5 DISCUSSION

In the present study the effects of histone modifications and their interactions with transcription factors were investigated and placed into the context of heart development and function. The studies were carried out in cardiomyocytes (HL-1 cells) and compared with differentiated as well as undifferentiated skeletal muscle cells (C2C12 cells). First of all, the occurrence of four different histone modifications and their combinatorial effects on transcription were elucidated using ChIP-chip experiments and expression arrays. In addition, these findings were extended through the analysis of four different key transcription factors known to be essential for heart development employing ChIP-chip, RNA silencing, and expression arrays. Finally, the interplay between histone modifications and transcription factor binding was investigated and the results were compiled into regulatory networks.

5.1 Development of Open Source Tools for ChIP-chip

5.1.1 The Design of the Muscle Regulatory chips for ChIP

The investigations of histone modifications and transcription factor binding on a transcriptome-wide level were made possible by the design of custom ChIP-chip arrays. Although murine genome-wide expression arrays have been available for several years[316] , ChIP-chip arrays covering the entire genome of a higher eukaryote are currently unavailable. In this study a set of ChIP-chip arrays has been tailored to specifically meet the requirements of research on any muscle tissue, including heart and skeletal muscle. The presented array represents 8,585 genes out of the ≈ 27,000 encoded in the mouse genome (mouse genome build NCBI mm8 v36).

In the first study a probe tiling in a nucleosomal range of approximately 150 bp was used. This was sufficient to analyze histone modifications because 147 bp of DNA are wrapped around the histone core in one nucleosome. The binding sites of transcription factors are, however, typically only 10-20 bp long. Therefore, a smaller spacing was preferable for the investigation of transcription factors and a second array version was designed. However, twice the number of probes was required thus approximately doubling the cost of the arrays. The design is accessible through a web site. Whether particular transcripts are contained on the array and which genomic regions are represented, as well as the number of probes for a given region can be queried there. After publication the site will be made publicly available through www.molgen.mpg.de/~heart/.

5.1.2 Data Analysis – *Ringo*

In order to asses the quality of the array hybridizations and to analyze the raw data obtained from the ChIP-chip experiments a number of algorithms were designed. From these the freely available, open-source software module *Ringo* was developed[257]. *Ringo* can be used for the import of the raw microarray data as supplied by NimbleGen, their quality assessment, normalization, visualization, and for the detection and quantitation of ChIP-enriched regions.

Software for the analysis of ChIP-chip existed previously (For example, mpeak[317], TiMAT http://bdtnp.lbl.gov/TiMAT, MAT[318], TileMap[319], ACME[320], HGMM[321], and ChIPOTle[322]). However, these programs can only be used to find ChIP-enriched regions on ChIP-chip data that have been already normalized and quality controlled. *Ringo* is complementary to this existing software for ChIP microarray analysis, because its functionality covers the complete primary analysis for ChIP-chip tiling microarrays, especially those from the company NimbleGen.

Because *Ringo* is integrated with the Bioconductor project[252] it can be easily combined with other software or packages. This determines a unique aspect of *Ringo*: It facilitates the construction of more automated programmed workflows and offers benefits in the scalability, reproducibility and methodical scope of the analyses.

The R-package *Ringo* is available from the Bioconductor web site at http://www.bioconductor.org and runs on Linux, Mac OS and MS-Windows. The easiest way to obtain the most recent version of the software, with all its dependencies, is to follow the instructions at http://www.bioconductor.org/download.

5.1.3 Integrated Database of Results

In this study the global effect of four histone modifications and of four cardiac transcription factors on expression were analyzed. Through this analysis 8,585 genes were characterized regarding their histone modification patterns and binding of Gata4, Mef2a, Nkx2.5, and Srf. The raw array data were deposited at the public data base *ArrayExpress* [http://www.ebi.ac.uk]. The raw data from the histone project can already be freely downloaded, the data from the transcription factor project will be made available after publication. However, researchers without the opportunity of high-throughput raw data analysis may still be interested to assess the results of this analysis for their particular gene of interest. Therefore, a database was developed where a gene of interest may be queried by MGI symbol or ENSEMBL ID.

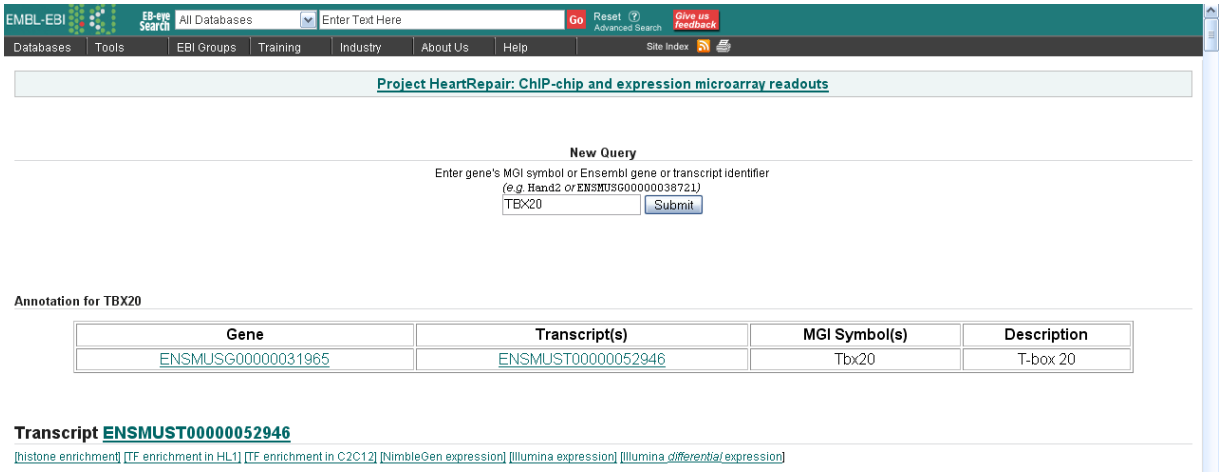

Figure 5-1. Query page for ChIP-chip and expression array readouts. The results of histone modification and transcription factor ChIP arrays as well as expression arrays are visualized. Here, Tbx20 is given as an example.

If the gene was one of the 8,585 analyzed genes the following information is graphically displayed: the expression levels and histone modification pattern in myoblasts, myotubes and HL-1 cells, as well as if binding sites of Gata4, Mef2a, Nkx2.5, and Srf were identified. In case of several TSSs the information for each is displayed separately. The data are processed into an intuitive graphic presentation facilitating the access by a broad (non-bioinformatic) community (Figure 5-2). However, it must be noted, that as for all high-throughput data the results for single genes can only serve as a starting point and must be verified. Since the transcription factor data are not yet published, currently the web address and password are only available upon request.

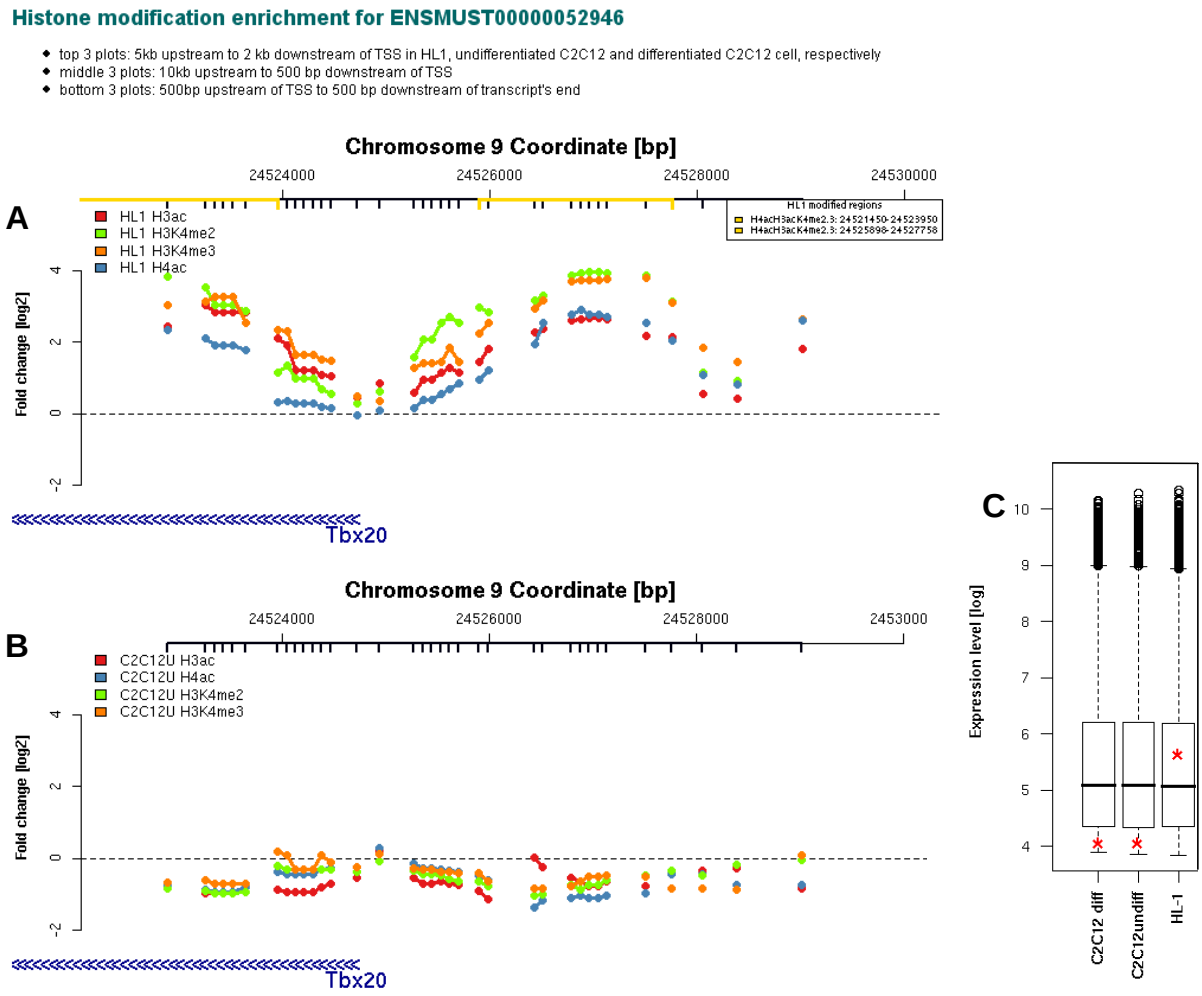

Figure 5-2. Three out of 19 plots that are shown on the web-page querying for *Tbx20*. The graphs show histone modification pattern near the TSS of *Tbx20* in A) HL-1 cells, B) C2C12 undiff. cells, and C) the expression level of *Tbx20* (red crosses) relative to the transcriptome of the cell lines is shown as box plots.

5.2 Regulation of Transcription in Heart and Skeletal Muscle Cells by Histone Modifications

Nucleosomes are more than a static entity providing the scaffold for the compaction of DNA and efforts have been made to elucidate their further functions. Since the formulation of the *histone code hypothesis* in 2000 by Strahl and Allis[78] it has been an open point of discussion whether histone modifications act synergistically to form a *histone code* and to which extent they function as signaling marks for the recruitment of effectors. The existence of a *histone code* is expected to manifest itself by distinguishable effects of different combinations of histone modifications. Furthermore, modifications might influence each other in such a way that their single effects are non-additive. In this context it is furthermore of interest whether histone modification marks are primary or secondary to transcription. If histone marks were placed before the transcription process took place this would indicate a major function as a signaling mark. If, however, these modifications were formed only in the

course of transcription a high interdependence of modification change and expression change would be the expected consequence.

5.2.1 Modified Sites and Literature Comparison

Analysis of the ChIP-chip experiments led to the identification of genomic locations where histones carried the investigated modifications. Interestingly, the number of sites for each of the histone marks was similar in myotubes, myoblasts and cardiomyocytes, although only few sites occurred at the same position in two or all three cell types. In good agreement with the literature[228], the average lengths of DNA sequence associated with histone modifications were found to be approximately 600 bp. As approximately 147 bp of DNA is wrapped around one nucleosome[1], this translates to four consecutive modified nucleosomes. Although shorter stretches of enriched sequence were also found, approximately 75% of all modified sites were longer than 300 bp demonstrating that single modified nucleosomes rarely occur in mouse, as was previously reported also for yeast[79]. These findings correspond well to models where histone modifications are placed mainly by spreading mechanisms and, consequently, single modified nucleosomes rarely occur. Only about 25% of sites were above ≈ 900 bp long giving a stretch of six nucleosomes as upper range. Although from the size of the sites calculation of the number of nucleosomes is possible, single nucleosomes could not be distinguished on the arrays. The picture obtained in this study may be refined in the future by technological advances that may provide higher spatial resolution allowing the characterization of modifications on the level of single nucleosomes or histones. Such methods could employ combinations of ChIP-chip with nuclease digest[184] or mass spectrometry[23,323,324].

Co-occurrence of Histone Modifications

Assignment of the histone modified sites to transcription start sites showed that approximately half of the genes contributing to each transcriptome were associated with modified histones. Investigation into the combinatorial occurrence showed that the different modifications frequently occurred together, as was expected from previous studies[25,226-230] where pairwise correlations were observed to be high. However, in these studies a detailed analysis of the composition of modifications at one location or possible functional consequences of colocalizations compared to single occurrence were not investigated. In detail, the highest correlations were found between H3ac and H3K4me3 as well as between the two methylation states while the correlation of H3ac and H3K4me2 was only slightly lower. The modifications on histone 3 mainly occurred together, acetylation of histone 4

frequently occurred as single-code domain, that is without enrichment of any of the other modifications. Interestingly, the frequency of co-occurrences were highly similar in the investigated cell types.

5.2.2 Domains and TSSs

To reflect the modifications at one genomic location the term *modified domain* was introduced and each domain was assigned a code to represent the modifications at that particular position. The majority of TSSs were associated with only one modified domain, as modifications were typically clustered together. Enrichment of histone marks was found within ± 1 kb of the TSSs, as described previously[25,226-230]: However, only low levels of modifications occurred directly at the TSSs. This indicates that these regions are either rarely modified or contain few nucleosomes. From studies in yeast nucleosome-depleted regions at the TSSs are well known[24,325,326]. Recently, a study mapping the positions of nucleosomes in seven human cell lines showed that nucleosome free regions of approximately 250 bp upstream of the TSSs exist in higher eukaryotes as well[327], corresponding to the observed stretch of no modification observed in this study.

Different modified domains show positional preferences relative to the TSS above and beyond those of the individual modifications. In particular, H4ac is observed up- and downstream of the TSSs. However, on the basis of domains it could now be shown that if H4ac occurs downstream other modifications on histone 3 are also present, while single H4ac appears nearly always upstream. For yeast it has been reported that H3K4me2 occurs throughout transcribed regions. In higher eukaryotes, however, only a general association of this mark with transcribed regions has been observed so far[228,328]. It could now be demonstrated that in mouse the sites exclusively marked by H3K4me2 are distributed throughout transcribed regions, as in yeast.

5.2.3 Domains & Expression

Data presented in section 4.2.3 demonstrate that, indeed, histone tail modifications form a combinatorial code. Looking at the modifications individually, they were approximately equally associated with higher transcript levels, as has been previously reported[226-230,329]. This picture changes when the combinatorial nature of the modifications is considered. The analysis of the relationship between expression and modified domains from the perspective of transcript levels and from the perspective of modification associations gave consistent results. H3ac was found to have the strongest association with higher transcript levels, while additional activating modifications reduced this effect. Interestingly, when

investigating domains marked only by either H3ac or H4ac without additional modifications, H3ac is clearly associated with much higher transcript levels than H4ac (p value = 9×10^{-6}). This observation indicates that electrostatic interactions between histone tails and DNA do not solely determine the effect of acetylations.

H3K4me2 has been reported to correlate with active chromosomal regions[62,228] and to mark monoallelically expressed genes[330]. However, a recent study[331] employing an artificial system of an episome in human HEK 293 cells did not find any correlation of H3K4me2 with either gene expression or the presence of polymerase II. Furthermore, Schneider *et al.*[66] detected this modification in the inactive β-globin locus and an analysis of two HoxB genes has shown, that, although both *Hoxb1* and *Hoxb9* show acetylation of H3K9 and H3K4me2, only *Hoxb1* is expressed. This is attributed to the higher-order structure of chromatin, where the successive pattern of gene expression is achieved by sequential looping out of the decondensed chromatin state. Hence, in this case, although histone modifications may lead to a transcriptionally poised state, the temporal pattern of expression occurs as a consequence of gene extrusion and nuclear relocation. This transcriptome-wide study now shows, that H3K4me2 by itself or its combination with H3K4me3 is not positively correlated with transcript levels.

H3K4me3 has been employed as a marker to identify actively transcribed genes[332]. However, several studies have reported that a substantial percentage of transcripts associated with this modification were not expressed (31%[332] 24%[59,229,333]). Our results offer an explanation for these findings. It seems that H3K4me3 is not an optimal marker to identify transcribed genes and that the activating effect ascribed to it is mainly a result of its frequent co-localization with acetylations. Based on this analysis, histone acetylations, in particular H3ac, seem to be better predictors of elevated expression levels. Similar observations were made in two independent cell lines, one of which was considered in two differentiation stages. This suggests that these results are generally valid.

5.2.4 Differentiation

Finally, the role of these modifications in the process of differentiation was investigated. Histone 4 acetylation was found to be highly dynamic during this process; only half of the sites showing this modification were retained during the differentiation of myoblasts to myotubes. This indicates that H4ac plays an important role in this process. This observation was further substantiated by the analysis of differentially expressed transcripts comparing myoblasts and myotubes: In the vicinity of the TSSs of upregulated transcripts, H4ac was significantly often gained. However, a much larger number of TSSs was associated

with changes of histone modification than with significant differential expression. When analyzing the expression levels of all transcripts where modification conversions at the TSSs occurred, no significant changes were found. Although a large number of regions showed modification alterations, mostly these were not associated with changes in expression. A possible explanation is that modifications precede transcription, but are not by themselves sufficient for its regulation.

It is conceivable that the histone modifications that are gained and lost during differentiation could function as signaling marks. Single-gene studies have shown that significant levels of modifications are present in the vicinity of genes prior to transcription, resulting in a poised chromatin state[328,334]. Furthermore, histone modifications may serve as recognition sites for the recruitment of effector modules. Several bromo-, PHD-, and chromodomain proteins have been reported to bind to specifically acetylated and methylated lysine residues[76,223,335,336]. In this study, sequences associated with gain of H4ac in differentiation showed overrepresentation of Mef2 binding sites relative to sequences losing H4ac. Mef2 is a key transcription factor in differentiation and is known to be inhibited by histone deacetylases in myoblasts. During differentiation to myotubes Mef2 recruits histone acetyl transferases essential for differentiation, such as p300 and GRIP[337].

5.3 Transcriptional Regulation in the Heart by Transcription Factors

5.3.1 The Majority of TFBSs is Located Outside the Core Promoter

Using a ChIP-chip approach, in cardiomyocytes several hundred novel binding sites of the transcription factors Gata4, Mef2a, Nkx2.5, and Srf were identified and known binding sites confirmed. Several genes known to be dysregulated in cell culture or mouse models could now be shown to be direct targets of the investigated TFs. Although ChIP-chip experiments have been previously performed for Mef2[285,338,339] and Srf[340], these studies used arrays covering a much smaller portion of the genome; moreover, the binding of TFs can be cell-type-dependent and now cardiomyocytes were investigated for the first time. In case of Gata4 and Nkx2.5 this is the first study where in a large scale approach directly bound target genes could be identified.

In the array design 10 kb upstream and the first intron of each TSS (gene) was represented. This allowed a precise characterization of the binding of the TFs relative to the TSSs of target genes. The absolute number of target genes varied widely between the four TFs the position of the TFBS relative to TSSs was similar. The TFBSs are approximately symmetrically distributed relative to the TSSs. A recent study[341] also found regulatory elements to be symmetrically distributed around TSSs in the human genome. These findings indicate that such a positioning might be a common eukaryotic feature. Additionally, more than one third of the TFBSs was identified within transcribed regions, in particular within the first intron of target genes. This implies that intronic sequences may have a more influential role in transcriptional regulation than previously assumed.

5.3.2 Gene Ontologies of Direct TF Targets Match Mouse Model Phenotypes

To understand the pathways in which the investigated TFs are involved it is necessary to understand the function of their direct targets. Therefore, gene ontology terms (GOs) based on biological processes were assigned to the genes where TF binding was observed. The terms found to be significantly overrepresented for each TF ($p \leq 1 \times 10^{-4}$) were in excellent agreement with previously reported phenotypes of mouse models, as discussed in the following paragraphs. Moreover, additional GOs may point to novel functions in pathways hitherto not associated with the TFs.

It has been reported, that abolishing of Gata4 activity in cell culture leads to apoptosis and a reduced ability to differentiate to cardiac myocytes[128,129]; in accordance with this finding the term 'cardiac cell differentiation' is overrepresented. In addition, GOs overrepresented for Gata4 target genes include 'heart development', 'striated and skeletal

muscle development', 'muscle contraction' and 'regulation of muscle contraction', 'cardiac muscle morphogenesis', and 'cardiac inotropy'. In agreement with Gata4 expression in T-cells[342] genes annotated to be important for 'immune system development' could also be found.

Similarly, 15 GO terms for Mef2a targets were related to muscle or heart development and function. Overrepresentation of 'muscle contraction' and 'sarcomere organization' are in accordance with the phenotype of murine *Mef2a*$^{(-/-)}$ mice[159] where myofibrillar fragmentation was observed; the activation of the fetal cardiac gene program observed in the knockout mice is reflected in overrepresentation of genes involved in 'embryonic heart tube development'. In addition a role in 'blood vessel formation' and 'transcriptional regulation from RNA polymerase II promoter' could be identified.

In *Nkx2.5* hypomorphs looping morphogenesis is not initiated in the linear heart tube stage[343] and cell proliferation[146,344] is diminished. Furthermore, the recruitment of myocytes to the conduction system is reduced[344]. These functions of Nkx2.5 are reflected in the overrepresentation of GOs for 'heart looping', 'positive regulation of cell proliferation' and 'cell motility'. Interestingly, genes annotated for 'biomineral formation' and 'bone mineralization' were also found significantly often, implying novel functions of this TF.

Srf targets are known to play a role in a broader number of pathways including cell growth, migration, cytoskeletal organization, and myogenesis. This is reflected in the GO terms overrepresented for Srf targets which cover a broad range of biological processes. Skeletal muscle cells lacking *Srf* have defective formation of cytoskeletal structures, in particular actin fibers[189]. Mice lacking skeletal muscle *Srf* expression form muscle fibers but fail to undergo hypomorphic growth after birth leading to skeletal muscle hypoplasia and death[192]. Genes involved in 'actin cytoskeleton organization' and 'biogenesis' were found to be overrepresented. Similarly, in *Srf-null* cardiomyocytes severe defects in the contractile apparatus including stress fiber formation, as well as mislocalization and /or attenuation of sarcomeric proteins is reported[190]. Srf target genes were significantly associated with 'muscle contraction' and 'regulation of heart contraction'. Mice lacking cardiac *Srf* expression are reported to be embryonic lethal due to impaired chamber maturation and reduced cellularity[345]. Among Srf target genes GO terms for 'positive regulation of cell proliferation' and 'embryonic heart tube development' as well as 'tube morphogenesis' are overrepresented. Additionally genes playing a role in 'transcription' as well as 'nucleic acid metabolic processes', 'RNA biosynthetic processes', and 'DNA-dependent transcription' were identified.

It was observed that the investigated TFs frequently bind together. Consequently, the target genes were classified depending on how many and which combination of TFs they were bound by. This resulted in eleven groups of genes regulated by more than one TF (compare Table 4-12). For each group GO terms relating to heart development and function are significantly overrepresented. In particular, among those genes bound by all four investigated TFs nearly one quarter (15 out of 63) are either transcription factors or cofactors several of which have been implicated in heart development and disease, suggesting that the investigated TFs can be placed at the top of several cardiac transcriptional pathways.

5.3.3 Gata4, Mef2a, Nkx2.5, and Srf Frequently Bind Together

It is known that each of the four investigated TFs can interact with at least one of the other four. However, this knowledge was based on single-gene studies or artificial systems employing overexpression. Therefore, the extent to which co-regulation occurs *in vivo* was unknown. The current investigation now demonstrates that the analyzed TFs have a high number of target genes in common and also frequently bind at the same position.

Gata4 and Nkx2.5 are mutual cofactors and only overexpression of both leads to cardiogenesis in lineage committed precursors[106]. In Gata4 the interaction maps to the C-terminal zinc finger and a C-terminus extension, similarly, in Nkx2.5 to a C-terminally extended homeodomain. Possibly the interaction induces a conformational change that unmasks Nkx2.5 activation domains. Although Gata4 and Nkx2.5 each have a relatively low number of target genes; their binding is highly correlated as they bind together at 204 genes and at 163 of these even within a 500 bp window.

Srf and Gata4 bind together at a similar number of 162 genes within 500 bp. The interaction of Srf and Gata4 has been shown to occur through the MADS box domain of Srf and the second zinc finger of Gata4, while the first zinc finger acts inhibitory[110]. The zinc finger of Gata4 is also necessary for the interaction with the activation domains of Mef2 proteins[105].

Srf is known to interact with a wide variety of proteins. These include members of the Nkx2.5 family of homeodomain proteins[172] which can form complexes both with Srf and their own adjacent recognition site[346]. In this study the binding was highly correlated with 151 genes where these two proteins bind within 500 bp.

A direct physical interaction between any of the Mef2 proteins and Nkx2.5 has so far not been reported. On the one hand, based on the high odds ratio of ≈ 4 and the high number of 226 genes where binding of Mef2a and Nkx2.5 is observed within a 500 bp window it is probable that here also direct physical interaction occurs. On the other hand, it is also possible

that the interaction is mediated by a third protein such as Gata4. This appears probable, because in nearly all cases where Mef2a and Nkx2.5 bound, Gata4 binding was also observed.

The occurrence of Srf and Mef2a, however, was anti-correlated, as the odds ratio to find binding sites of these two proteins in the vicinity of the same gene was low (≈ ⅛). A previous study[347] reported that human MEF2 and SRF compete for binding at least one target gene. On the one hand, this may be possible as both proteins interact with DNA through a MADS-box. On the other hand, the binding motifs differ strongly. Additionally, the calculated anti-correlation may be distorted; although the relative number of shared genes between Mef2 and Srf is low (291 out of 1655 and 1509 direct target genes, respectively), the absolute number of 291 is in a comparative range to the physically interacting proteins.

5.3.4 Few TFBSs Show Evolutionary Conservation

Previous reports have determined binding motifs for Gata4, Mef2a, Nkx2.5, and Srf and deposited these in the data base TRANSFAC[348]. These motifs were matched against the sequences underlying the enriched sites in ChIP-chip. This led to several conclusions valid for Gata4, Mef2a, and Nkx2.5: First, the TRANSFAC motifs can be retrieved in nearly all binding sites (> 80%) and frequently more than once. As the motifs are generally very short, this however, is not surprising. Second, only ≈ 30% of TFBSs were found within conserved elements of whole vertebrate alignments between eleven species (PhastCons elements[247]). It has generally been thought that sequences harboring regulatory motifs are highly conserved, however, current investigations challenge this belief. Third, this observation of little conservation also holds true when requiring an exact match of the motif sequence in the mouse and human alignment. These results indicate that regulatory regions evolve far more rapidly than previously assumed.

Recently, results for the first 1% of the human genome of the Encyclopedia of DNA Elements (ENCODE) have been published[341]. The ENCODE project aims to identify and catalogue the functional elements encoded in the genome. Overall it was found, that about 40% of evolutionarily constrained regions were without detectable biological function. On the flipside about half the functional elements found in non-coding DNA are unconstrained. This study suggests the possibility of a large pool of neutral elements that are biochemically active but provide no specific benefit to the organism. This pool may serve as a 'warehouse' for natural selection, potentially acting as the source of lineage specific elements. Such a model is supported by the data presented here.

5.3.5 Refinement of TF Binding Motifs

The *de novo* search for motifs in the enriched sequences for Gata4, Mef2a, and Nkx2.5 retrieved motifs highly similar to those previously determined by *in vitro* selection (motifs were obtained from TRANSFAC[260]: Gata4[305] , Mef2a[307], and Nkx2.5[141]). The sequences could be refined and extended by further base pairs. Additional investigations will be necessary to determine whether alternative motifs may be related to cell type, function or cofactor and if the transcription factors bind to these with equal affinity.

Recently, Cooper *et al*.[340] investigated SRF binding in three different human cell types by ChIP-chip and were unable to retrieve the CArG-box as binding motif. They ascribe this to the CG-rich nature of these sequences and the amount of input sequence confounding all used motif-finders. However, in the present study the *de novo* search yielded convincing results for the three other TFs. Therefore, it appears improbable, that the inability to retrieve the Srf motif is due to the search algorithm. Srf is known to bind to sequences differing from the CArG box and thereby distinguishing genes involved in cell growth, which often have perfect CArG-boxes, and genes for myogenesis, which often have one or more nucleotide substitutions[349]. Investigations focusing on individual target genes will show if the identified motifs are related to different functionalities or might be cardiomyocyte specific.

5.3.6 TFBSs are Marked by Histone Modifications

Determining the co-occurrence of TFBSs and four histone modifications (H3ac, H4ac, H3K4me2, and H3K4me3) revealed that these modifications occur at 60-80% of all TFBSs which is far more frequently than would be expected by chance (≈ 30%). However, a preference for one particular histone modification could not be observed for any of the TFs due to the high co-occurrence of the investigated histone modifications. The frequent co-occurrences between histone modifications and TF binding prompted the question whether this influences the expression levels of the target genes. Therefore, the target genes were grouped according to whether only TF binding or also histone modifications were observed. It was tested whether the additional occurrence of histone modifications was associated with higher transcript levels.

Transcripts showing binding of Gata4 and Srf as well as acetylation of histone 3 at the binding site were significantly elevated. A possible reason could be that the acetylation of the histone tails leads to loosening of the chromatin structure resulting in a higher number of TF binding occurrences. On the other hand, this should also apply to Mef2a and Nkx2.5 as well as for histone 4 acetylation. However, here this effect was not (or less) pronounced.

Gata4 and Srf are both known to physically interact with the histone acetyl transferase p300. Indeed, human GATA4 only obtains its full DNA-binding potential when acetylated by this enzyme[114]. In case of Srf the co-activator Myocardin (Myocd) is known to recruit p300 to Srf binding sites and thereby promotes gene activation[350]. Although p300 can also acetylate other lysines it is known to acetylate the lysine residues K14 and K18 on histone 3[33]. The opposing function is carried out by histone deacetylases (HDACs). Interestingly, if embryonic stem cells are treated with inhibitors of HDACs the levels of acetylated GATA4 increase and the cells differentiate into cardiac myocytes[283]. This implies that the action of p300 on histone modifications, possibly mediated by Gata4 and Srf, is a critical step in heart development.

Mef2 proteins are also known to interact with p300 but target genes with both Mef2a binding and additional H3ac did not show elevated levels compared to those with only Mef2 binding. However, in case of Mef2 proteins the regulation by histone acetyltransferases and deacetylases is far more complex. Mef2 proteins bind to their target genes but are frequently repressed by interaction with class I and class II HDACs[351]. In response to differentiation signals, HDACs are exported out of the nucleus and the MEF2 proteins are free to recruit HAT enzymes, in particular GRIP[352] and p300[353], leading to gene activation. Furthermore, members of the Mef2 family have been reported to interact with enzymes involved in changing the methylation state of histones such as the arginine methyl transferase CARM1[354] and the cardiac transcriptional repressor Jumonji (Jmj)[355]. These mechanisms might indicate why no trend for genes with both binding of Mef2a and histone modifications could be observed. However, it is clear, that mice with conditional knockout of *Mef2* display chamber dilation. Mice lacking *HDAC5* or *HDAC9*[88] display enhanced hypertrophy; deletion of both enzymes leads to early postnatal death from a spectrum of cardiac abnormalities including ventricular septal defects and thin-walled myocardium[89]. Deletion of both *HDAC1* and *HDAC2* results in neonatal lethality and a variety of cardiac abnormalities[90]. These findings underline the critical function of histone-modifying enzymes in heart development.

In case of Nkx2.5 targets, all four investigated histone modifications were associated with a minor additional elevation of transcript levels. Not much is known of interactions between Nkx2.5 and histone-modifying enzymes, however, an association with Jumonji (Jmj) has been described[356]. Recently, one of the human members of the Jumonji family has been shown to be a histone demethylase specifically converting tri- to di-methylation on H3K9me3 and H3K36me3[357]. H3K9me3 has been associated with transcriptional repression. Mice with a homozygous knockout of the *jmj* gene showed abnormal heart development including ventricular septal defects, double-outlet right ventricle, and thin ventricular wall[358].

The interaction of the TF with histone-modifying enzymes and the associated cardiac phenotypes emphasize the importance of the interplay between transcription factors and histone modifications. Currently two models are possible. In one case, first the transcription factor binds because a specific binding motif is recognized, and then the histone-modifying enzymes are recruited. Histone modifications would then lead to a more open chromatin structure making the DNA accessible for the transcriptional apparatus. In the second scenario which would be in accordance with the *histone code hypothesis*[78], the histone modifications are first placed and then act as signaling marks for the recruitment of transcription factors. This would also offer an explanation why nearly all binding sites contain the respective TF binding motif while only a fraction of all motifs existent in the genome are bound. Currently, a rapidly growing number of proteins that recognizes specific histone modification marks is being identified[33]. A picture is emerging in which a tight interplay between histone modification, histone-modifying enzymes, and transcription factors plays a central role in transcriptional regulation.

5.3.7 The TFs Function Mainly as Activators

As different cofactors influence whether a transcription factor functions as activator or repressor, knowledge of binding does not equal knowledge of regulatory function. In a first approach the expression levels of the target genes as identified by ChIP-chip in untreated HL-1 cells was investigated. Approximately 80% of the targets were expressed. Compared to the average expression level of the HL-1 transcriptome the average expression level of the TF target genes was significantly higher indicating that these TFs act primarily as activators.

To gain this information for each individual target gene the cardiomyocytes were treated with siRNAs directed against the TFs and the transcriptome was analyzed on genome-wide expression arrays. The majority of differentially expressed target genes appears to be activated by binding of the investigated TFs. This activating function of the TFs is not only reflected in the number of targets but also in the higher fold changes of the corresponding transcripts. However, the transcription factors can also function as repressors of transcription as is expected from the known interactions with corepressors such as histone deacetylases.

5.3.8 Direct Targets and Differential Expression

The overlap between the genes identified to be direct targets in ChIP-chip analysis and the differentially expressed transcripts in siRNA treated cells was determined. The comparison revealed that for each TF approximately one third of the bound genes is not differentially expressed. Especially for essential target genes evolution has favored systems in

which the loss of one regulatory transcription factor may be compensated by another; this may explain the low overlap. In future, depletion of several TF at once may shed light on redundantly regulated genes.

In case of Mef2a 999 binding sites were identified, but only 119 transcripts were differentially expressed in the siRNA treated cells. Previous studies reported[280], that members of the Mef2 family, can at least partially take over each others functions. In case of Srf the number of differentially expressed transcripts (519) was also found to be lower than the number of binding sites (1,335). In case of Nkx2.5 and Gata4 a higher number of differentially expressed transcripts compared to the TF binding sites was found (Nkx2.5: 383 binding sites, 782 differentially expressed and Gata4: 447, 621) indicating that a high number of downstream targets was also dysregulated in knockdown experiments. The discrepancy between bound target genes and regulated genes has also been observed in other studies[359-361]; the observed overlap for Gata4, Mef2a, Nkx2.5, and Srf was low but in a comparable range with published data Table 5-1.

Table 5-1. Overlap between genes bound in ChIP-chip and differential expression in siRNA treated cells for Gata4, Mef2a, Nkx2.5, and Srf and comparison to published data.

	Oct4[359]	Nanog[359]	HSF1[360] (no heat shock)	HSF1[360] (heat-shock induced)	TAL1[361]	Gata4	Mef2a	Nkx2.5	Srf
siRNA signifi-cantly different-tially expressed transcripts	4,711	2,264	354	449	6	2,741	472	3,360	1,543
ChIP direct target TSSs	1,083	3,006	75	417	36	468	1,655	392	1,509
Overlap (% of direct targets)	394 (36%)	475 (16%)	2 (3%)	50 (12%)	6 (17%)	62 (13%)	32 (2%)	55 (14%)	112 (8%)

There are several different explanations for this phenomenon which are not mutually exclusive. For one, it is possible that as in the case of the Mef2 family, the loss of one protein can be compensated by another similar protein which is also present in the cell[280]. Similarly, the expression of genes which are redundantly regulated even under normal conditions may be hardly influenced if only one of the regulators is lost. In these two models the TF binding would be functional, but only detectible if all TFs capable of binding the target are removed from the cells.

However, it is also conceivable that the TF binding is not functional. A ubiquitously expressed TF, e.g. Srf might bind wherever the binding site is present and accessible but may need cofactors for functionality. If the cofactors are expressed in a more cell type-specific

manner, this might be a mechanism for bestowing cell type-specificity. In the light of recent findings by Cooper *et al*.[340], however, this model appears less probable. In this study binding of Srf in three different human cell lines was investigated, but only few binding events occurred at the same position in two or all three cell types.

Possibly, functionality in binding requires the presence of cofactors which may only be expressed in response to external signals, for example differentiation signals. In this scenario the binding could be considered as a poised state. Further investigations studying knockdown of several regulators at once and the binding pattern during dynamic processes such as differentiation will be necessary to resolve this important issue.

5.3.9 A Gene Regulatory Network

Using chromatin immunoprecipitation analysis a comprehensive list of *bona fide* target genes was obtained. This data resulted in the generation of a first network focused on the depiction of co-binding and biological process gene ontology annotation. This large graph visualizes the number of target genes and the number of genes shared between the TFs. To enable a convenient study of the gene ontologies the network will be made available online. This data delineates the architecture of a cardiac regulatory network.

Regarding the subnetwork of Gata4, Mef2a, Nkx2.5, and Srf it has previously been shown in several studies that to some extent regulation between these TFs takes place (Figure 5-3 A). However, here a tight interplay could be demonstrated and new interactions could be demonstrated (Figure 5-3 B).

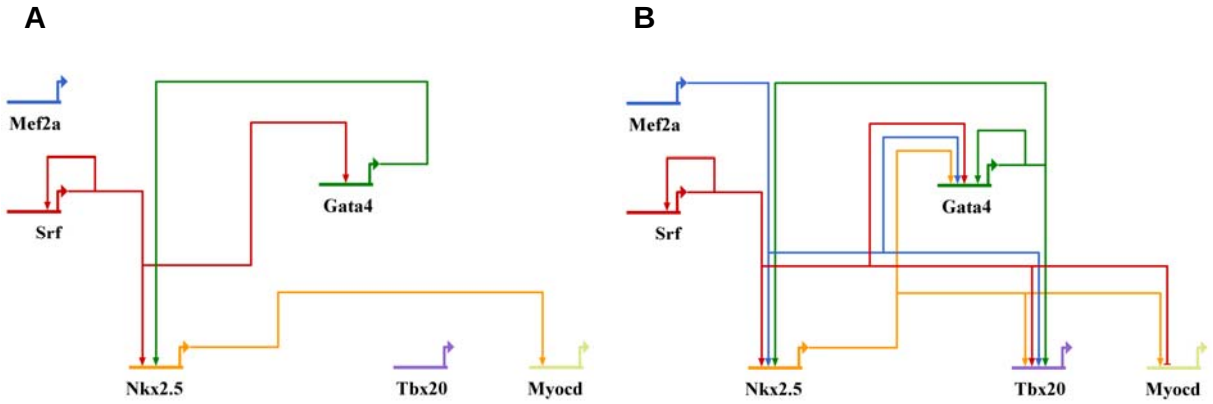

Figure 5-3. Core regulatory network of Gata4, Mef2a, Nkx2.5, and Srf according to published[362] data (A) and supplemented with data from this study (B). Gata4, Mef2a, Nkx2.5, and Srf regulate of *Tbx20*. Expression of *Myocd* is activated by Nkx2.5 and repressed by Srf. Myocd is an activating cofactor of Nkx2.5 and Srf.

Each of the investigated TFs has been linked to severe cardiac diseases. However, the phenotypes of patients vary widely for one and the same mutation or resemble each other for different mutations. This observation is less surprising when viewing the intricate regulatory mechanisms that have emerged from this study, although only four TFs were investigated.

The ChIP data was complemented by expression array data obtained from siRNA knockdown experiments of the same TFs. By integration of the two data sets the list of direct target genes could be reduced to a comprehensive set of regulated direct targets. The resultant network sheds light on the pathways regulated by the TFs. Interestingly, for genes which were regulated by two or more TFs the regulation was either all activating or all repressing. Only two exceptions were found: First, *Rbpms* (RNA binding protein gene with multiple splicing) is activated by Nkx2.5 and repressed by Gata4, and second, *Myocd* (myocardin) is activated by Nkx2.5 and repressed by Srf. The regulation of *Myocd* is of particular interest, as Myocd has been reported to act as activating co-factor for both Nkx2.5 and Srf. Myocd has long been known to be essential for smooth muscle development[176] and a recent investigation suggests that Myocd may function as a switch between smooth and skeletal differentiation programs[177]. The opposing effects of Nkx2.5 and Srf binding on *Myocd* expression point to a novel regulatory circuit. While the gene ontology terms overrepresented in the lists of direct targets match well with the observed mouse phenotypes it will be interesting to elucidate the contribution of each target in the context of muscle function and development.

5.3.10 Expression of *Tbx20* is Activated by Gata4, Mef2a, Nkx2.5, and Srf

One of the genes appearing at the center of the gene regulatory network of Gata4, Mef2a, Nkx2.5, and Srf was the T-box transcription factor *Tbx20*. Several members of this family are known to play a role in the development of CHDs. Deletions of *TBX1* have been reported in patients with *DiGeorge Syndrome* and mutations or haploinsufficiency of *TBX5* are frequent causes of *Holt-Oram Syndrome* associated with atrial septal defects as well as first and second degree atrioventricular block. *Tbx20* is one of the first genes expressed in the vertebrate cardiac lineage and its expression pattern is conserved form Drosophila to humans. Reduced *Tbx20* expression results in abnormal heart morphogenesis in zebrafish[313] and mouse[314,315]models, indicating an essential role for Tbx20 in the formation of the outflow-tract and right ventricle. Heart biopsies from patients with *Tetrology of Fallot* (*TOF*) show increased *TBX20* levels (unpublished results). Furthermore, the Tbx20 protein is known to interact with two of the investigated TFs: Gata4 and Nkx2.5. While these previous investigations clearly showed that dysregulation of *Tbx20* leads to severe cardiac phenotypes, so far little is known of the processes regulating its expression.

Here, it could now be demonstrated that Gata4, Mef2a, Nkx2.5, and Srf are functional regulators of *Tbx20* expression. In cardiomyocytes the promoter of *Tbx20* is marked by activating histone modifications. In skeletal muscle cells, where *Tbx20* is not expressed, these modifications are absent. H4ac, H3ac, H3K4me2, and H3K4me3 are generally believed to decrease the compaction of chromatin, thereby making is accessible for the binding of TFs. Indeed, the binding of Gata4, Mef2a, Nkx2.5, and Srf was observed precisely at the regions of maximal modification. Three binding sites were observed and confirmed in qPCR, each site showing enrichment for all four transcription factors. Analysis of the expression levels in siRNA treated HL-1 cells showed that the TF binding has an activating effect on transcription. It can therefore be speculated, that in *TOF* patients with reduced *TBX20* levels one of the TFs or a binding site in the *TBX20* promoter is mutated. *GATA4* and *Nkx2.5* mutations have also been reported as disease genes of familiar atrial septal defect[134,135] and implicated in causing *Tetralogy of Fallot* (TOF)[136]. However, it remains to be elucidated whether all three binding sites are equally functional and whether mutations within these binding sites can be causative for congenital heart diseases such as *TOF*.