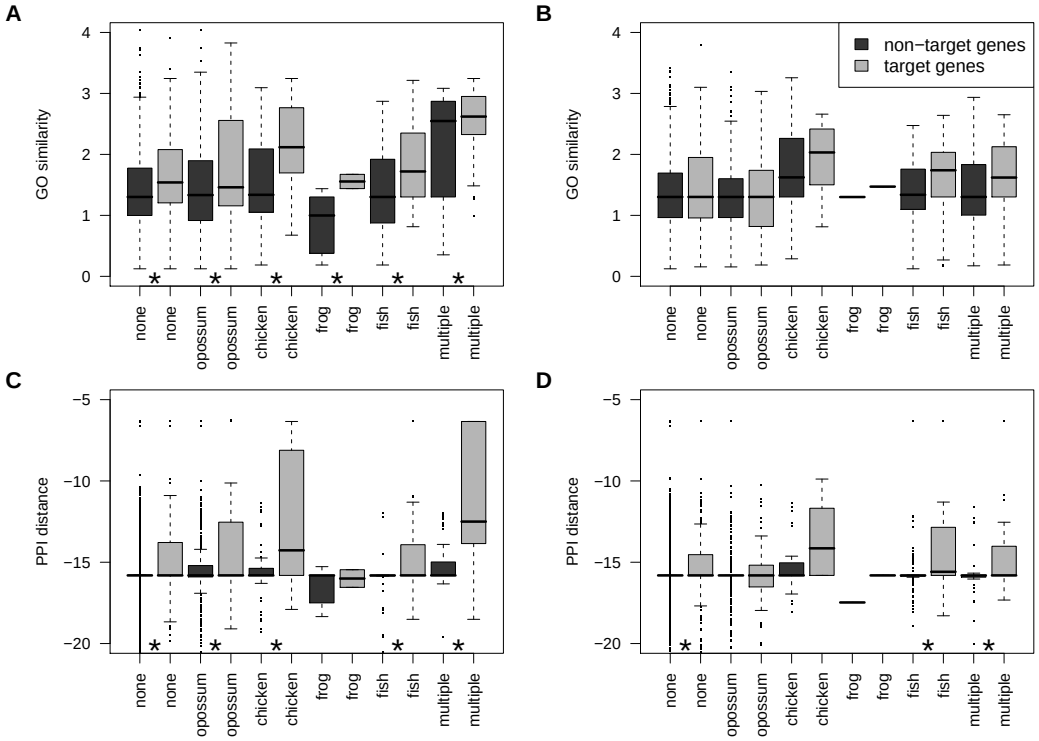


Figure 5.6: Since conserved synteny often fails to unambiguously predict a target gene, we tested whether GO similarity and PPI distances may help to resolve cases where multiple genes exhibit equal degrees of conserved synteny. For p300 enhancers from limb and forebrain we identified all genes in intervals at $\Theta = 500$kb with highest CSS but that cannot uniquely be assigned to the enhancer. We grouped this set into CSS classes that correspond to evolutionary distances from Figure 5.2.9 with the exception that the label 'fish' indicates CSS $\in]1.72, 2.3]$ and 'multiple' corresponds to CSS > 2.3 . (A,B) GO similarities for target and non-target genes for p300 limb (A) and forebrain (B) enhancers. (C,D) PPI-distance for target and non-target genes for p300 limb (C) and forebrain (D) enhancers. Comparison of target genes vs. non-target genes within these subsets showed for all subclasses that target genes show a tendency for higher GO similarity and closer distances in PPI networks. $'*' := P < 0.05$, Wilcoxon-test with Benjamini-Hochberg multiple testing correction.

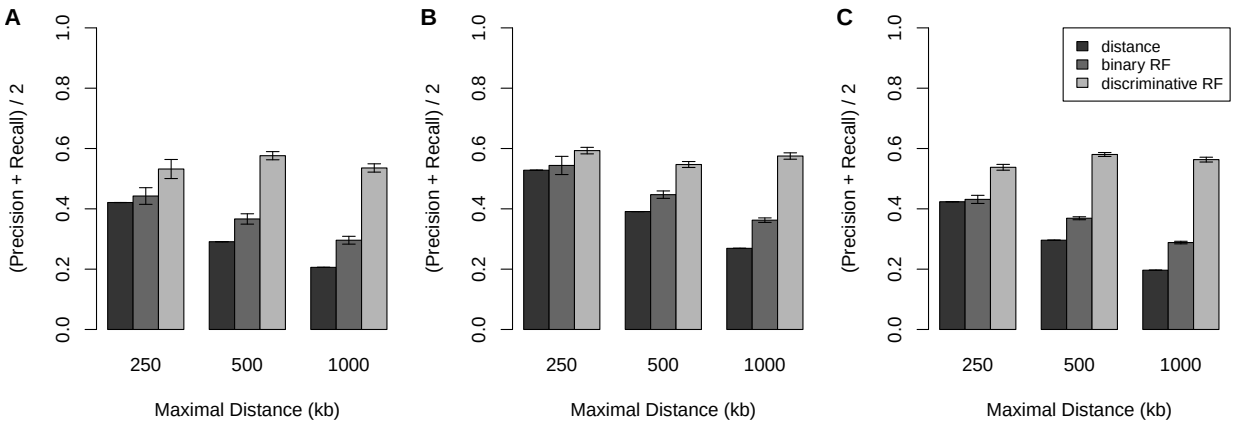


Figure 5.7: Evaluation of random forest classifier predictions on p300 ChIP-seq data for A) limb, B) forebrain, and C) 2430 Gli3 bound regions identified from ChIP-chip experiments [108]. Average precision-recall values for predictions based on distance, and two random forest classifiers are shown. For the random forest models, the data was split into 80% training and 20% validation sets. The results shown are mean values after 10 repeated evaluations and standard errors. Combination of genomic, functional, and protein interactome data allows correct target gene identification in 56-61% of cases for genomic intervals of 2Mb.

5.3.3 Accurate target gene prediction using random forest classifiers and combination of features

Decision tree induction is a supervised learning method for classifying data. During the learning phase a tree is constructed iteratively, whereby at each node a test is derived that splits the local training set into two subsets so that the heterogeneity of the resulting subsets is minimized. Typically the learning phase is stopped as soon as the heterogeneity falls below a certain threshold. Random forests (RF) are an extension of decision trees to collections of trees that use randomization in the selection of features for splitting the learning sample at each node [119]. The final classification is made by taking the majority vote for all trees in the forest [118]. Alternatively, classification probabilities can be defined as the ratio of trees voting for a certain class.

We evaluated two random forest approaches (Figure 5.3). The first was a binary RF classifier which separately calculates the probability of each gene of being a target; these probabilities can then be used to rank the k genes in a given interval according to the probability of being a target gene. The second approach involved a discriminative RF classifier which compares all gene pairs in the interval and chooses the gene that was the most frequently predicted target gene in the set of all pairs (see *Methods*).

In order to make the methods more easily comparable we will use the average precision-recall ($\frac{\text{Precision+Recall}}{2}$) as performance measure, analogously to Schweikert *et al.* [141]. Individual precision and recall values for limb and forebrain target gene predictions can be found in Tables 5.2 and 5.3. Both classifiers were compared to the genomic distance-based method. Figure 5.7 (A,B) shows the average precision-recall values of the three approaches for $\Theta = \{250, 500, 1000\}$kb. Increasing Θ leads to higher number of genes in an interval. For $\Theta = 250$kb an average of 8.5 genes is located in the genomic window around the putative p300 from the limb dataset. This number increases to

25 genes for Θ = 1000kb. Binary and discriminative random forest classifiers show substantially better performance than the distance based approach. This is true for all comparisons of random forest classifiers vs. predictions based on any single feature 5.2 and 5.3).

Since cases have been reported where the distance to the target genes exeeds 1Mb, we also applied the classifier on a putative p300 enhancers that are located up to 2Mb away of the nearest differentially expressed gene. The average precision-recall value stays almost constant at a level of 58% [1].

We also applied the classifier on 1372 p300 forebrain enhancers and 1324 limb enhancers that are not in proximity of an upregulated gene (Θ = 1000kb). The predictions include previously reported *Bmp7* limb enhancer and a *Sox2* enhancer, that is active in rhombencephalon [142, 124], suggesting that at least a proportion of the predictions is valid.

Method	**Θ** kb	**Precision**	**Recall**	**Average** $\frac{\mathbf{P+R}}{\mathbf{2}}$
Random	250	0.35 ± 0.02	0.35 ± 0.02	0.35 ± 0.02
Distance	250	0.42	0.42	0.42
CSS	250	0.62	0.2	0.41
GO	250	0.42	0.41	0.41
PPI	250	0.48	0.41	0.44
binary RF	250	0.44 ± 0.09	0.44 ± 0.09	0.44 ± 0.09
discriminative RF	250	0.53 ± 0.1	0.53 ± 0.1	0.53 ± 0.1
Random	500	0.22 ± 0.01	0.22 ± 0.01	0.22 ± 0.01
Distance	500	0.29	0.29	0.29
CSS	500	0.53	0.13	0.33
GO	500	0.31	0.31	0.31
PPI	500	0.28	0.26	0.27
binary RF	500	0.37 ± 0.05	0.36 ± 0.05	0.37 ± 0.05
discriminative RF	500	0.58 ± 0.04	0.58 ± 0.04	0.58 ± 0.04
Random	1000	0.11 ± 0.01	0.11 ± 0.01	0.11 ± 0.01
Distance	1000	0.21	0.21	0.21
CSS	1000	0.35	0.06	0.2
GO	1000	0.22	0.22	0.22
PPI	1000	0.19	0.19	0.19
binary RF	1000	0.3 ± 0.04	0.29 ± 0.04	0.3 ± 0.04
discriminative RF	1000	0.54 ± 0.04	0.54 ± 0.04	0.54 ± 0.04
Random	2000	0.05 ± 0.01	0.05 ± 0.01	0.05 ± 0.01
Distance	2000	0.15	0.15	0.15
CSS	2000	0.23	0.03	0.13
GO	2000	0.15	0.15	0.15
PPI	2000	0.14	0.13	0.13
binary RF	2000	0.24 ± 0.05	0.24 ± 0.05	0.24 ± 0.05
discriminative RF	2000	0.58 ± 0.03	0.58 ± 0.03	0.58 ± 0.03

Table 5.2: Target gene prediction on putative p300 limb enhancers. Precision and recall values for predictions based genomic distance, conserved synteny score (CSS), the binary random forest classifier (RF), and discriminative RF classifier. For random selection of target genes and random forest classifiers means and standard deviation are shown for 10 repeated evaluations.

Method	**Θ** kb	**Precision**	**Recall**	**Average** $\frac{\mathbf{P+R}}{\mathbf{2}}$
Random	250	0.39 ± 0.01	0.39 ± 0.01	0.39 ± 0.01
Distance	250	0.53	0.53	0.53
CSS	250	0.69	0.32	0.51
GO	250	0.46	0.46	0.46
PPI	250	0.51	0.44	0.47
binary RF	250	0.55 ± 0.09	0.54 ± 0.1	0.54 ± 0.1
discriminative RF	250	0.59 ± 0.03	0.59 ± 0.03	0.59 ± 0.03
Random	500	0.25 ± 0.01	0.25 ± 0.01	0.25 ± 0.01
Distance	500	0.39	0.39	0.39
CSS	500	0.63	0.22	0.43
GO	500	0.34	0.34	0.34
PPI	500	0.4	0.36	0.38
binary RF	500	0.45 ± 0.04	0.44 ± 0.04	0.45 ± 0.04
discriminative RF	500	0.55 ± 0.03	0.55 ± 0.03	0.55 ± 0.03
Random	1000	0.13 ± 0.01	0.13 ± 0.01	0.13 ± 0.01
Distance	1000	0.27	0.27	0.27
CSS	1000	0.43	0.12	0.28
GO	1000	0.23	0.22	0.22
PPI	1000	0.26	0.24	0.25
binary RF	1000	0.36 ± 0.02	0.36 ± 0.02	0.36 ± 0.02
discriminative RF	1000	0.58 ± 0.03	0.58 ± 0.03	0.58 ± 0.03
Random	2000	0.07 ± 0.01	0.07 ± 0.01	0.07 ± 0.01
Distance	2000	0.2	0.2	0.2
CSS	2000	0.31	0.06	0.18
GO	2000	0.14	0.14	0.14
PPI	2000	0.15	0.15	0.15
binary RF	2000	0.29 ± 0.03	0.29 ± 0.03	0.29 ± 0.03
discriminative RF	2000	0.58 ± 0.02	0.58 ± 0.02	0.58 ± 0.02

Table 5.3: Target gene prediction on putative p300 forebrain enhancers. Precision and recall values for predictions based genomic distance, conserved synteny score (CSS), the binary random forest classifier (RF), and discriminative RF classifier. For random selection of target genes and random forest classifiers means and standard deviation are shown for 10 repeated evaluations.

5.3.4 Prediction of Gli3 target genes in a limb ChIP-chip dataset

To test whether our method might be applicable to experiments with other enhancers, we analyzed a ChIP-chip data set for Gli3 in mouse limbs [108]. Gli3 is a transcription factor that is activated upon Shh signaling which specificies the anterior posterior axis in the developing limb bud and thus regulates the number of digits. Vokes *et al.* defined a high quality set of 5274 Gli3 bound regions of which 2430 are located $<$ 1Mb away of an Shh repsonsive gene as identified by differential expression [108]. Similar to the p300 data, conserved synteny predicts targets with higher precision but lower recall and the random forest approaches showed substantially better performance than any single feature based prediction (Figure 5.7 C, Table 5.4).

Method	**Θ** kb	**Precision**	**Recall**	**Average** $\frac{\mathbf{P+R}}{\mathbf{2}}$
Random	250	0.34 ± 0.02	0.34 ± 0.02	0.34 ± 0.02
Distance	250	0.42	0.42	0.42
CSS	250	0.55	0.16	0.35
GO	250	0.41	0.4	0.4
PPI	250	0.39	0.35	0.37
binary RF	250	0.43 ± 0.04	0.43 ± 0.04	0.43 ± 0.04
discriminative RF	250	0.54 ± 0.03	0.54 ± 0.03	0.54 ± 0.03
Random	500	0.21 ± 0.01	0.21 ± 0.01	0.21 ± 0.01
Distance	500	0.3	0.3	0.3
CSS	500	0.38	0.09	0.24
GO	500	0.31	0.3	0.3
PPI	500	0.28	0.25	0.26
binary RF	500	0.37 ± 0.01	0.37 ± 0.02	0.37 ± 0.02
discriminative RF	500	0.58 ± 0.02	0.58 ± 0.02	0.58 ± 0.02
Random	1000	0.11 ± 0.01	0.11 ± 0.01	0.11 ± 0.01
Distance	1000	0.2	0.2	0.2
CSS	1000	0.24	0.04	0.14
GO	1000	0.23	0.23	0.23
PPI	1000	0.18	0.17	0.17
binary RF	1000	0.29 ± 0.01	0.29 ± 0.01	0.29 ± 0.01
discriminative RF	1000	0.56 ± 0.03	0.56 ± 0.03	0.56 ± 0.03
Random	2000	0.07 ± 0.01	0.07 ± 0.01	0.07 ± 0.01
Distance	2000	0.14	0.14	0.14
CSS	2000	0.16	0.02	0.09
GO	2000	0.18	0.18	0.18
PPI	2000	0.13	0.12	0.13
binary RF	2000	0.24 ± 0.02	0.24 ± 0.02	0.24 ± 0.02
discriminative RF	2000	0.52 ± 0.08	0.51 ± 0.08	0.52 ± 0.08

Table 5.4: Target gene prediction on putative Gli3 bound sites. Precision and recall values for predictions based genomic distance, conserved synteny score (CSS), the binary random forest classifier (RF), and discriminative RF classifier. For random guessing and random forest classifiers means and standard deviation are shown for 10 repeated evaluations.

5.4 Discussion

Current research on long-range regulatory interactions is strongly focused on the computational detection and experimental validation of *cis*-regulatory elements [91, 94]. ChIP-seq experiments on the transcriptional coactivator p300 have proven to be a highly reliable method for experimental detection of enhancer regions in various tissues [67, 20], but the identified sequences still have to be linked to their transcriptional targets.

Previous studies have postulated that evolutionarily constraints on enhancer-target gene interactions are likely to be responsible for the maintainance the conserved synteny in large genomic intervals [100, 101, 107]. Kikuta *et al.* and Akalin *et al.* defined target genes as transcription factors with an HCNE density peak in human-zebrafish conserved-syntenic regions that were termed genomic regulatory blocks (GRBs) [101, 107]. This is in agreement with the observations that HCNEs are clustered around developmental genes and transcription factors [143, 32], but it may not reflect the general pattern of enhancer-target gene interactions since previously defined GRBs [107] do not represent an exhaustive genome-wide collection of genes targeted by long-range regulation. For example, only 579 (12.7%) of ChIP-seq peaks from limb and forebrain overlap with aligned regions between mouse and zebrafish genomes and only 64 (7.7%) of upregulated genes in limb and forebrain overlap with the mouse orthologs of the GRB target genes. Therefore p300 ChIP-seq data defines a more general class of enhancer-target gene interactions that are less conserved and not exclusively restricted to transcription factors.

Two observations prompted us to use proximity in PPI networks and GO functional similarity as features for predicting enhancer targets. Feed forward loops and autoregulatory loops are common in gene regulatory networks [144]. p300 binding in genomic regions that display conserved synteny with the Sox9 promoter suggests that it could be involved in the activation of SOX9 transcription (Figure 5.5). p300 also directly interacts with the SOX9 protein [140] and shares a number of known or predicted intermediary interaction partners in the protein interaction network [116]. This suggested the hypothesis that the enhancer binding protein might display a relative proximity to its targets in the protein interaction network. Second, we hypothesized that the regulator would have a higher GO similarity to its targets than to non-target genes. Although GO similarity alone predicts target genes at larger distances ($\Theta > 500$kb) with comparable recall values as genomic distance (Tables 5.2, 5.3, and 5.4), it cannot be utilized to predict non-transcription factor targets of very specific functions of p300 targets that are involved in cell adhesion [145] or erythropoiesis [146]. The motivation of the random forest approach was therefore to exploit the complementary aspects of the four features. Our results demonstrate that the combination of features dramatically improved the prediction of target genes in genomic intervals of up to 2Mb centered at the location of a p300 enhancer, with a recall of 58% compared to only 27% for genomic proximity and 12% for conserved synteny (Table 5.3). The analysis of a second data set on Gli3 binding in embryonic mouse limbs displayed a similar advantage for the random forest predictions.

Since available data on enhancer-target gene interactions is extremely limited, we chose to interpret an upregulation of a gene in the vicinity of an enhancer to be the effect of direct regulation. This represents a limitation of our study, as the assumption that genes not found to be differentially expressed are not target genes may be incorrect, for instance because the differential expression may occur at a time point that was not measured. Another limitation is the assumption of our model that enhancers can regulate only one target gene. Enhancers may be active in various tissues [32] and multiple enhancers may coordinate the expression of one target gene [47]. Nevertheless, under the assumptions of our study, we have shown that genomic distance, conserved synteny, PPI

distance, and functional similarity can be combined to dramatically improve predictions of the target genes.

The random forest classifiers that have been trained on limb and forebrain enhancers cannot be directly transferred to enhancer-target gene prediction in other tissues. Using the limited data now available, we have observed that the random forest classifiers are specific for the immunoprecipitated factor but also for the tissue [1]. However, more data will be needed to evaluate if this reflects variability between experiments or tissue-specificity characteristics of regulatory interactions. With this proviso, our methodology can be applied to new ChIP-seq data to prioritize candidate enhancer-target gene interactions for validation experiments, and may also be useful for assessing the most biologically relevant hits identified by high-throughput chromosome conformation capture assays that have been developed to globally map chromatin interactions [106, 147, 148]. ChIP-seq is still a relatively new protocol and contains biases that are poorly understood [25], however with more experimental data sets becoming publicly available, more detailed analyses can be performed to further evaluate how to combine functional classification of binding events and the association to target genes into an integrative downstream analysis of ChIP-seq experiments.

Chapter 6

Discussion and Conclusion

Genome-wide approaches to map protein-DNA interactions as well as histone modifications have been shown to greatly enhance our knowledge of transcriptional regulation. It has partially shifted the focus from investigation of promoter sequences of coregulated genes [6] to the impact of *cis*-regulatory modules that may be located hundreds of kilobases away of their target gene. Currently one of the greatest challenges in the analysis of ChIP-seq data seems to be the classification of several thousands of binding events [43, 23, 34] into functional and non-functional binding [31]. In the context of this study a binding event was considered as functional if it is associated with the regulation of a target gene. It has to be noted that transcription factors may also serve other functions than just directly regulating gene expression levels. The transcription factor Runx2 for example was initially identified as a nuclear matrix protein [149], which does not exclude that this role also has some impact on its regulatory function.

In chapter 2 we have looked more closely into the observed ChIP-seq signals that are generated. Most previous studies that employed ChIP-based approaches ignored the ChIP-seq signal strength after the peak-calling step, which basically classifies the genome into bound and unbound states. Theoretically the signal strength corresponds to the fraction of cells that is bound by the transcription factor at the timepoint of immunoprecipitation. This fraction may be influenced by many factors such as the heterogeneity of cell types, the abundances of the transcription factor in each of the cell types, the accessibility of chromatin, and the sequence content of the bound regions. The analysis in chapter 2 shows that predicted binding affinities of highly overrepresented DNA motifs only weakly correlate with the observed ChIP-seq signal ($r \approx 0.3$ with $\log_2$ fold change *vs.* control). When a method was applied that tries to maximize the correlation by greedily selecting k-mers ignoring any synergistic effects between different motif combinations, it was able to improve the correlation between predicted binding affinity and the ChIP-seq signal to $r \approx 0.5$. However, based on the analysis of two different data sets, this suggest that still a large fraction of the variance in ChIP-seq signals cannot be explained from the DNA sequence alone. We tested whether the accessibility of chromatin also influences the strength of the signal, which is supported by three observations. First we found that the number of neighboring peaks weakly correlates ($r \approx 0.11$) with ChIP-seq signal, second we found that promoter peaks with much higher signal than expected from sequence show a significant enrichment in genes that are upregulated in cells overexpressing Runx2 (this is under the assumption that binding events at open chromatin regions are more likely to be functional), third regions that show Stat1 binding in cells that are not stimulated by IFN-γ show elevated levels of histone marks H3K4me3 and H3K4me1 that are generally associated with

open chromatin. In the case of Runx2 we showed that the closer analysis of the bound sequences identifies an excess of signal that is sequence unspecific and which can be used to prioritize binding events in promoters with respect to their regulatory role.

In chapter 3 we applied existing methods to analyze differential binding between Hoxd13 wildtype and mutant proteins. The analysis showed that not only alterations in the recognition motif of the DNA binding domain can be detected, but also that differences in sequence preference due to loss of cooperative binding with cofactors can be detected. Current databases such as UniPROBE [59], TRANSFAC [150], and JASPAR [151] only provide models for monomer or homodimer binding. Only in a few examples like for Hoxa9 and Meis1, distinct models exist for heterodimer binding [56]. Thus in the lack of a comprehensive data set on combinatorial binding, *de novo* motif analysis of *in vivo* bound regions as detected by ChIP-seq is a powerful tool for the investigation of combinatorial binding. Furthermore we have shown, that combined analysis of motifs and colocalized binding from ChIP-seq experiments for the initial transcription factor and its putative cofactor is a suitable method to validate cofactor interactions. The analysis showed that although the degree of overlap for the Hoxd13 mutant Q317K showed a decreased enrichment relative to the Hoxd13 wildtype peaks (2.5 fold *vs.* 4 fold), the R298Q mutant showed a significant 1.7 fold depletion in colocalized peaks. This indicates that although the interaction is likely to be affected by the Q317K mutation, the effect is much stronger for the Hoxd13 R298Q mutant. It might also suggest that the degree of peak overlaps between two different ChIP-seq experiments may be a better measure to detect quantitative changes in cofactor interactions as compared to an assay such as co-immunoprecipitation.

The question about the functionality of binding events is closely linked to the problem of identification of target genes. Ideally these two problems have to be solved simultaneously. In chapter 4 we have developed a method that ranks classes of putative *cis*-regulatory modules with respect to their impact on gene expression of predicted target genes. we used bound regions from ChIP-seq experiments for six different transcription factors to explore the degree of colocalized binding events and to measure the effect on gene expression. Keeping in mind, that each of these experiments have been carried out in a different cell culture which overexpresses the transcription factor of interest, we interpreted the colocalization of binding events only as a potential combinatorial binding event. The bound regions were divided into classes of putative *cis*-regulatory regions and extended traditional gene set enrichment analysis to the application on ChIP-seq data. The method, termed ChIP-seq enrichment analysis (CSEA), corrects for the influence of variable sizes of gene regulatory domains and in addition allows the comparison of enrichment scores across different module classes. Using this method it is possible to rank module classes with respect to their association with differential expression of the predicted target genes, defined by the nearest transcription start sites. This allowed us to prioritize target genes sets and to identify relevant biological processes that are regulated by the studied transcription factors. Moreover the lack of coherent biological processes in target genes that show binding of Runx2 alone and are upregulated in Runx2 overexpressing cells suggests that this gene set might be heavily biased by dysregulation due to Runx2 overexpression. This is supported by the finding that single factor bound regions are less conserved than regions that can be bound by multiple factors.

This approach does not account for the synergistic effects between regulatory elements. This leads to overlaps in the predicted target gene sets. Synergistic events between enhancers have been shown recently using chromatin conformation capturing (3C) assays [47]. This is also supported by the finding that the number of neighboring peaks is to a certain degree predictive for the ChIP-seq

signal strength.

Thus future extensions would have to integrate the effect of multiple *cis*-regulatory modules within a regulatory domain of a gene. However already in its current form, CSEA can be used to distinguish bound regions with high regulatory impact from non-functional bound regions with low impact.

In chapter 5 we have introduced an alternative approach to target gene assignment based on genomic distance. For ChIP-seq data of the p300 coactivator in mouse embryonic limb and forebrain tissues it had been shown that 75 of 86 (87%) p300 ChIP-seq peaks showed enhancer activity in lacZ reporter assays [20], indicating that most p300 bound sites are indeed functional binding events in contrast to binding events of many other transcription factors. p300 is considered as a coactivator and is recruited to genomic loci by other DNA-binding factors. This has been shown in mouse embryonic stem cells, where a knockout of Nanog, Oct4, and Sox2 by RNAi leads to a loss of p300 binding at locations that are typically bound by these factors [67]. The p300 protein contains an acetyltransferase domain. Acetylation of histone tails has been shown to be correlated with active chromatin [52, 50]. Thus p300 ChIP-seq is a good indicator of enhancer regions but the binding of p300 is more an effect of combinatorial binding events at enhancer regions rather than a dominant factor in the control of combinatorial regulation. The high functionality of p300 bound regions was exploited in a way, that all p300 ChIP-seq peaks were assumed to be indeed functional. In addition it was also assumed, that differential gene expression of a gene in the vicinity of a p300 bound region is an indicator for an enhancer-target gene interaction. Based on these two assumptions we developed an integrative approach that can be used to predict the majority of target genes for p300 peaks. This approach takes the genomic distance between enhancer and target gene, conservation of synteny [100], functional similarities between regulator and target genes, as well as protein interactions into account. Although the trained model shows specificity for tissue and transcription factor, we was able to apply this approach also on ChIP-chip data for the sonic hedgehog signaling effector Gli3 [108]. This suggests that integrative approaches that use available functional as well as biochemical knowledge can help to predict links between regulatory regions and their target genes. Further analysis on the assignment of target genes could involve binding of CTCF. CTCF is a zinc-finger protein with transcriptional activator and repressor activity [152]. More recently a function of CTCF as insulator protein was found, with the role of separating regulatory domains and protecting distant non-target genes from interactions with enhancers [153]. Although the full spectrum of function and the mechanisms of action are only poorly understood, CTCF binding has been used to define regulatory regions for the target gene assignment in ChIP-seq experiments [96]. One interesting property of CTCF is that it is ubiquitously expressed and its binding sites stay largely invariant over developmental tissues and stages [152]. If that was the case, then the role of CTCF as insulator protein should be used with care since several p300 bound limb enhancers that are likely to regulate Sox9 (Figure 5.5) and also known Sox9 enhancers [127] show a number of intervening CTCF bound regions.

In summary we have developed a number of methods and concepts that can be used to distinguish functional binding events over binding events with no obvious impact on gene expression. we have shown that depending on the type of control, the strength of the ChIP-seq signal can be used to identify binding events in regions that correlate with open chromatin features. This knowledge can be used to analyze functional binding for a single experiment. For multiple experiments the ChIP-seq enrichment analysis can be applied to detect classes of putative *cis*-regulatory modules with high functional impact. In conjunction with methods we proposed a novel method to assign

binding events to their target genes. Thus, the methods described in this thesis provide a useful framework to guide the analysis of ChIP-seq data.

A number of questions could be answered by future analysis and more experimental data. For instance what is the impact of various histone marks on the binding of transcription factors. More data than could be analyzed in chapter 2 would be needed to investigate the mutual influence between epigenetic marks and transcription factor binding. With respect to the cooperative binding of transcription factors, additional sequence analysis could reveal deeper insights into the architecture of cooperative binding. Important questions would be, what are the differences between colocalized binding events shared by Runx2, Dlx5, and Msx2 in contrast to regions that are bound by Runx2 alone. Is this dependent on the presence of binding sites for all the factors or is this depending on the mode of Runx2-DNA interaction. One could ask if the interaction with cofactors requires a strong Runx2 motif or rather a weak one. In a number of cases it has been reported that the biologically important binding sites are not the ones with maximal affinity [154, 155].

If the presence of binding sites for each of the factors is needed, what configurations are allowed? Does for example collinearity of binding sites matter? Finally it would be interesting to see whether such a model for combinatorial binding could be used to predict regulatory regions in other genomes and then to check whether such regions show a stronger conservation of transcription factor binding than was found by Schmidt and Wilson *et al.* [156]. The analysis of ChIP-seq data raises a number of questions and especially the filtering of regulatory binding events is still challenging, but it is worth to invest more work into the analysis since ChIP-seq data constitutes a promising resource for asking more fundamental questions about transcription factor binding and regulatory aspects that are associated with it.

Bibliography

[1] Christian Rödelsperger, Gao Guo, Mateusz Kolanczyk, Angelika Pletschacher, Sebastian Köhler, Sebastian Bauer, Marcel H Schulz, and Peter N Robinson. Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions. *Nucleic Acids Res*, Nov 2010.

[2] J. Hecht, V. Seitz, M. Urban, F. Wagner, P. N. Robinson, A. Stiege, C. Dieterich, U. Kornak, U. Wilkening, N. Brieske, C. Zwingman, A. Kidess, S. Stricker, and S. Mundlos. Detection of novel skeletogenesis target genes by comprehensive analysis of a Runx2(-/-) mouse model. *Gene Expr Patterns*, 7(1-2):102–112, Jan 2007.

[3] Daniel W Young, Mohammad Q Hassan, Xiao-Qing Yang, Mario Galindo, Amjad Javed, Sayyed K Zaidi, Paul Furcinitti, David Lapointe, Martin Montecino, Jane B Lian, Janet L Stein, Andre J van Wijnen, and Gary S Stein. Mitotic retention of gene expression patterns by the cell fate-determining transcription factor Runx2. *Proc Natl Acad Sci U S A*, 104(9):3189–3194, Feb 2007.

[4] Mitsuko Suzuki, Naoto Ueno, and Atsushi Kuroiwa. Hox proteins functionally cooperate with the GC box-binding protein system through distinct domains. *J Biol Chem*, 278(32):30148–30156, Aug 2003.

[5] Gao Guo, Sebastian Bauer, Jochen Hecht, Marcel H Schulz, Andreas Busche, and Peter N Robinson. A short ultraconserved sequence drives transcription from an alternate FBN1 promoter. *Int J Biochem Cell Biol*, 40(4):638–650, 2008.

[6] Helge G Roider, Thomas Manke, Sean O'Keeffe, Martin Vingron, and Stefan A Haas. Pastaa: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, 25(4):435–442, Feb 2009.

[7] W. F. Shen, C. P. Chang, S. Rozenfeld, G. Sauvageau, R. K. Humphries, M. Lu, H. J. Lawrence, M. L. Cleary, and C. Largman. Hox homeodomain proteins exhibit selective complex stabilities with Pbx and DNA. *Nucleic Acids Res*, 24(5):898–906, Mar 1996.

[8] Michael F Berger, Anthony A Philippakis, Aaron M Qureshi, Fangxue S He, Preston W Estep, and Martha L Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11):1429–1435, Nov 2006.

[9] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M Vaquerizas, Jian Yan, Mikko J Sillanp, Martin Bonke, Kimmo Palin, Shaheynoor Talukder, Timothy R Hughes, Nicholas M Luscombe, Esko Ukkonen, and Jussi

Taipale. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome Res*, 20(6):861–873, Jun 2010.

[10] A. E. Kel, E. Gssling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*, 31(13):3576–3579, Jul 2003.

[11] Barrett C Foat, Alexandre V Morozov, and Harmen J Bussemaker. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, 22(14):e141–e149, Jul 2006.

[12] Helge G Roider, Aditi Kanhere, Thomas Manke, and Martin Vingron. Predicting transcription factor affinities to dna from a biophysical model. *Bioinformatics*, 23(2):134–141, Jan 2007.

[13] Masashi Yamauchi, Shinji Kawai, Takahiro Kato, Takashi Ooshima, and Atsuo Amano. Odd-skipped related 1 gene expression is regulated by Runx2 and Ikzf1 transcription factors. *Gene*, 426(1-2):81–90, Dec 2008.

[14] V. Jackson. Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent. *Cell*, 15(3):945–954, Nov 1978.

[15] Heng Li, Jue Ruan, and Richard Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851–1858, Nov 2008.

[16] O. Owolabi and D.R. McGregor. Fast approximate string matching. *Softw Pract Exper*, 18:287–393, 1988.

[17] P. Jokinen and E. Ukonen. Two algorithms for approximate string matching in static texts. *Lect Notes in Comput Sci*, 520:240–248, 1991.

[18] David Weese, Anne-Katrin Emde, Tobias Rausch, Andreas Dring, and Knut Reinert. RazerS–fast read mapping with sensitivity control. *Genome Res*, 19(9):1646–1654, Sep 2009.

[19] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[20] Axel Visel, Matthew J Blow, Zirong Li, Tao Zhang, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, Veena Afzal, Bing Ren, Edward M Rubin, and Len A Pennacchio. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, Feb 2009.

[21] Gordon Robertson, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, Bridget Bernier, Richard Varhol, Allen Delaney, Nina Thiessen, Obi L Griffith, Ann He, Marco Marra, Michael Snyder, and Steven Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8):651–657, Aug 2007.

[22] Zhengdong D Zhang, Joel Rozowsky, Michael Snyder, Joseph Chang, and Mark Gerstein. Modeling ChIP sequencing in silico with applications. *PLoS Comput Biol*, 4(8):e1000158, 2008.

[23] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol*, 27(1):66–75, Jan 2009.

[24] Elizabeth G Wilbanks and Marc T Facciotti. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, 5(7):e11471, 2010.

[25] Yong Zhang, Tao Liu, Clifford A Meyer, Jerome Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nussbaum, Richard M Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based analysis of ChIP-seq (MACS). *Genome Biol*, 9(9):R137, 2008.

[26] Ghia M Euskirchen, Joel S Rozowsky, Chia-Lin Wei, Wah Heng Lee, Zhengdong D Zhang, Stephen Hartman, Olof Emanuelsson, Viktor Stolc, Sherman Weissman, Mark B Gerstein, Yijun Ruan, and Michael Snyder. Mapping of transcription factor binding regions in mammalian cells by chip: comparison of array- and sequencing-based technologies. *Genome Res*, 17(6):898–909, Jun 2007.

[27] Teng Fei, Kai Xia, Zhongwei Li, Bing Zhou, Shanshan Zhu, Hua Chen, Jianping Zhang, Zhang Chen, Huasheng Xiao, Jing-Dong J Han, and Ye-Guang Chen. Genome-wide mapping of SMAD target genes reveals the role of BMP signaling in embryonic stem cell fate determination. *Genome Res*, 20(1):36–44, Jan 2010.

[28] Deneen M Wellik and Mario R Capecchi. Hox10 and Hox11 genes are required to globally pattern the mammalian skeleton. *Science*, 301(5631):363–367, Jul 2003.

[29] Elizabeth D Wederell, Mikhail Bilenky, Rebecca Cullum, Nina Thiessen, Melis Dagpinar, Allen Delaney, Richard Varhol, YongJun Zhao, Thomas Zeng, Bridget Bernier, Matthew Ingham, Martin Hirst, Gordon Robertson, Marco A Marra, Steven Jones, and Pamela A Hoodless. Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res*, 36(14):4549–4564, Aug 2008.

[30] Ziv Bar-Joseph, Georg K Gerber, Tong Ihn Lee, Nicola J Rinaldi, Jane Y Yoo, Franois Robert, D. Benjamin Gordon, Ernest Fraenkel, Tommi S Jaakkola, Richard A Young, and David K Gifford. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, 21(11):1337–1342, Nov 2003.

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

[32] Adam Woolfe, Martin Goodson, Debbie K Goode, Phil Snell, Gayle K McEwen, Tanya Vavouri, Sarah F Smith, Phil North, Heather Callaway, and Krys Kelly *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, 3(1):e7, Jan 2005.

[33] Michael D Wilson, Nuno L Barbosa-Morais, Dominic Schmidt, Caitlin M Conboy, Lesley Vanes, Victor L J Tybulewicz, Elizabeth M C Fisher, Simon Tavar, and Duncan T Odom. Species-specific transcription in mice carrying human chromosome 21. *Science*, 322(5900):434–438, Oct 2008.

[34] Maya Kasowski, Fabian Grubert, Christopher Heffelfinger, Manoj Hariharan, Akwasi Asabere, Sebastian M Waszak, Lukas Habegger, Joel Rozowsky, Minyi Shi, Alexander E Urban, Mi-Young Hong, Konrad J Karczewski, Wolfgang Huber, Sherman M Weissman, Mark B Gerstein, Jan O Korbel, and Michael Snyder. Variation in transcription factor binding among humans. *Science*, 328(5975):232–235, Apr 2010.

[35] Thomas Manke, Helge G Roider, and Martin Vingron. Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS Comput Biol*, 4(3):e1000039, Mar 2008.

[36] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.

[37] G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17 Suppl 1:S207–S214, 2001.

[38] Laurence Ettwiller, Benedict Paten, Marcel Souren, Felix Loosli, Jochen Wittbrodt, and Ewan Birney. The discovery, positioning and verification of a set of transcription-associated motifs in vertebrates. *Genome Biol*, 6(12):R104, 2005.

[39] J. van Helden, M. del Olmo, and J. E. Prez-Ortn. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res*, 28(4):1000–1010, Feb 2000.

[40] Chaim Linhart, Yonit Halperin, and Ron Shamir. Transcription factor and microRNA motif discovery: the amadeus platform and a compendium of metazoan target sets. *Genome Res*, 18(7):1180–1189, Jul 2008.

[41] Ming Hu, Jindan Yu, Jeremy M G Taylor, Arul M Chinnaiyan, and Zhaohui S Qin. On the detection and refinement of transcription factor binding sites using ChIP-seq data. *Nucleic Acids Res*, 38(7):2154–2167, Apr 2010.

[42] Stoyan Georgiev, Alan P Boyle, Karthik Jayasurya, Xuan Ding, Sayan Mukherjee, and Uwe Ohler. Evidence-ranked motif identification. *Genome Biol*, 11(2):R19, 2010.

[43] A. Gordon Robertson, Mikhail Bilenky, Angela Tam, Yongjun Zhao, Thomas Zeng, Nina Thiessen, Timothee Cezard, Anthony P Fejes, Elizabeth D Wederell, Rebecca Cullum, Ghia Euskirchen, Martin Krzywinski, Inanc Birol, Michael Snyder, Pamela A Hoodless, Martin

Hirst, Marco A Marra, and Steven J M Jones. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res*, 18(12):1906–1917, Dec 2008.

[44] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–171, Feb 2001.

[45] Tong Ihn Lee, Sarah E Johnstone, and Richard A Young. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nat Protoc*, 1(2):729–748, 2006.

[46] Toshihisa Komori. Regulation of skeletal development by the Runx family of transcription factors. *J Cell Biochem*, 95(3):445–453, Jun 2005.

[47] Barbara D'haene, Catia Attanasio, Diane Beysen, Jose Dostie, Edmond Lemire, Philippe Bouchard, Michael Field, Kristie Jones, Birgit Lorenz, and Björn Menten *et al.* Disease-causing 7.4 kb cis-regulatory deletion disrupting conserved non-coding sequences and their interaction with the FOXL2 promotor: implications for mutation screening. *PLoS Genet*, 5(6):e1000522, Jun 2009.

[48] Paul G Giresi, Jonghwan Kim, Ryan M McDaniell, Vishwanath R Iyer, and Jason D Lieb. Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res*, 17(6):877–885, Jun 2007.

[49] Raymond K Auerbach, Ghia Euskirchen, Joel Rozowsky, Nathan Lamarre-Vincent, Zarmik Moqtaderi, Philippe Lefranois, Kevin Struhl, Mark Gerstein, and Michael Snyder. Mapping accessible chromatin regions using sono-seq. *Proc Natl Acad Sci U S A*, 106(35):14926–14931, Sep 2009.

[50] Jonathan A R Gordon, Mohammad Q Hassan, Sharanjot Saini, Martin Montecino, Andre J van Wijnen, Gary S Stein, Janet L Stein, and Jane B Lian. Pbx1 represses osteoblastogenesis by blocking Hoxa10-mediated recruitment of chromatin remodeling factors. *Mol Cell Biol*, 30(14):3531–3541, Jul 2010.

[51] L. Adamkova, K. Souckova, and J. Kovarik. Transcription protein STAT1: biology and relation to cancer. *Folia Biol (Praha)*, 53(1):1–6, 2007.

[52] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R. David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, Wei Wang, Zhiping Weng, Roland D Green, Gregory E Crawford, and Bing Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 39(3):311–318, Mar 2007.

[53] R. L. Johnson and C. J. Tabin. Molecular models for vertebrate limb development. *Cell*, 90(6):979–990, Sep 1997.

[54] Frances R Goodman. Limb malformations and the human hox genes. *Am J Med Genet*, 112(3):256–265, Oct 2002.

[55] Xiuli Zhao, Miao Sun, Jin Zhao, J. Alfonso Leyva, Hongwen Zhu, Wei Yang, Xuan Zeng, Yang Ao, Qing Liu, Guoyang Liu, Wilson H Y Lo, Ethylin Wang Jabs, L. Mario Amzel, Xiangnian Shan, and Xue Zhang. Mutations in HOXD13 underlie syndactyly type V and a novel brachydactyly-syndactyly syndrome. *Am J Hum Genet*, 80(2):361–371, Feb 2007.

[56] W. F. Shen, J. C. Montgomery, S. Rozenfeld, J. J. Moskow, H. J. Lawrence, A. M. Buchberg, and C. Largman. AbdB-like Hox proteins stabilize DNA binding by the Meis1 homeodomain proteins. *Mol Cell Biol*, 17(11):6448–6458, Nov 1997.

[57] Cecilia B Moens and Licia Selleri. Hox cofactors in vertebrate development. *Dev Biol*, 291(2):193–206, Mar 2006.

[58] Frank Plger, Petra Seemann, Mareen Schmidt von Kegler, Katarina Lehmann, Jrg Seidel, Klaus W Kjaer, Jens Pohl, and Stefan Mundlos. Brachydactyly type a2 associated with a defect in progdf5 processing. *Hum Mol Genet*, 17(9):1222–1233, May 2008.

[59] Kimberly Robasky and Martha L Bulyk. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res*, 39(Database issue):D124–D128, Jan 2011.

[60] Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. Weblogo: a sequence logo generator. *Genome Res*, 14(6):1188–1190, Jun 2004.

[61] Olivier Pourqié. *Hox Genes*. Academic Press Inc., 2009.

[62] Thomas M Williams, Melissa E Williams, Joanne H Heaton, Thomas D Gelehrter, and Jeffrey W Innis. Group 13 HOX proteins interact with the MH2 domain of R-Smads and modulate smad transcriptional activation functions independent of hox dna-binding capability. *Nucleic Acids Res*, 33(14):4475–4484, 2005.

[63] S. Audic and J. M. Claverie. The significance of digital gene expression profiles. *Genome Res*, 7(10):986–995, Oct 1997.

[64] Timothy Ravasi, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, Nicolas Bertin, Piero Carninci, Carsten O Daub, Alistair R R Forrest, Julian Gough, Sean Grimmond, Jung-Hoon Han, Takehiro Hashimoto, Winston Hide, Oliver Hofmann, Atanas Kamburov, Mandeep Kaur, Hideya Kawaji, Atsutaka Kubosaki, Timo Lassmann, Erik van Nimwegen, Cameron Ross MacPherson, Chihiro Ogawa, Aleksandar Radovanovic, Ariel Schwartz, Rohan D Teasdale, Jesper Tegnr, Boris Lenhard, Sarah A Teichmann, Takahiro Arakawa, Noriko Ninomiya, Kayoko Murakami, Michihira Tagami, Shiro Fukuda, Kengo Imamura, Chikatoshi Kai, Ryoko Ishihara, Yayoi Kitazume, Jun Kawai, David A Hume, Trey Ideker, and Yoshihide Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, Mar 2010.

[65] Kai Wang, Masumichi Saito, Brygida C Bisikirska, Mariano J Alvarez, Wei Keat Lim, Presha Rajbhandari, Qiong Shen, Ilya Nemenman, Katia Basso, Adam A Margolin, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nat Biotechnol*, 27(9):829–839, Sep 2009.

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

[67] Xi Chen, Han Xu, Ping Yuan, Fang Fang, Mikael Huss, Vinsensius B Vega, Eleanor Wong, Yuriy L Orlov, Weiwei Zhang, Jianming Jiang, Yuin-Han Loh, Hock Chuan Yeo, Zhen Xuan Yeo, Vipin Narang, Kunde Ramamoorthy Govindarajan, Bernard Leong, Atif Shahab, Yijun Ruan, Guillaume Bourque, Wing-Kin Sung, Neil D Clarke, Chia-Lin Wei, and Huck-Hui Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, Jun 2008.

[68] F. Otto, A. P. Thornell, T. Crompton, A. Denzel, K. C. Gilmour, I. R. Rosewell, G. W. Stamp, R. S. Beddington, S. Mundlos, B. R. Olsen, P. B. Selby, and M. J. Owen. Cbfa1, a candidate gene for cleidocranial dysplasia syndrome, is essential for osteoblast differentiation and bone development. *Cell*, 89(5):765–771, May 1997.

[69] Sigmar Stricker, Reinald Fundele, Andrea Vortkamp, and Stefan Mundlos. Role of Runx genes in chondrocyte differentiation. *Dev Biol*, 245(1):95–108, May 2002.

[70] V. Geoffroy, P. Ducy, and G. Karsenty. A PEBP2 alpha/AML-1-related factor increases osteocalcin promoter activity through its binding to an osteoblast-specific cis-acting element. *J Biol Chem*, 270(52):30973–30979, Dec 1995.

[71] Qi Shen and Sylvia Christakos. The vitamin D receptor, Runx2, and the Notch signaling pathway cooperate in the transcriptional regulation of osteopontin. *J Biol Chem*, 280(49):40589–40598, Dec 2005.

[72] K. Shirakabe, K. Terasawa, K. Miyama, H. Shibuya, and E. Nishida. Regulation of the activity of the transcription factor Runx2 by two homeobox proteins, Msx2 and Dlx5. *Genes Cells*, 6(10):851–856, Oct 2001.

[73] Mi-Hye Lee, Youn-Jeong Kim, Won-Joson Yoon, Jee-In Kim, Byung-Gyu Kim, Yoo-Seok Hwang, John M Wozney, Xin-Zi Chi, Suk-Chul Bae, Kang-Young Choi, Je-Yoel Cho, Je-Yong Choi, and Hyun-Mo Ryoo. Dlx5 specifically regulates Runx2 type II expression by binding to homeodomain-response elements in the Runx2 distal promoter. *J Biol Chem*, 280(42):35579–35587, Oct 2005.

[74] Nicolas Holleville, Stéephanie Mat'eos, Martine Bontoux, Karine Bollerot, and Anne-H'elène Monsoro-Burq. Dlx5 drives Runx2 expression and osteogenic differentiation in developing cranial suture mesenchyme. *Dev Biol*, 304(2):860–874, Apr 2007.

[75] Uri Alon. Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8(6):450–461, Jun 2007.

[76] Peter Bialek, Britt Kern, Xiangli Yang, Marijke Schrock, Drazen Sosic, Nancy Hong, Hua Wu, Kai Yu, David M Ornitz, Eric N Olson, Monica J Justice, and Gerard Karsenty. A Twist code determines the onset of osteoblast differentiation. *Dev Cell*, 6(3):423–435, Mar 2004.

[77] Hernan Roca, Mattabhorn Phimphilai, Rajaram Gopalakrishnan, Guozhi Xiao, and Renny T Franceschi. Cooperative interactions between RUNX2 and homeodomain protein-binding sites

are critical for the osteoblast-specific expression of the bone sialoprotein gene. *J Biol Chem*, 280(35):30845–30855, Sep 2005.

[78] Mohammad Q Hassan, Rahul Tare, Suk Hee Lee, Matthew Mandeville, Brian Weiner, Martin Montecino, Andre J van Wijnen, Janet L Stein, Gary S Stein, and Jane B Lian. Hoxa10 controls osteoblastogenesis by directly activating bone regulatory and phenotypic genes. *Mol Cell Biol*, 27(9):3337–3352, May 2007.

[79] Pablo Villavicencio-Lorini, Pia Kuss, Julia Friedrich, Julia Haupt, Muhammed Farooq, Seval Trkmen, Denis Duboule, Jochen Hecht, and Stefan Mundlos. Homeobox genes d11-d13 and a13 control mouse autopod cortical bone and joint formation. *J Clin Invest*, 120(6):1994–2004, Jun 2010.

[80] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstrle, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, Todd R Golub, Pablo Tamayo, Bruce Spiegelman, Eric S Lander, Joel N Hirschhorn, David Altshuler, and Leif C Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–273, Jul 2003.

[81] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005.

[82] Leila Taher and Ivan Ovcharenko. Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics*, 25(5):578–584, Mar 2009.

[83] Sebastian Bauer, Steffen Grossmann, Martin Vingron, and Peter N Robinson. Ontologizer 2.0–a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651, Jul 2008.

[84] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

[85] Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nat Biotechnol*, 28(5):495–501, May 2010.

[86] Ivan Ovcharenko, Gabriela G Loots, Marcelo A Nobrega, Ross C Hardison, Webb Miller, and Lisa Stubbs. Evolution and functional classification of vertebrate gene deserts. *Genome Res*, 15(1):137–145, Jan 2005.

[87] D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, and F. Hsu *et al.* The UCSC genome browser database: 2008 update. *Nucleic Acids Res*, 36(Database issue):D773–D779, Jan 2008.

[88] Nadiya M Teplyuk, Ying Zhang, Yang Lou, John R Hawse, Mohammad Q Hassan, Viktor I Teplyuk, Jitesh Pratap, Mario Galindo, Janet L Stein, Gary S Stein, Jane B Lian, and Andre J van Wijnen. The osteogenic transcription factor Runx2 controls genes involved in sterol/steroid metabolism, including CYP11A1 in osteoblasts. *Mol Endocrinol*, 23(6):849–861, Jun 2009.

[89] Axel Visel, Simon Minovitsky, Inna Dubchak, and Len A Pennacchio. VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Res*, 35(Database issue):D88–D92, Jan 2007.

[90] Len A Pennacchio, Nadav Ahituv, Alan M Moses, Shyam Prabhakar, Marcelo A Nobrega, Malak Shoukry, Simon Minovitsky, Inna Dubchak, Amy Holt, and Keith D Lewis *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, Nov 2006.

[91] Len A Pennacchio, Gabriela G Loots, Marcelo A Nobrega, and Ivan Ovcharenko. Predicting tissue-specific enhancers in the human genome. *Genome Res*, 17(2):201–211, Feb 2007.

[92] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, Jun 2007.

[93] M. Merika, A. J. Williams, G. Chen, T. Collins, and D. Thanos. Recruitment of CBP/P300 by the IFN beta enhanceosome is required for synergistic activation of transcription. *Mol Cell*, 1(2):277–287, Jan 1998.

[94] Leelavati Narlikar, Noboru J Sakabe, Alexander A Blanski, Fabio E Arimura, John M Westlund, Marcelo A Nobrega, and Ivan Ovcharenko. Genome-wide discovery of human heart enhancers. *Genome Res*, 20(3):381–392, Mar 2010.

[95] Jason B Warner, Anthony A Philippakis, Savina A Jaeger, Fangxue Sherry He, Jolinta Lin, and Martha L Bulyk. Systematic identification of mammalian regulatory motifs' target genes and functions. *Nat Methods*, 5(4):347–353, Apr 2008.

[96] Yi Cao, Zizhen Yao, Deepayan Sarkar, Michael Lawrence, Gilson J Sanchez, Maura H Parker, Kyle L MacQuarrie, Jerry Davison, Martin T Morgan, Walter L Ruzzo, Robert C Gentleman, and Stephen J Tapscott. Genome-wide myod binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev Cell*, 18(4):662–674, Apr 2010.

[97] Dirk A Kleinjan, Anne Seawright, Greg Elgar, and Veronica van Heyningen. Characterization of a novel gene adjacent to Pax6, revealing synteny conservation with functional significance. *Mamm Genome*, 13(2):102–107, Feb 2002.

[98] Dirk A Kleinjan, Anne Seawright, Sebastien Mella, Catherine B Carr, David A Tyas, T. Ian Simpson, John O Mason, David J Price, and Veronica van Heyningen. Long-range downstream enhancers are essential for Pax6 expression. *Dev Biol*, 299(2):563–581, Nov 2006.

[99] Laura A Lettice, Simon J H Heaney, Lorna A Purdie, Li Li, Philippe de Beer, Ben A Oostra, Debbie Goode, Greg Elgar, Robert E Hill, and Esther de Graaff. A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*, 12(14):1725–1735, Jul 2003.

[100] Nadav Ahituv, Shyam Prabhakar, Francis Poulin, Edward M Rubin, and Olivier Couronne. Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. *Hum Mol Genet*, 14(20):3057–3063, Oct 2005.

[101] Hiroshi Kikuta, Mary Laplante, Pavla Navratilova, Anna Z Komisarczuk, Pär G Engström, David Fredman, Altuna Akalin, Mario Caccamo, Ian Sealy, and Kerstin Howe *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res*, 17(5):545–555, May 2007.

[102] Xiaokang Pan, Lincoln Stein, and Volker Brendel. Synbrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, 21(17):3461–3468, Sep 2005.

[103] Martin C Frith, Michael C Li, and Zhiping Weng. Cluster-buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*, 31(13):3666–3668, Jul 2003.

[104] Saurabh Sinha, Yupu Liang, and Eric Siggia. Stubb: a program for discovery and analysis of cis-regulatory modules. *Nucleic Acids Res*, 34(Web Server issue):W555–W559, Jul 2006.

[105] F. Müller, B. Chang, S. Albert, N. Fischer, L. Tora, and U. Strähle. Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord. *Development*, 126(10):2103–2116, May 1999.

[106] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, Feb 2002.

[107] Altuna Akalin, David Fredman, Erik Arner, Xianjun Dong, Jan Bryne, Harukazu Suzuki, Carsten Daub, and Yoshihide Hayashizaki *et al.* Transcriptional features of genomic regulatory blocks. *Genome Biol*, 10(4):R38, Apr 2009.

[108] Steven A Vokes, Hongkai Ji, Wing H Wong, and Andrew P McMahon. A genome-scale analysis of the cis-regulatory circuitry underlying sonic hedgehog-mediated patterning of the mammalian limb. *Genes Dev*, 22(19):2651–2663, Oct 2008.

[109] Scott Schwartz, W. James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C Hardison, David Haussler, and Webb Miller. Human-mouse alignments with blastz. *Genome Res*, 13(1):103–107, Jan 2003.

[110] Webb Miller, Kate Rosenbloom, Ross C Hardison, Minmei Hou, James Taylor, Brian Raney, Richard Burhans, David C King, Robert Baertsch, and Daniel Blankenberg *et al.* 28-way vertebrate alignment and conservation track in the UCSC genome browser. *Genome Res*, 17(12):1797–1808, Dec 2007.

[111] Dayana Krawchuk and Artur Kania. Identification of genes controlled by LMX1B in the developing mouse limb bud. *Dev Dyn*, 237(4):1183–1192, Apr 2008.

[112] Daniel H Huson, Daniel C Richter, Christian Rausch, Tobias Dezulian, Markus Franz, and Regula Rupp. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, 8:460, 2007.

[113] Christian Rödelsperger, Sebastian Köhler, Marcel H Schulz, Thomas Manke, Sebastian Bauer, and Peter N Robinson. Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts. *Genomics*, 94(5):308–316, Nov 2009.

[114] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[115] Sebastian Köhler, Marcel H Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*, 85(4):457–64, Oct 2009.

[116] Lars Juhl Jensen, M Kuhn, M Stark, S Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, Peer Bork, and Christian von Mering. String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucl. Acids Res.*, Jan 2009.

[117] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*, 82(4):949–58, Apr 2008.

[118] Pierre Geurts, Alexandre Irrthum, and Louis Wehenkel. Supervised learning with decision tree-based methods in computational and systems biology. *Mol Biosyst*, 5(12):1593–1605, Dec 2009.

[119] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

[120] Albin Sandelin, Peter Bailey, Sara Bruce, Pär G Engström, Joanna M Klos, Wyeth W Wasserman, Johan Ericson, and Boris Lenhard. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5(1):99, 2004.

[121] Gill Bejerano, Craig B Lowe, Nadav Ahituv, Bryan King, Adam Siepel, Sofie R Salama, Edward M Rubin, W. James Kent, and David Haussler. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, 441(7089):87–90, May 2006.

[122] B. Kammandel, K. Chowdhury, A. Stoykova, S. Aparicio, S. Brenner, and P. Gruss. Distinct cis-essential modules direct the time-space pattern of the pax6 gene activity. *Dev Biol*, 205(1):79–97, Jan 1999.

[123] Dirk A Kleinjan, Anne Seawright, Andrew J Childs, and Veronica van Heyningen. Conserved elements in Pax6 intron 7 involved in (auto)regulation and alternative transcription. *Dev Biol*, 265(2):462–477, Jan 2004.

[124] Masanori Uchikawa, Yoshiko Ishida, Tatsuya Takemoto, Yusuke Kamachi, and Hisato Kondoh. Functional analysis of chicken sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev Cell*, 4(4):509–519, Apr 2003.

[125] Daisuke Kurokawa, Hiroshi Kiyonari, Rika Nakayama, Chiharu Kimura-Yoshida, Isao Matsuo, and Shinichi Aizawa. Regulation of Otx2 expression and its functions in mouse forebrain and midbrain. *Development*, 131(14):3319–3331, Jul 2004.

[126] Chiharu Kimura-Yoshida, Kuniko Kitajima, Izumi Oda-Ishii, E. Tian, Misao Suzuki, Masayuki Yamamoto, Tohru Suzuki, Makoto Kobayashi, Shinichi Aizawa, and Isao Matsuo. Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development*, 131(1):57–71, Jan 2004.

[127] Sabina Benko, Judy A Fantes, Jeanne Amiel, Dirk-Jan Kleinjan, Sophie Thomas, Jacqueline Ramsay, Negar Jamshidi, Abdelkader Essafi, Simon Heaney, and Christopher T Gordon *et al.* Highly conserved non-coding elements on either side of SOX9 associated with pierre robin sequence. *Nat Genet*, 41(3):359–364, Mar 2009.

[128] Fedik Rahimov, Mary L Marazita, Axel Visel, Margaret E Cooper, Michael J Hitchler, Michele Rubini, Frederick E Domann, Manika Govil, Kaare Christensen, Camille Bille, Mads Melbye, Astanand Jugessur, Rolv T Lie, Allen J Wilcox, David R Fitzpatrick, Eric D Green, Peter A Mossey, Julian Little, Regine P Steegers-Theunissen, Len A Pennacchio, Brian C Schutte, and Jeffrey C Murray. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet*, 40(11):1341–1347, Nov 2008.

[129] Kerry Ann Miller, Scott Davidson, Angela Liaros, John Barrow, Marissa Lear, Danielle Heine, Stefan Hoppler, and Alasdair MacKenzie. Prediction and characterisation of a highly conserved, remote and camp responsive enhancer that regulates msx1 gene expression in cardiac neural crest and outflow tract. *Dev Biol*, 317(2):686–694, May 2008.

[130] Vronique Bourdeau, Julie Deschnes, Raphal Mtivier, Yoshihiko Nagai, Denis Nguyen, Nancy Bretschneider, Frank Gannon, John H White, and Sylvie Mader. Genome-wide identification of high-affinity estrogen response elements in human and mouse. *Mol Endocrinol*, 18(6):1411–1427, Jun 2004.

[131] Osamu Uemura, Yohei Okada, Hideki Ando, Mickael Guedj, Shin-Ichi Higashijima, Takuya Shimazaki, Naoichi Chino, Hideyuki Okano, and Hitoshi Okamoto. Comparative functional genomics revealed conservation and diversification of three enhancers of the isl1 gene for motor and sensory neuron-specific expression. *Dev Biol*, 278(2):587–606, Feb 2005.

[132] Jeffrey A Magee, Li wei Chang, Gary D Stormo, and Jeffrey Milbrandt. Direct, androgen receptor-mediated regulation of the fkbp5 gene via a distal enhancer element. *Endocrinology*, 147(1):590–598, Jan 2006.

[133] Sarah De Val, Joshua P Anderson, Analeah B Heidt, Dustin Khiem, Shan-Mei Xu, and Brian L Black. Mef2c is activated directly by ets transcription factors through an evolutionarily conserved endothelial cell-specific enhancer. *Dev Biol*, 275(2):424–434, Nov 2004.

[134] Kenta Sumiyama and Frank H Ruddle. Regulation of dlx3 gene expression in visceral arches by evolutionarily conserved enhancer elements. *Proc Natl Acad Sci U S A*, 100(7):4030–4034, Apr 2003.

[135] R. Forghani, L. Garofalo, D. R. Foran, H. F. Farhadi, P. Lepage, T. J. Hudson, I. Tretjakoff, P. Valera, and A. Peterson. A distal upstream enhancer from the myelin basic protein gene regulates expression in myelin-forming schwann cells. *J Neurosci*, 21(11):3780–3787, Jun 2001.

[136] W. James Kent. Blat–the blast-like alignment tool. *Genome Res*, 12(4):656–664, Apr 2002.

[137] R. M. Kuhn, D. Karolchik, A. S. Zweig, T. Wang, K. E. Smith, K. R. Rosenbloom, B. Rhead, B. J. Raney, A. Pohl, M. Pheasant, L. Meyer, F. Hsu, A. S. Hinrichs, R. A. Harte, B. Giardine, P. Fujita, M. Diekhans, T. Dreszer, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent. The UCSC genome browser database: update 2009. *Nucleic Acids Res*, 37(Database issue):D755–D761, Jan 2009.

[138] Asish K Ghosh and John Varga. The transcriptional coactivator and acetyltransferase p300 in fibroblast biology and fibrosis. *J Cell Physiol*, 213(3):663–671, Dec 2007.

[139] Noriko Shikama, Werner Lutz, Ralph Kretzschmar, Nadine Sauter, Jeanne-Franoise Roth, Silvia Marino, Jonas Wittwer, Alexander Scheidweiler, and Richard Eckner. Essential function of p300 acetyltransferase activity in heart, lung and small intestine formation. *EMBO J*, 22(19):5175–5185, Oct 2003.

[140] Takayuki Furumatsu, Masanao Tsuda, Kenji Yoshida, Noboru Taniguchi, Tatsuo Ito, Megumi Hashimoto, Takashi Ito, and Hiroshi Asahara. Sox9 and p300 cooperatively regulate chromatin-mediated transcription. *J Biol Chem*, 280(42):35203–35208, Oct 2005.

[141] Gabriele Schweikert, Alexander Zien, Georg Zeller, Jonas Behr, Christoph Dieterich, Cheng Soon Ong, Petra Philips, Fabio De Bona, Lisa Hartmann, Anja Bohlen, Nina Krger, Sören Sonnenburg, and Gunnar Rätsch. mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Res*, 19(11):2133–2143, Nov 2009.

[142] Derek Adams, Michele Karolak, Elizabeth Robertson, and Leif Oxburgh. Control of kidney, eye and limb expression of Bmp7 by an enhancer element highly conserved between species. *Dev Biol*, 311(2):679–690, Nov 2007.

[143] Gill Bejerano, Michael Pheasant, Igor Makunin, Stuart Stephen, W. James Kent, John S Mattick, and David Haussler. Ultraconserved elements in the human genome. *Science*, 304(5675):1321–1325, May 2004.

[144] Szymon M Kielbasa and Martin Vingron. Transcriptional autoregulatory loops are highly conserved in vertebrate evolution. *PLoS One*, 3(9):e3210, 2008.

[145] Yong-Bae Kim, Sung-Yul Lee, Sang-Kyu Ye, and Jung Weon Lee. Epigenetic regulation of integrin-linked kinase expression depending on adhesion of gastric carcinoma cells. *Am J Physiol Cell Physiol*, 292(2):C857–C866, Feb 2007.

[146] J. D. Engel and K. Tanimoto. Looping, linking, and chromatin activity: new insights into beta-globin locus regulation. *Cell*, 100(5):499–502, Mar 2000.

[147] Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, and Poh Huay Mei *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, Nov 2009.

[148] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, and Michael O Dorschner *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct 2009.

[149] S. K. Zaidi, A. Javed, J. Y. Choi, A. J. van Wijnen, J. L. Stein, J. B. Lian, and G. S. Stein. A specific targeting signal directs runx2/cbfa1 to subnuclear domains and contributes to transactivation of the osteocalcin gene. *J Cell Sci*, 114(Pt 17):3093–3102, Sep 2001.

[150] 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 Res*, 34(Database issue):D108–D110, Jan 2006.

[151] Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, 36(Database issue):D102–D106, Jan 2008.

[152] Tae Hoon Kim, Ziedulla K Abdullaev, Andrew D Smith, Keith A Ching, Dmitri I Loukinov, Roland D Green, Michael Q Zhang, Victor V Lobanenkov, and Bing Ren. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 128(6):1231–1245, Mar 2007.

[153] A. C. Bell, A. G. West, and G. Felsenfeld. The protein *ctcf* is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3):387–396, Aug 1999.

[154] J. Jiang and M. Levine. Binding affinities and cooperative interactions with bhlh activators delimit threshold responses to the dorsal gradient morphogen. *Cell*, 72(5):741–752, Mar 1993.

[155] Sari Tuupanen, Mikko Turunen, Rainer Lehtonen, Outi Hallikas, Sakari Vanharanta, Teemu Kivioja, Mikael Bjrklund, Gonghong Wei, Jian Yan, Iina Niittymki, Jukka-Pekka Mecklin, Heikki Jrvinen, Ari Ristimki, Mariachiara Di-Bernardo, Phil East, Luis Carvajal-Carmona, Richard S Houlston, Ian Tomlinson, Kimmo Palin, Esko Ukkonen, Auli Karhu, Jussi Taipale, and Lauri A Aaltonen. The common colorectal cancer predisposition snp rs6983267 at chromosome 8q24 confers potential to enhanced wnt signaling. *Nat Genet*, 41(8):885–890, Aug 2009.

[156] Dominic Schmidt, Michael D Wilson, Benoit Ballester, Petra C Schwalie, Gordon D Brown, Aileen Marshall, Claudia Kutter, Stephen Watt, Celia P Martinez-Jimenez, Sarah Mackay, Iannis Talianidis, Paul Flicek, and Duncan T Odom. Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–1040, May 2010.

List of Figures

List of Tables

A Zusammenfassung

Grundlegende biologische Prozesse wie Wachstum und Differenzierung werden durch die koordinierte Regulation von Genen durch Transkriptionsfaktoren gesteuert. Genomweite experimentelle Ansätze zur Quantifizierung von Genexpression mittels Microarrays haben unser Wissen über genregulatorische Netzwerke und deren Dynamik über verschiedene Entwicklungsstadien und Gewebe hinweg substantiell erweitert. Die Technik, über spezifische Antikörper, von einem Transkriptionsfaktor gebundene DNA, bzw. Chromatin zu immunoprezipitieren und dann die angereicherte DNA zu sequenzieren (ChIP-seq), hat die Möglichkeit geschaffen, zu einem bestimmten Zeitpunkt nahezu alle genomischen Regionen, die von einem Transkriptionsfaktoren gebunden sind, zu detektieren und mit diesem Wissen viel tiefere Einblicke in die Mechanismen der Genregulation zu gewinnen. Die Interpretation der gewonnenen Daten gestaltet sich jedoch schwierig, weil erstens Transkriptionsfaktoren genomische Regionen, sogenannte *cis*-regulatorische Bereiche, binden können, die hunderte von Kilobasen von einem Gen entfernt liegen und dessen Expression beeinflussen und weil zweitens nicht jedes Bindungsereignis die Expression eines Gens beeinflusst.

In dieser Arbeit werden Methoden entwickelt, um genomweite Bindungsprofile besser zu charakterisieren und zu vergleichen. Darüber hinaus beschreiben wir einen Algorithmus, der die Bindungsprofile von mehreren Transkriptionsfaktoren integriert und Klassen von kombinatorisch gebundenen Regionen definiert und deren Funktionalität über ihre Assoziation mit differentieller Expression von benachbarten Genen bestimmt. Die Methode lässt sich dazu verwenden, aus den tausenden von gebunden Regionen, Klassen von einigen hunderten zu definieren, die mit höherer Wahrscheinlichkeit eine regulatorische Rolle spielen. Dies wird durch Analysen auf funktionelle Kohärenz und speziesübergreifende Sequenzkonservierung bestätigt.

Im letzten Abschnitt dieser Arbeit stellen wir eine Methode vor, die ausgehend von funktionellen *cis*-regulatorischen Bereichen die Vorhersage der Zielgene verbessert, indem sie nicht nur die genomische Distanz berücksichtigt sondern zusätzlich die Konservierung der Syntenie zwischen *cis*-regulatorischer Region und Zielgen, die funktionelle Ähnlichkeit zwischen dem immunoprezipitierten Transkriptionsfaktor und Zielgen und deren Nähe in Proteininteraktionsnetzwerken berücksichtigt und damit die Korrektheit der Vorhersagen für Zielgene im Vergleich zu Vorhersagen, die nur auf genomischer Distanz basieren, um das Zweifache auf $\approx 58\%$ verbessert.

Die vorgestellten Methoden ermöglichen es somit, mehr biologisch relevante Informationen aus den ChIP-seq Daten zu ziehen und damit die Wirkungsweise der untersuchten Transkriptionsfaktoren besser zu verstehen.

B Summary

Fundamental biological processes such as differentiation and proliferation depend on the coordinated regulation of genes by transcription factors. Genome-wide experimental approaches for quantification of gene expression have substantially extended our knowledge about gene-regulatory networks and their dynamics across developmental stages and tissues. The technique of using specific antibodies in order to enrich DNA that is bound by a transcription factor and to subsequently sequence the immunoprecipitated DNA (ChIP-seq) has facilitated the genome-wide mapping of protein-DNA interactions. The generated data can be used to gain deeper insights into the mechanisms of gene regulation. However, the interpretation of the data is complicated by the fact that transcription factors bind genomic regions, so called *cis*-regulatory modules, that may regulate the expression of a target gene that is located several hundred kilobases away. Furthermore not every binding event shows a direct effect on the expression of a gene.

In this work, we develop methods for characterization and comparison of genome-wide binding profiles. In addition we describe an algorithm which integrates binding profiles for multiple transcription factors and defines classes of combinatorial binding events in order to assess their functional impact on differential expression of neighboring genes. This method can be used to filter the thousands of binding events for classes of a few hundred events that are more likely to have a regulatory function. This is confirmed by analysis of functional coherence and cross-species sequence conservation.

In the last part of this thesis, we present a method, that improves the prediction of target genes for a set of functional *cis*-regulatory regions by not only relying on genomic distance but also integrating information about conserved synteny between *cis*-regulatory region and target gene, functional similarity between regulator and target gene, and vicinity in protein-interaction networks. Our method predicts the correct target genes in $\approx 58\%$ of cases, which is a two fold improvement over an approach that only relies on genomic distance.

In summary, the presented methods allow to gain more biologically relevant insights from the analysis of ChIP-seq data and to improve our understanding of the function of the analyzed transcription factors.

C Ehrenwörtliche Erklärung

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

Berlin, Dezember 2011 Christian Rödelsperger