4 RESULTS

4.1 Development of ChIP Arrays and Analysis Algorithms

4.1.1 The Design of Heart and Skeletal Muscle ChIP Arrays

The most robust and favorable array design for ChIP-chip experiment would be an array tiling the entire genome of the organism of interest, including all non-coding regions. However, for higher eukaryotes this would require a huge number of spots on a large number of arrays at immense cost, making it necessary to focus on genomic regions of special interest. An array design tailored to investigations of muscle development and differentiation was not commercially available. Therefore, a first goal of this study was to design an array representing all genes expressed either in heart, smooth or skeletal muscle either in mouse or human. Expressed transcripts were identified using several sources listed in Chapter 3.2, Table 3-1.

A ChIP Array Design for Investigation of Histone Modifications

The next step was to select regions most likely to contain regulatory elements. These are thought to be predominantly located in the promoter region of approximately 5 kb upstream and in the first intron of a gene. Furthermore, several studies reported that regulatory sequences tend to be evolutionarily conserved[269-271]. In particular, sequence conservation in non-coding, upstream regions of orthologous genes from man and mouse is likely to reflect common regulatory DNA sites[272]. Accordingly, the array was designed to represent the man-mouse conserved regions of 5 kb upstream of each selected gene, the first exon and first intron.

This resulted in the compilation of a ChIP-chip array design covering 8,585 genes (corresponding to 10,976 transcripts) designed for investigations of histone modifications in murine muscle. In addition an expression array representing the same set of genes was designed. The array designs are now available from NimbleGen with the design ID 2389 in case of ChIP-chip arrays and ID 2390 in case of expression arrays, respectively.

A ChIP Array Design for Investigation of TF Binding

The NimbleGen array ID 2389 is the first array usable for the ChIP investigation of histone modifications in heart and skeletal muscle cells. However, to investigate TF binding a new design was necessary for three reasons: First, the spacing had to be reduced. As 147 bp of DNA is wrapped around a nucleosome, a spacing of 85 bp was sufficient for the investigation of histone modifications, but the binding sites of TFs are usually only few nucleotides long.

Second, technical advances made a higher array content at constant cost possible. Therefore, a greater upstream portion of each gene could be represented. Third, as one aim of the investigation was to elucidate to what extent binding occurs in non-conserved regions or repeats, it was desirable also to represent regions with very low conservation. Therefore, an array was designed for the same set of genes as before but covering 10 kb upstream, the first exon and intron at high resolution. For technical reasons probes were distributed on two arrays first array contains chromosomes 1 to 8 and the second 9 to X. The array design is now available from NimbleGen with the with the design ID 4853/4854.

4.1.2 *Ringo* – an R/Bioconductor Package for Analyzing ChIP-chip Readouts

ChIP-chip is a high-throughput assay with which the binding of protein to DNA can be detected. However, the raw microarray intensity readings themselves are not immediately informative; enriched regions need to be bioinformatically annotated and compared to related datasets. At the time of the experiments no program for the analysis of the NimbleGen ChIP-chip raw data was supplied or available. Therefore, an algorithm was developed for the identification of high-confidence signals on the arrays. Several factors may lead to noise in the intensity of the probe levels. The procedure to identify enriched sites will be explained using the ChIP-chip signals for the acetylation of histone 3 in the vicinity of the gene Hand2 as an example (Figure 4-1 A).

First intensities of probes from biological replicate samples are averaged and a running median over the probe intensities in a fixed window is calculated (here a 800 bp was used, Figure 4-1 B). To allow for different efficiencies of antibodies, the cut-off must be determined for each antibody (histone modification) separately. For this purpose, the above smoothing procedure is repeated on the same data but with randomly permuted probe positions. In this example, it was decided to use the 99% quantile as threshold, which allows for a nominal false discovery rate of one percent. Averaged probes with intensities greater than the threshold are considered to be enriched by the ChIP procedure (Figure 4-1 C). It is still possible, that single probe outliers are identified as enriched sites. Therefore, depending on the experimental set-up, a minimum number of consecutive probes is required to call a site enriched. In this case, only such locations that contained at least three enriched probes less than 600 bp apart were considered to represent histone-modified sites (Figure 4-1 D). This number was based on the consideration that generally two or more consecutive histones with one modification occur[79,228]. The parameters of the analysis have to be adjusted according to the experimental set-up. In case of the transcription factor binding analysis, for example, the

expected binding sites are shorter and the resolution higher (for details compare Chapter 3.3.2).

The algorithms were published[257] as an open source package called *Ringo* which is available from the Bioconductor project[252]. Besides the detection algorithm for enriched sites additional functionalities of *Ringo* include import, quality assessment and preprocessing of the raw data provided by NimbleGen as well as visualization of the raw and processed data.

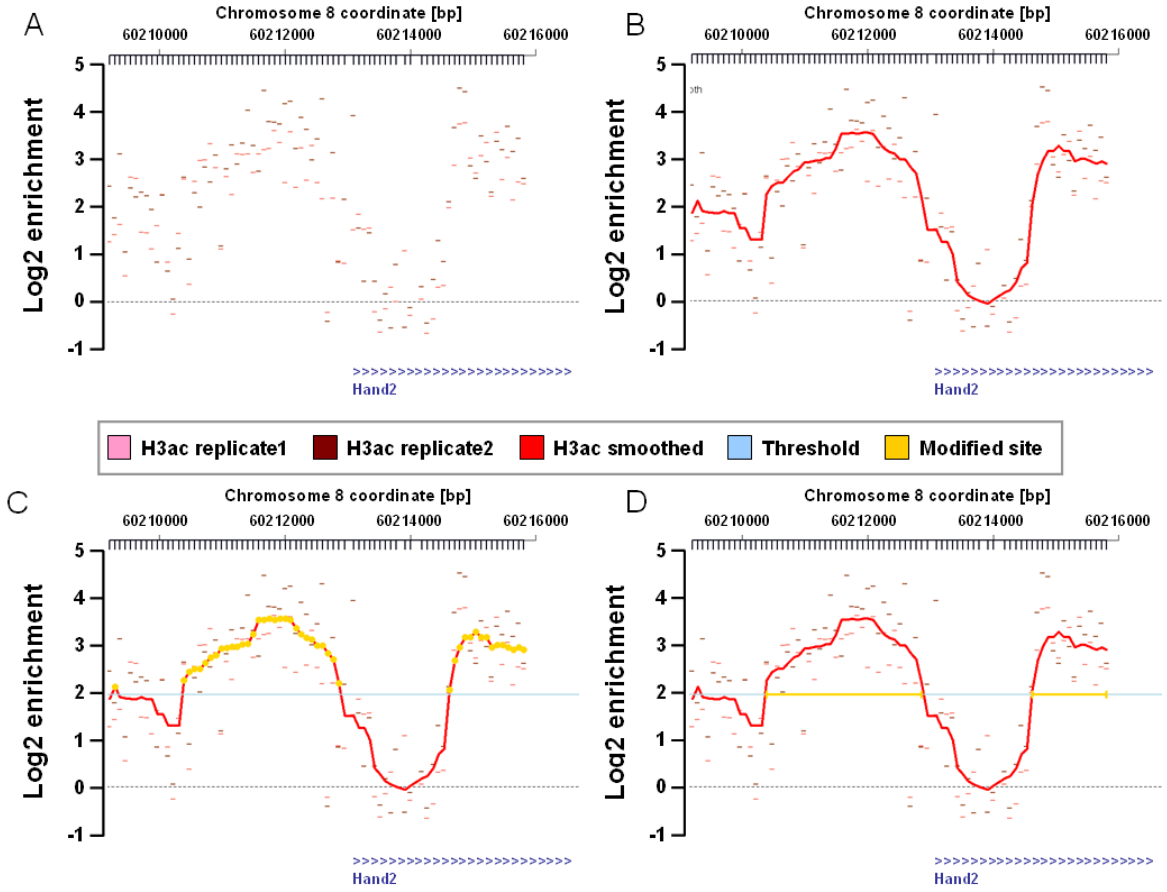


Figure 4-1. Identification of sites marked by modified histones exemplarily shown by histone 3 acetylation in the vicinity of the TSS of Hand2 gene. A) Normalized array data for H3ac replicate 1 and 2. B) From probe intensities of replicate 1 and 2 a running median is calculated (red line, H3ac Smooth). C) A permutation-based threshold with a nominal false discovery rate of 1% is defined (light blue line, threshold), average probe intensities above this threshold are considered to be enriched (yellow). D) More than three consecutive probes with intensities above the threshold define a modified site.

4.2 Combinatorial Properties of Histone Modifications

4.2.1 Gene Expression Patterns of Heart and Skeletal Muscle Cells

Function of Differentially Expressed Genes

To understand the effect of the histone modifications H3ac, H4ac, H3K4me2, and H3K4me3 on expression a model system of three cell types was used: skeletal muscle C2C12 cells in undifferentiated (myoblasts) and differentiated state (myotubes) and HL-1 cells (cardiomyocytes) as shown in Figure 4-2. The first step was to characterize the transcriptome of each cell type and to investigate the differences. As the expression pattern of a cell line may change with serial passaging[273] RNA was prepared from each cell type at three different passages and the RNA from these isolations was pooled[274], so that every pool represents a cross section of passages 2 to 20. The resultant RNA preparations were analyzed using the NimbleGen arrays ID 2389; significantly differentially expressed genes were identified and overrepresented gene onotologies were determined.

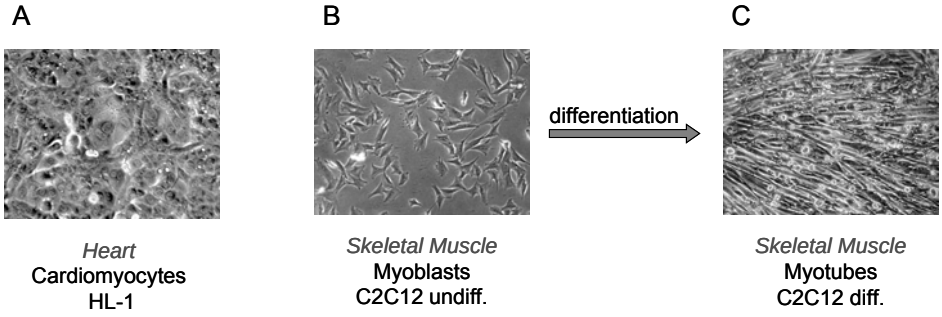


Figure 4-2. Cell lines used in the study. A) HL-1 cells are cardiomyocytes which contract in cell culture, a video is available on request. B) C2C12cells are derived from skeletal muscle and can be differentiated from myoblasts (B) to myotubes (C).

Between the undifferentiated C2C12 myoblasts and differentiated C2C12 myotubes 299 transcripts were found to be differentially expressed. The significantly overrepresented gene ontology terms of the upregulated genes in this differentiation process were: 'regulation of muscle contraction', 'contractile fiber' and 'actin cytoskeleton' (Supplementary Table 1). The observation, that upregulated genes are associated with these terms clearly indicates that muscle fibers were formed. In differentiation, cells cease to divide. Consequently, genes down-regulated in this process are associated with gene ontology terms such as: 'mitotic cell cycle', 'DNA replication', 'cell division', 'cell cycle progression' (Supplementary Table 2). These findings are in good agreement with previous studies which reported genes involved in cell cycle withdrawal, muscle differentiation and apoptosis to be differentially expressed in the course of C2C12 differentiation[233,275,276].

More than twice as many (632) differentially expressed transcripts were found between C2C12 undifferentiated and HL-1 cells, as would be expected from comparison between two different, but similar tissues. Gene ontology annotations overrepresented for significantly upregulated genes in HL-1 compared to C2C12 undifferentiated cells include 'muscle contraction' and several metabolic pathways (Supplementary Table 3) in accordance with the higher energy turnover of contracting cardiomyocytes. For genes downregulated in HL-1 cells compared to C2C12 undifferentiated cells 'skeletal muscle development' is overrepresented (Supplementary Figure 4). Between differentiated C2C12 and HL-1 cells even more transcripts (926) were found to be differentially expressed. This reflects the higher similarity of cardiomyocytes to myoblasts than to myotubes.

Verification of Expression Array Analysis by Real-Time PCR

Although transcript levels were calculated by averaging the values of approximately 15 probes per transcript, results from microarrays may be falsified by confounding factors such as probe GC-content or hybridization artifacts. Therefore the array intensities as well as the fold changes calculated between two cell types were verified for 15 transcripts in each cell type. The microarray data was confirmed in all cases (Supplementary Figure 1 and Supplementary Figure 2).

4.2.2 Histone Modification Patterns of Heart and Skeletal Muscle Cells

4.2.2.1 Specificity of Antibodies used in ChIP Experiments

Antibody specificity is a major confounding factor in chromatin immunoprecipitation experiments. Although the antibodies used in this study were previously described to be specific in the same application[25] the quality may be lot-dependent. Therefore, Western blot analysis using calf thymus histones was carried out. Figure 4-3 shows that each antibody only gives one band corresponding to the expected size.

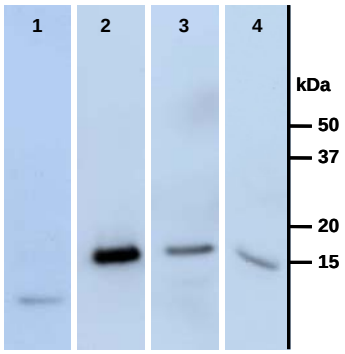


Figure 4-3. Control of antibody specificity by Western blot analysis of calf thymus histones. Lane 1: anti-H4ac (14 kDa), lane 2: anti-H3ac (17 kDa), lane 3: anti-H3K4me3 (17 kDa), lane 4: anti-H3K4me2 (17 kDa). The figure was kindly supplied by Ilona M. Dunkel.

4.2.2.2 The Median Modified Site Contains Four Nucleosomes

Chromatin immunoprecipitation of the four histone modifications H4ac, H3ac, H3K4me2, and H3K4me3 was carried out in each of the three cell types C2C12 undifferentiated, C2C12 differentiated and HL-1. The first step in the array analysis was to identify the chromosomal regions where modified histones are located. In total, in each cell type between 2,900 and 3,600 sites for each of the four histone modifications was identified (Table 4-1). Interestingly, the number of histone 4 acetylated sites was consistently lower than for the other three modifications. The median sizes of the sites were approximately 600 bp, corresponding to four consecutive nucleosomes containing histones with the modification of interest. This value corresponds well to a previous study[228], where the distribution of H3ac, H3K4me2 and H3K4me3 in HepG2 cells across human chromosomes 21 and 22 was investigated.

Table 4-1. Number and median size of histone modified sites per modification and cell type and comparison to published data.

Modifications	Number of Modified Sites			Median Site Size [bp]			
	C2C12 undiff	C2C12 diff	HL-1	C2C12 undiff	C2C12 diff	HL-1	HepG2[228]
H3ac	3,059	3,248	3,210	637	645	609	703
H4ac	2,925	3,026	2,940	561	561	543	No data
H3K4me2	3,297	3,205	3,378	621	613	608	605
H3K4me3	3,493	3,538	3,357	647	647	607	659

To compare the number of identified sites with previous results[228] the average number of modified sites per gene was calculated. This was based on the absolute number of identified modified sites identified in each study (Table 4-1 and Table 1 from Bernstein *et al*.[228]) divided by the number of genes represented on the arrays (8,585 genes on the ChIP arrays used in this study and a total of 1,397 genes on Chr 21 and Chr 22 according to Ensembl v36).

Table 4-2. Sites enriched for histone modifications and comparison to published data.

Modifications	Average Number of Sites per Gene			
	C2C12 undiff.	C2C12 diff.	HL-1	HepG2[228]
H3ac	0.34	0.36	0.35	0.50
H4ac	0.32	0.33	0.32	no data
H3K4me2	0.36	0.35	0.37	0.31
H3K4me3	0.39	0.39	0.37	0.36

4.2.2.3 Co-occurrence of Histone Modifications

To gain an understanding of the combinatorial co-occurrence of histone modifications, it was analyzed how often genomic locations were enriched for one, two, three or all four modification types. The number of possible combinations of modifications, without counting non-modified as combination, that can occur at one genomic position is given by 2^N-1 with N the number of investigated modifications. Consequently, the four investigated histone marks can occur in 15 different combinations at one position. However, in the investigated cell lines only a few of these combinations occurred frequently. While modifications on histone 3 predominantly appeared together, histone 4 acetylation mainly occurred either by itself or in conjunction with all three other modifications (Figure 4-4 A). The distributions of combinations were similar for the three cell types (Figure 4-4 B).

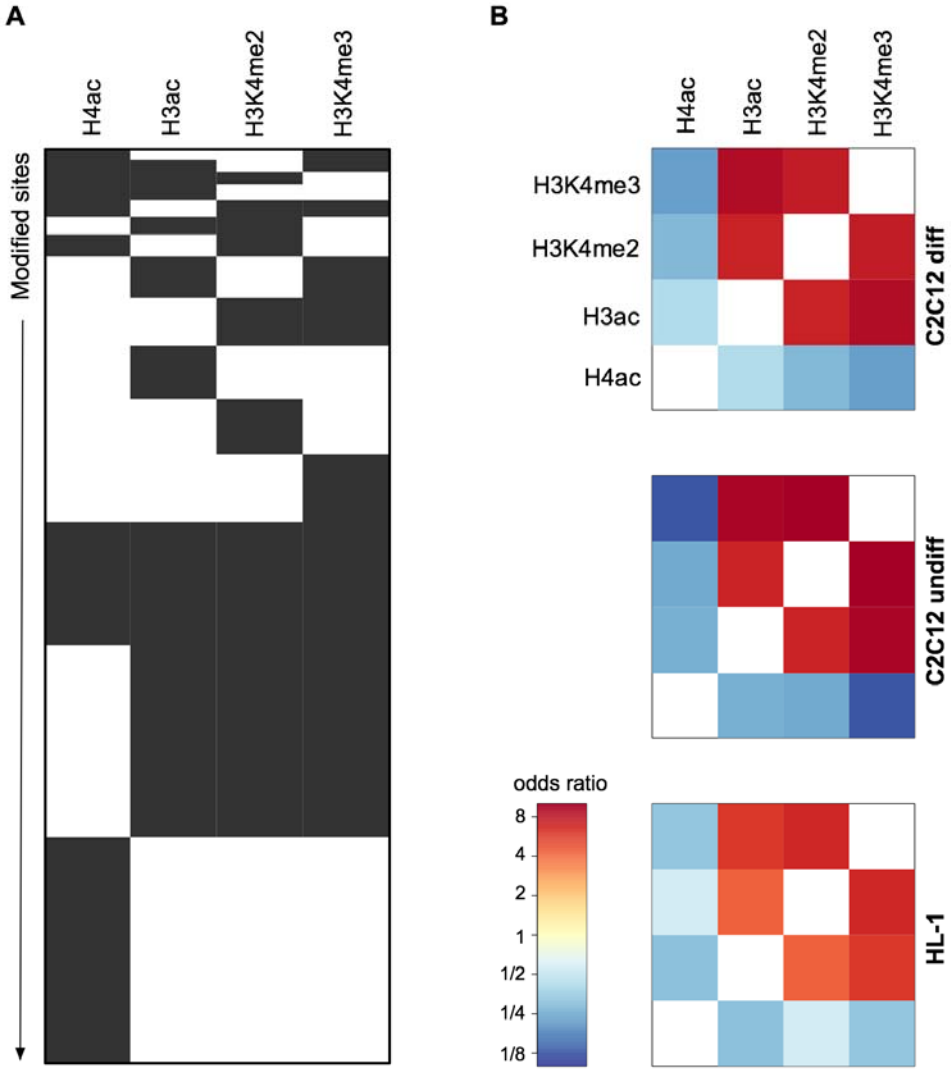


Figure 4-4. Analysis of co-occurrence of histone modifications. A) The combinatorial occurrence of histone modified sites. Each row in the heatmap corresponds to a combination of modified sites. Dark indicates presence, white absence. The height of the rows is proportional to the number of occurrences of a combination, summed over the three cell types. The most frequent cases are H4ac alone, H3acK4me2/3 and the combination of both. B) Odds ratios of pair-wise contingency tables of the occurrence of modified sites in domains. The pattern is similar for all three cell types. Red indicates positively correlated occurrence, blue corresponds to anti-correlation.

To further characterize the combinatorial occurrence of modifications, a modification code was assigned to each modified location, so that the code describes which of the 15 different possible combinations is present at that position. The sequence stretch associated with the modifications is now termed *modified domain*, or more concisely, a *domain*. Figure 4-5 illustrates this for the example of Hand2: in HL-1 cells two domains with the *modification code* H4ac-H3acK4me2/3 were found, while none was found in C2C12 undifferentiated cells. Merging histone modified sites to modified domains resulted in a total of approximately 6,000 domains per cell type.

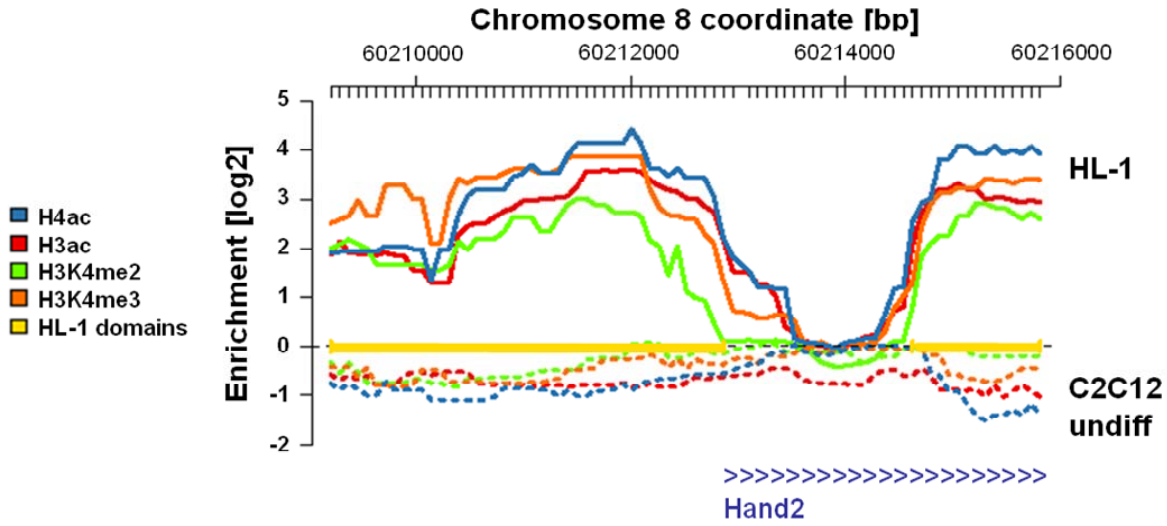


Figure 4-5. Normalized and smoothed ChIP-chip intensities around the TSS of the Hand2 gene. In C2C12 undifferentiated cells (dashed lines) no modifications are associated with the Hand2 gene. In HL-1 (solid lines) two domains with the modification code H4ac-H3acK4me2/3 were identified. Each domain consists of four modified sites.

4.2.2.4 Cell Type-Specific Domains

Next the histone modification pattern in the vicinity of each of the represented TSSs was analyzed. For this purpose, modified domains were assigned to a TSS of a gene if they were located within 5 kb upstream of the TSS or within the transcribed region. For genes with multiple TSSs, TSS and the respective transcribed region were considered separately. Consequently, one modified domain may be assigned to several TSSs.

First, the data from each of the three cell types was evaluated separately and the degree of overlap was determined. Not only was there a large number of differentially expressed transcripts, as discussed in Chapter 4.2, but also the modification patterns of about two thirds of the TSSs with modifications differed between cell types. Figure 4-6 is a Venn diagram representation of the number of TSSs associated with domains of the same modification code in the three cell types; positional identity of the domains was not required. Out of the approximately 5,000 TSSs per cell type found to be associated with domains 1,267 TSSs were marked by the same modifications in myoblasts, myotubes and cardiomyocytes. Around 1,500 TSSs showed a modification pattern that only occurred in one of the three cell types. Although only one third of the domains was the same between the three cell types, the overall results linking modification patterns to transcript levels were found to be highly comparable. Therefore, a composite data set that includes domains and transcript levels from all three cell types was used in the following sections to illustrate the results.

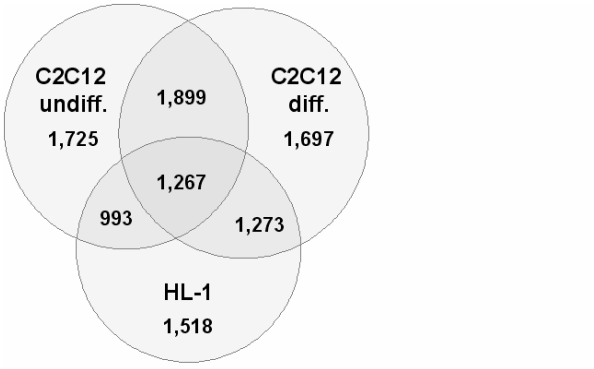


Figure 4-6. Venn diagram representing the number of TSSs marked by the same modified domains in one, two or all three cell types. All TSSs represented on the arrays were considered. In the following graphs only TSSs were considered for which the corresponding expression data was also obtained.

4.2.2.5 Association of TSSs and Modified Domains

One third of the TSSs were associated with the same modification code in all three cell types, while more than 1,500 appeared only in one cell type. It was now investigated how often TSSs are marked by particular combinations and if the frequency of combinations differed between the cell types. For this purpose, TSSs were categorized depending on the number of domains found in their direct vicinity (Figure 4-7). In general, about half of the analyzed genes were associated with modified domains. Two thirds of these genes were marked by only one modified domain and approximately one fourth by two domains. Subsequently, the TSSs associated with two domains were further classified into whether these domains had identical or two different codes; different codes were found to be far more frequent. Then the distribution of codes in each class was investigated and found to show little variation between the cell types. In the bottom line of Figure 4-7 exemplary results as obtained from C2C12 undifferentiated cells are shown.

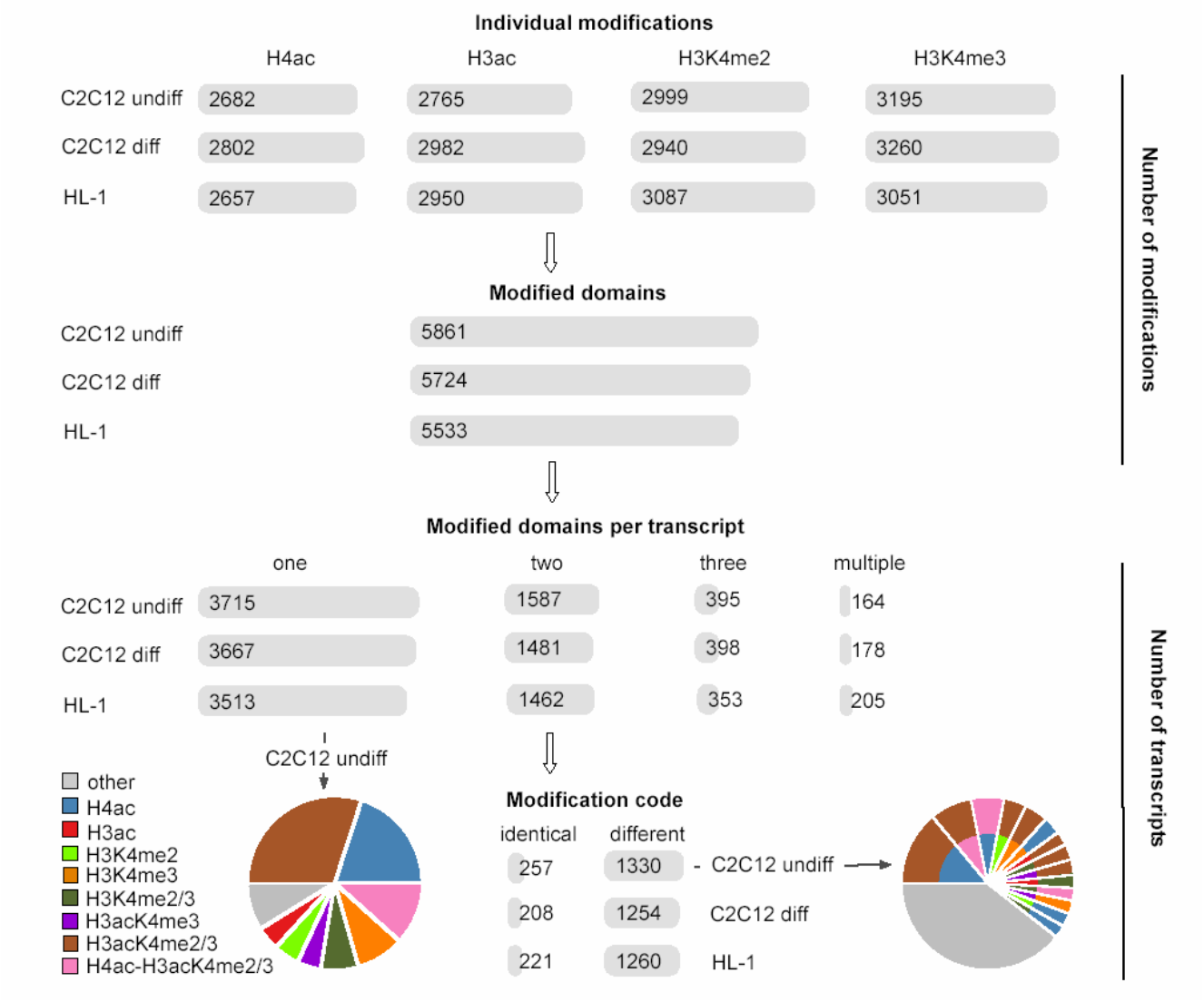


Figure 4-7. The distribution of modifications shows little variation over the cell types. The horizontal bars show the number of individual modifications (top row) and of modified domains (second row) for the three cell types. The third row shows the distribution of the number of modified domains per TSS. TSSs with one or two domains are most frequent. The pie charts show the distributions of modification codes for one-domain transcripts and for two-domain transcripts with different codes in C2C12 undifferentiated cells.

Out of the 15 possible domain codes only eight occurred frequently. In cases where one domain per TSS was identified, these were predominantly multi-code H3acK4me2/3, H4ac-H3acK4me2/3 or single-code H4ac. If two modified domains per TSS occurred, these generally had different modification codes, one of them characteristically H3acK4me2/3. Single-code H4ac was found to be typical for genes marked by two or more identical modified domains. Interestingly, although only few TSSs were associated with the same modifications in all three cell types, the number of occurrences for each modification and modification combination was similar.

4.2.2.6 Localization of Modifications Relative to TSSs

It was previously reported that the investigated modifications occur predominantly near TSSs[25,228]. Therefore, the first question was whether this is also observed in the analyzed cell lines. For this purpose, the positions of modifications relative to the TSSs were analyzed while ignoring co-occurrences, that is, based on modified sites (Figure 4-8). Enrichment of histone modifications was observed within ± 1 kb, although few modifications were found directly at the TSS. Histone 4 acetylation was predominantly localized upstream and the three modifications on histone 3 more frequently downstream of the TSS. These observations are in accordance with previously reported findings[25,228].

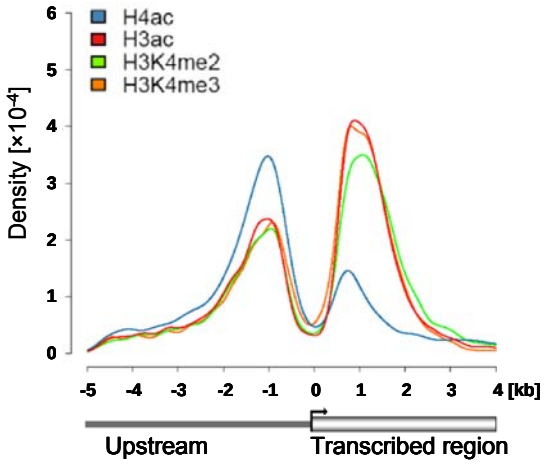


Figure 4-8. Shown are data for C2C12 undifferentiated cells. The 5 kb upstream and 4 kb downstream regions of each TSS were aligned by the TSS (x-axis). The y-axis shows the densities. Analysis is based on modified sites (i.e. without taking into account combinatorics). Modifications on histone 3 show similar distributions. H4ac occurs predominantly up- but also downstream.

Secondly, it was investigated if any of the modification combinations show a preferential localization relative to the TSSs. If particular combinations function as signaling marks, they might be overrepresented at specific genomic locations. Therefore, the analysis was repeated based on the eight most frequent modified domains. This refined the previous picture where the combinatorial nature of modifications had been ignored. It was now found that although multi-code domains containing H3K4me2 peaked close to the TSS, domains marked only by H3K4me2 were distributed throughout the transcribed region. Domains characterized by H4ac alone occurred almost exclusively upstream, while in conjunction with other modifications (H4ac-H3acK4me2me3) it was found as often up- as downstream. This analysis shows: if H4ac occurs in transcribed regions it is almost always accompanied by the three other modifications on histone 3; if H4ac occurs alone it almost always is located upstream.

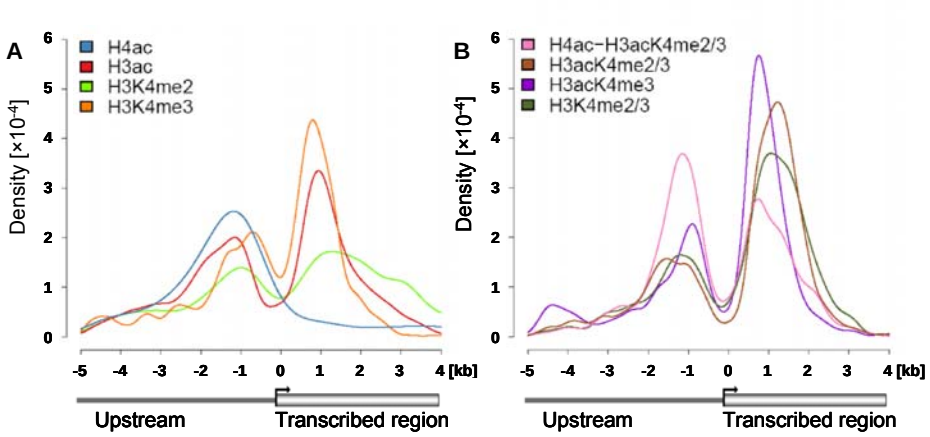


Figure 4-9. Analysis of the same data as shown in Figure 4-8, but on the basis of modified domains (i.e. considering co-occurrence of modifications). The eight most frequent domains are shown and split into two graphs for clarity. A) localization of the four single-code domains is shown. Single-code H4ac is positioned almost exclusively upstream, single-code H3K4me2 is distributed throughout the transcribed region. B) Localization of the four most frequent multi-code domains is shown. Domains containing all four modifications occur similarly often up- and downstream. Domains coding for histone 3 modifications are predominantly positioned in the transcribed region.

4.2.3 Transcript Levels and Histone Modifications Demonstrate Histone Code

The investigation of the localization of the most frequent modification combinations showed that some domains show clear positional preferences relative to the TSSs. This indicates, that particular combinations might have distinct functions. Bringing the descriptive information on the occurrence of the modifications together with the expression data was the next step in the analysis. Thereby the *histone code hypothesis* which postulates distinct combinations of modifications to result in specific read-outs, might be confirmed.

The investigation was carried out from two angles. First, transcripts were classified into expression categories and the frequency of modification combinations within these classes was compared. Second, the transcripts were classified according to their associated domains and the expression levels were investigated.

4.2.3.1 Analysis Based on Expression Categories

Transcripts were divided into four different expression categories depending on the intensities measured on the expression arrays. The cut-off values for the different classes (non-, low-, medium- and high-expressed) were determined using RT-PCR for genes of known transcriptional status (data not shown). Subsequently the occurrence of modifications within each class was investigated. First the analysis was conducted based on modified sites, that is without taking co-occurrences into account. As the transcription level classes contained a varying number of TSSs the relative percent increase of domains from one class to the next was calculated. Without taking combinations into account it was observed that all four histone modifications, in particular the two types of acetylation, coincided with elevated transcript

levels (Figure 4-10A). The number of sites relative to the number of TSSs in each expression class increased from non- to medium-expressed transcripts. The 10% most highly expressed transcripts, however, showed no further increase in modifications compared to medium expression levels.

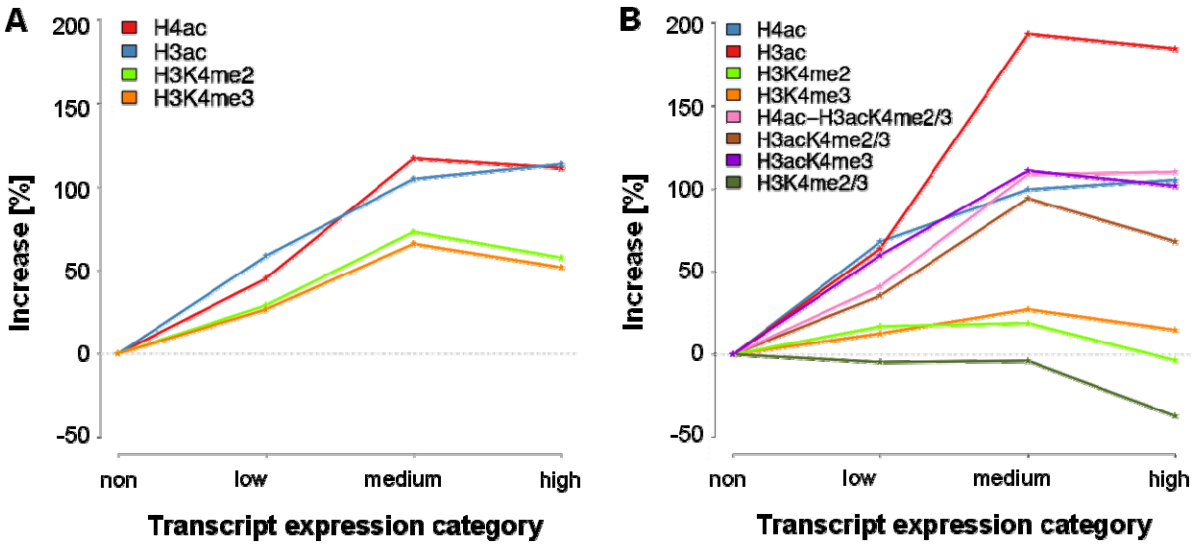


Figure 4-10. Higher expression levels of transcripts are associated with stronger histone acetylation. Transcripts were classified into four groups according to expression level: non-, low-, medium- and high-expressed (x-axis). The y-axis shows the percent increase of the frequency of modifications in each group, compared to the non-expressed group. The curves were obtained by taking the median across the three cell types. A) Analysis based on modified sites. Each individual modification increases over the expression groups. The frequencies of the two acetylation states and of the two methylation states, respectively, show similar increase. B) Analysis on the basis of domains. Shown are the eight most frequently occurring modified domains. The frequency of single-code H3ac shows strongest increase. Other acetylation containing domains behave similarly and increase by approximately 100%. Domains coding for one or both methylations show basically no change.

Next it was investigated, whether distinct combinations of modifications occur more frequently with higher expression levels. For this purpose the analysis was repeated based on modified domains. Figure 4-10A shows the eight most frequent combinations; when considering co-occurrences strikingly different results are obtained. The relative number of TSSs marked only by H3ac without additional modifications increased by 200% in the class of expressed transcripts compared to the non-expressed class. Interestingly, such a pronounced increase was not observed for such TSSs where other (activating) modifications in addition to H3ac were present. Furthermore, the proportion of domains solely containing one or both methylation modifications only slightly varied between three expression groups but was even decreased for the most highly expressed transcripts.

4.2.3.2 Analysis Based on Modification Categories

In the second analysis the TSSs were classified according to the associated modifications and the expression array intensities of the associated transcripts were analyzed. First the analysis was again performed on the basis of modified sites, that is, without taking combinations into account. TSSs associated with a domain containing e.g. H3ac and H3K4me3 was counted once in the class of TSSs associated with H3ac and then once more in the class on H3K4me3 associated TSSs. Such TSSs which were represented on the arrays by a sufficiently high number of probes, where by a modification could in theory have been detected but where nonetheless no enrichment for any modification was found were classified as non-modified.

Without considering combinations, all four modifications clearly, and to approximately the same degree, coincided with higher expression values compared to transcripts from non-modified TSSs (Figure 4-11A). To ascertain that the transcript levels in the four classes representing the four different modifications were approximately equal, pairwise *p* values were calculated. Table 4-3 gives *p* values for pairwise comparison of categories as shown in Figure 4-11A: Transcripts were categorized according to the modification status of the respective gene into five groups corresponding to the rows and columns of the table. Expression levels of the categories were compared by two-sided two-sample Wilcoxon tests and *p* values were adjusted for multiple testing using the Bonferroni procedure[251]. For each comparison, the Table 4-3 gives the sign of the difference and the *p* value: +, row category has higher levels than column category; -, row category has lower levels than column category; o, no rejection.

Table 4-3. Transcripts categorized by associated modified sites: *p* values for pairwise comparison of transcript categories as shown in Figure 4-11A.

Transcript Category	No modification	H3K4me2	H3K4me3	H4ac
H3ac	+ ($<10^{-30}$)	+ (4×10^{-8})	+ (8×10^{-7})	+ (2×10^{-3})
H4ac	+ ($<10^{-30}$)	+ (0.04)	o	
H3K4me3	+ ($<10^{-30}$)	o		
H3K4me2	+ ($<10^{-30}$)			

Subsequently the analysis was repeated on the basis of domains to investigate whether different combinations of modifications lead to distinct transcriptional outcomes. Figure 4-11B shows the results for the eight most frequent modification combinations, where a ranking of modification effects on expression levels was obtained. Calculation of pairwise

p values between the different modification classes showed that the association between different modification combinations and different transcript levels are indeed significant. Table 4-4 gives *p* values for pairwise comparison of categories as shown in Figure 4-11B: Table 4-4 is analogous to Table 4-3, but only the eight most frequent domain types are listed.

Table 4-4. Transcripts categorized by associated modified sites: *p* values for pairwise comparison of transcript categories as shown in Figure 4-11B.

Transcript Category	No modification	H3K4me2	H3K4me2/3	H3K4me3	H3ac-K4me3	H3ac-K4me2/3	H4acH3ac-K4me2/3	H4ac
H3ac	+ ($<10^{-30}$)	+ (1×10^{-16})	+ (2×10^{-19})	+ (6×10^{-8})	+ (3×10^{-2})	+ (4×10^{-5})	+ (2×10^{-2})	+ (9×10^{-6})
H4ac	+ ($<10^{-30}$)	+ (3×10^{-6})	+ (5×10^{-8})	o	o	o	o	
H4ac-H3ac-K4me2/3	+ ($<10^{-30}$)	+ (7×10^{-11})	+ (1×10^{-13})	+ (7×10^{-3})	o	o		
H3ac-K4me2/3	+ ($<10^{-30}$)	+ (8×10^{-9})	+ (2×10^{-11})	o	o			
H3ac-K4me3	+ (2×10^{-16})	+ (2×10^{-6})	+ (8×10^{-8})	o				
H3K4me3	+ (6×10^{-12})	+ (2×10^{-4})	+ (4×10^{-5})					
H3K4me2/3	o	o						
H3K4me2	o							

Single-code H3K4me2 and multi-code H3K4me2/3 modified transcripts' mean expression levels were not significantly higher than non-modified transcripts, whereas H3K4me3 alone was associated with only slightly higher expressed transcripts (mean fold change: 1.22, $p = 6 \times 10^{-12}$). TSSs associated with single-code H3ac coincided with the highest expression levels (mean fold change: 1.53, $p = 1 \times 10^{-30}$), while TSSs marked by other modifications in addition to H3ac showed lower expression ($p \leq 3 \times 10^{-2}$). However, none of the single code domains were repressive.

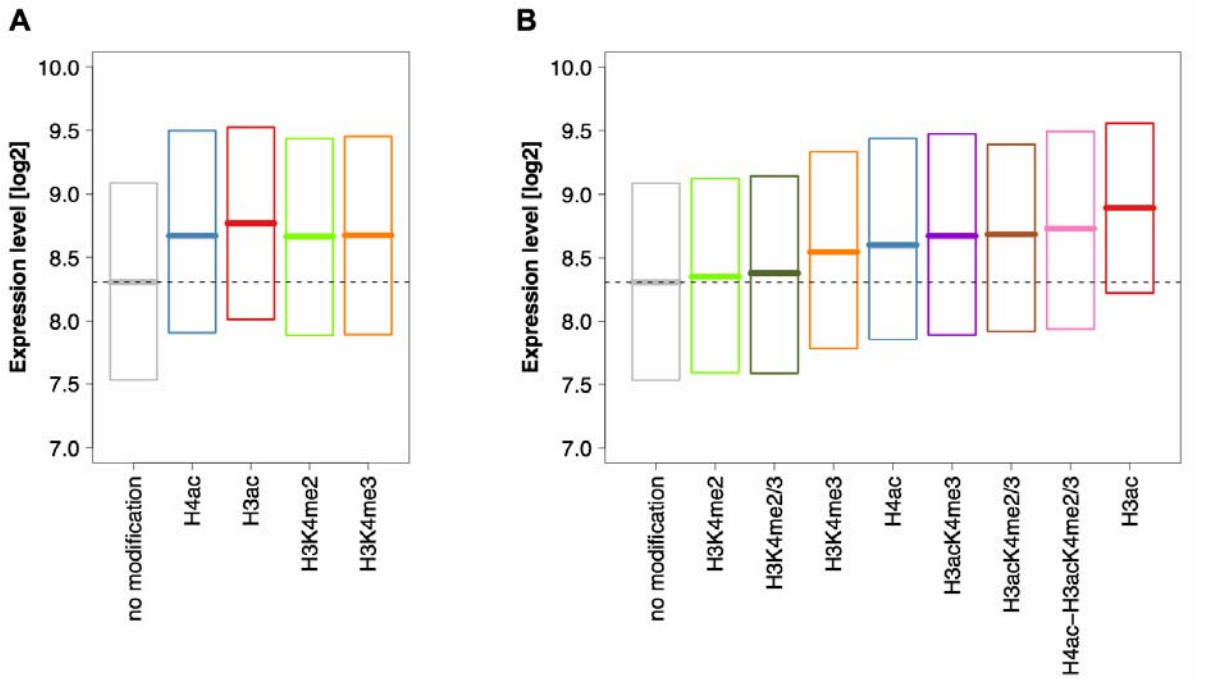


Figure 4-11. The combinatorial nature of the relationship between histone modifications and expression. Transcripts were grouped by their associated histone modifications. The expression level distribution in each group is represented by boxplots. A) Analysis based on modified sites. All four individual modifications are similarly associated with elevated expression levels compared to the no modification group ($p < 1 \times 10^{-30}$, Table 4-3). B) Analysis on the basis of domains. The highest levels are seen with H3ac alone, H3ac combined with the other three modifications shows comparatively lower levels ($p \leq 3 \times 10^{-2}$, Table 4-4). The levels with single-code H3K4me3 are comparatively low and even reduced in combination with H3K4me2. H3K4me2 associated transcripts are not significantly different from the no modification group.

In case of H3ac the co-occurrence of additional modifications was associated with lower expression levels compared to TSSs were H3ac occurred alone. Therefore a linear model (Model 3-1) was fitted to the data to investigate whether generally, the sum of individual modifications results in the values associated with their combinations or not (Table 4-5). Certain interaction terms between modifications were found to be significantly different from zero and the co-occurrence of modifications generally was associated with lower expression levels than expected from the sum of the individual modifications effects. For example, interaction between H4ac and H3K4me3 gave an estimate of -0.33 with a corrected *p* value of 3.1×10^{-3}, interaction between H3ac and H3K4me3 results in an estimate of -0.35 with a *p* value of 1.7×10^{-4}. These findings demonstrate the combinatorial, non-additive effect of histone modifications.

Table 4-5. Linear model and obtained coefficients. The table specifies for each predictor variable the coefficient estimate, its standard error and *p* value for the null hypothesis that the coefficient is equal to 0.

	Estimate	Std.Error	*t* value	*p* value
Intercept	4.26	0.06	76.24	$< 2\times10^{-16}$
H3ac	**0.58**	**0.05**	**11.07**	**$< 2\times10^{-16}$**
H4ac	**0.38**	**0.03**	**13.28**	**$< 2\times10^{-16}$**
H3K4me2	0.06	0.05	1.19	1
H3K4me3	**0.39**	**0.04**	**9.01**	**$< 2\times10^{-16}$**
GC	8.22	0.02	75.22	$< 2\times10^{-16}$
cell.type.C2C12U	-0.02	0.02	-0.99	1
cell.type.C2C12D	0	0.02	-0.25	1
H3ac:H4ac	-0.23	0.1	-2.34	0.37
H4ac:H3K4me2	0.08	0.09	0.93	0.35
H4ac:H3K4me3	**-0.33**	**0.09**	**-4.44**	**3.1×10^{-3}**
H3ac:H3K4me2	-0.23	0.11	-2.09	0.7
H3ac:H3K4me3	**-0.37**	**0.11**	**-3.11**	**1.7×10^{-4}**
H3K4me2:H3K4me3	**-0.33**	**0.08**	**-4.1**	**7.9×10^{-4}**
H4ac:H3K4me2:H3K4me3	0.05	0.15	0.31	1
H3ac:H4ac:H3K4me2	-0.11	0.18	-0.63	1
H3ac:H4ac:H3K4me3	0.24	0.16	2.33	0.38
H3ac:H3K4me2:H3K4me3	**0.46**	**0.14**	**3.4**	**1.3×10^{-2}**
H3ac:H4ac:H3K4me2:H3K4me3	-0.08	0.23	-0.33	1

4.2.4 Investigation of Skeletal Muscle Differentiation

To complete the static picture gained from the analysis of independent cell types, the dynamic of modification changes during differentiation and their correlation to expression changes was examined. For this purpose the occurrence of particular histone modifications in the earlier stage of myoblasts and in the later stage of differentiated myotubes was compared. This allowed insight into the placement and removal of these marks during differentiation. It was then investigated whether a change of modification in the proximity of the TSS was related to a change of the expression value of the corresponding transcript.

4.2.4.1 Modification Changes during Differentiation

Based on the definition of domains in undifferentiated and differentiated C2C12 cells as described in section 3.3.1 and their assignment to TSSs the change of histone modifications during differentiation was analyzed. A domain was classified as being unchanged if it retained the same composition of modifications (code) and positional identity on the genome (± 500 bp). For those domains where a change was observed, it was then investigated whether this consisted of the loss or the gain of one or more particular modifications at that genomic position. The domains were then assigned to the TSSs as described previously, thereby facilitating a correlation analysis of modification and expression changes.

Although the absolute number of modified domains found in myoblasts and in myotubes was fairly constant (Figure 4-6), a high number of changes occurred within the domains and some domains disappeared completely while others were newly formed. Overall, a substantial number of 3,498 TSSs was associated with some form of modification change, corresponding to approximately one third of all TSSs represented on the array. Strikingly, the majority of modification conversions involved H4ac both as a singly occurring modification and in the context of multi-code domains: 727 TSSs which showed only H4ac in myoblasts were associated with no modification at all in myotubes. For a similar number of 854 TSSs no modification was found in myoblasts while in myotubes domains consisting of only H4ac were identified.

In regard to changes within domains the major conversions were between domains coding for the three modifications of histone 3 (H3acK4me2/3) and domains containing H4acH3acK4me2/3, again involving loss and gain of H4ac. H3acK4me2/3 domains gained additional histone 4 acetylation in 299 cases, while 134 domains showing all four modifications lost H4ac while retaining H3acK4me2/3. In comparison, only 25 domains containing H3acK4me2/3 were completely lost and only 29 were gained. All other changes occurred even less frequently.

The modification changes mainly involved H4ac indicating that this modification plays a major role in the process of differentiation. To identify effectors involved in the conversion of this modification the sequences associated with the loss and gain of H4ac were examined for the binding matrices of transcription factors known to play a role in skeletal muscle differentiation. 854 TSSs were associated with a gain of H4ac. The surrounding sequences were associated with the presence of Mef2 binding sites (Group Specificity Score 4.5×10^{-4}). Mef2 is a histone acetyl transferase recruiting factor and is essential for skeletal muscle differentiation[276].

4.2.4.2 Expression Levels and Modification Changes

For transcripts where a change in histone modification near the TSSs was found, the fold changes between their expression level in myoblasts and in myotubes was calculated. The median fold change was determined both for all transcripts showing modification change and for each type of modification conversion separately. The average fold changes were not significantly different from zero in any case, as shown in Figure 4-12 for all changes and the three most frequent conversions. Furthermore, only few differentially expressed transcripts were represented in the group of genes associated with modification conversions. For

example, among the 854 transcripts associated with a gain of H4ac, only 29 showed significantly higher expression levels (Table 4-6).

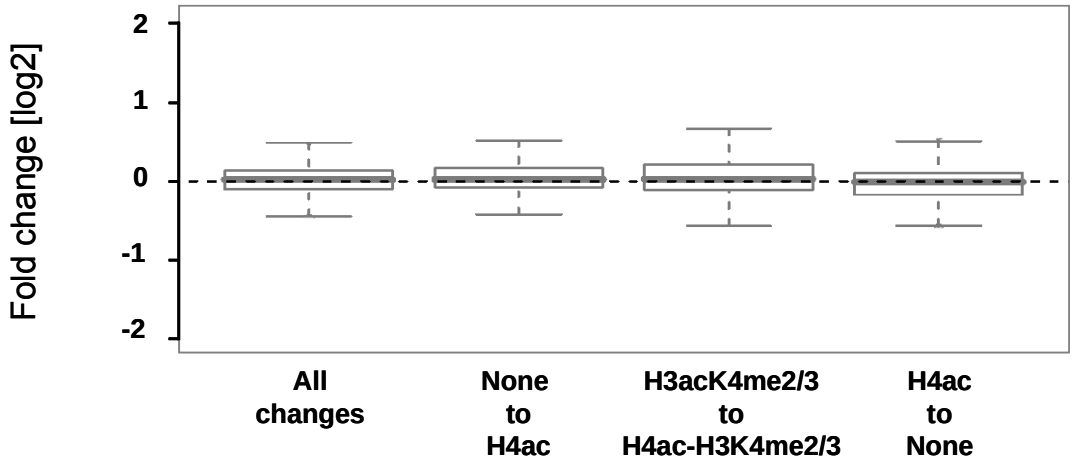


Figure 4-12. The figure shows a box plot representation of the fold changes of all changes and the three most frequently occurring modification conversions.

Table 4-6. Contingency table for upregulation in differentiation (rows) versus modification gains (columns).

Modification change assigned to TSS / Number of Genes	No differential expression	Significantly up regulated
H4ac no change	10,025	97
H4ac gain	825	29
H3ac no change	10,267	114
H3ac gain	583	12
H3K4me2 no change	10,259	109
H3K4me2 gain	591	17
H3K4me3 no change	10,281	112
H3K4me3 gain	569	14

4.2.4.3 Modification Changes of Differentially Expressed Genes

In comparison to the 3,498 TSSs associated with some form of modification change a relatively low number of 299 transcripts were significantly differentially expressed between myoblasts and myotubes. The gene ontology annotations overrepresented for these transcripts were in good agreement with published results[277] as show in chapter 4.2.

After examining the influence of modification changes near the TSSs on the expression of the corresponding transcripts, the opposite analysis was carried out. It was now investigated if the transcripts which showed significant differential expression were associated with a particular modification change.

Among the total of 126 upregulated transcripts, gain of H4ac was seen significantly more often than among not differentially expressed transcripts ($p = 1 \times 10^{-7}$), as could be shown by a logistic regression model (Model 3-2, Table 4-7). Other modification changes were not significant. For the downregulated transcripts, significant preferences for modification changes were observed (Model 3-3, Table 4-8). If histone modifications were a consequence of transcription, or alternatively, a sufficient driving force, it would be expected that modification changes are highly associated with changes of expression levels.

Table 4-7. Coefficients of model upregulation (Model 3-2).

	Estimate	Std.Error	z value	*p* value
Intercept	-4.73	0.11	-43.72	$<2\times10^{-16}$
H3ac.gain	0.12	0.33	0.35	1
H4ac.gain	**1.18**	**0.22**	**5.55**	**1.4×10^{-7}**
H3K4me2.gain	0.73	0.3	2.41	0.08
H3K4me3.gain	0.35	0.33	1.08	1

Table 4-8. Coefficients of model downregulation (Model 3-3).

	Estimate	Std.Error	z value	*p* value
Intercept	-4.23	0.09	-49.31	$<2\times10^{-16}$
H3ac.loss	0.49	0.3	1.67	0.48
H4ac.loss	0.18	0.27	0.69	1
H3K4me2.loss	-0.04	0.31	-0.14	1
H3K4me3.loss	0.64	0.28	2.28	0.11

4.2.5 Interim Summary: Combinatorics of Histone Modifications

Gene expression levels and the localization of four histone modifications H3ac, H4ac, H3K4me2, and H3K4me3 were characterized in myoblasts, myotubes and cardiomyocytes. Investigation of the association between the occurrence of the modifications and expression levels revealed that different modification combinations are associated with distinct transcriptional read-outs. This observation supports the *histone code* hypothesis. Comparison of expression levels and histone modifications in undifferentiated and differentiated C2C12 cells indicate that histone modifications are prior to transcription and may have functions as signaling marks.

4.3 A Transcriptional Regulatory Network of Gata4, Mef2a, Nkx2.5, and Srf

4.3.1 Detection of Transcription Factors in HL-1 Cells

4.3.1.1 Specificity of Antibodies Directed against Transcription Factors

In ChIP analysis specificity of the employed antibodies is a critical parameter to ensure enrichment of binding sites for only one particular transcription factor of interest. Western blot analysis with HL-1 whole cell lysate demonstrated that the antibodies used in this study give only one band corresponding to the expected protein sizes of 50 kDa in case of Gata4, 40 kDa in case of Nkx2.5 and 67 kDa in case of Srf (Figure 4-13). Western blots using an antibody directed against Mef2a showed two bands: one of ≈ 54 kDa corresponding to the molecular mass of Mef2a and a second band of ≈ 64 kDa which probably corresponds to a precursor variant of Mef2a[278]. According to the manufacturer's specifications a minor cross-reactivity with Mef2c or Mef2d may occur, however, additional bands were not detected.

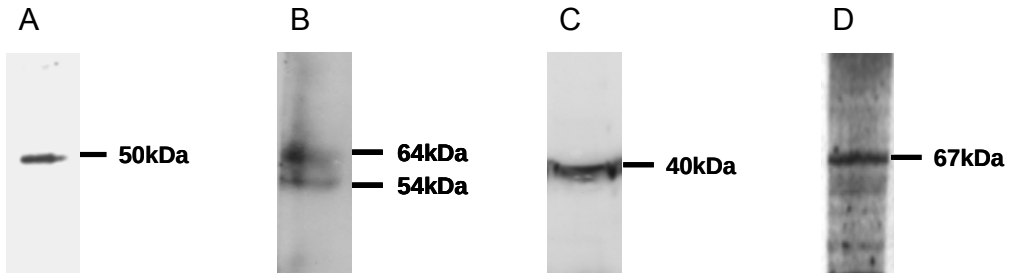


Figure 4-13. Specificity of antibodies used in ChIP experiments demonstrated by Western Blot analysis of HL-1 whole cell lysates. A) Gata4 B) Mef2a C) Nkx2.5 and D) Srf.

4.3.1.2 Cellular Localization of Transcription Factors

Nucleocytoplasmic shuttling is a common mechanism in transcriptional regulation influencing the concentration of the proteins in the nucleus and thereby the effect of transcription factors[279]. Therefore, the localization of the transcription factors within the HL-1 cells was verified using indirect immunofluorescence (Figure 4-14). For all four investigated TFs a clear signal in the nucleus was visible: in case of Mef2a and Srf a weak additional signal in the cytoplasm was observed (Figure 4-14 B, D). Moreover, in each case nearly all cells seem to contain the TF of interest, indicating a homogenous cell population.

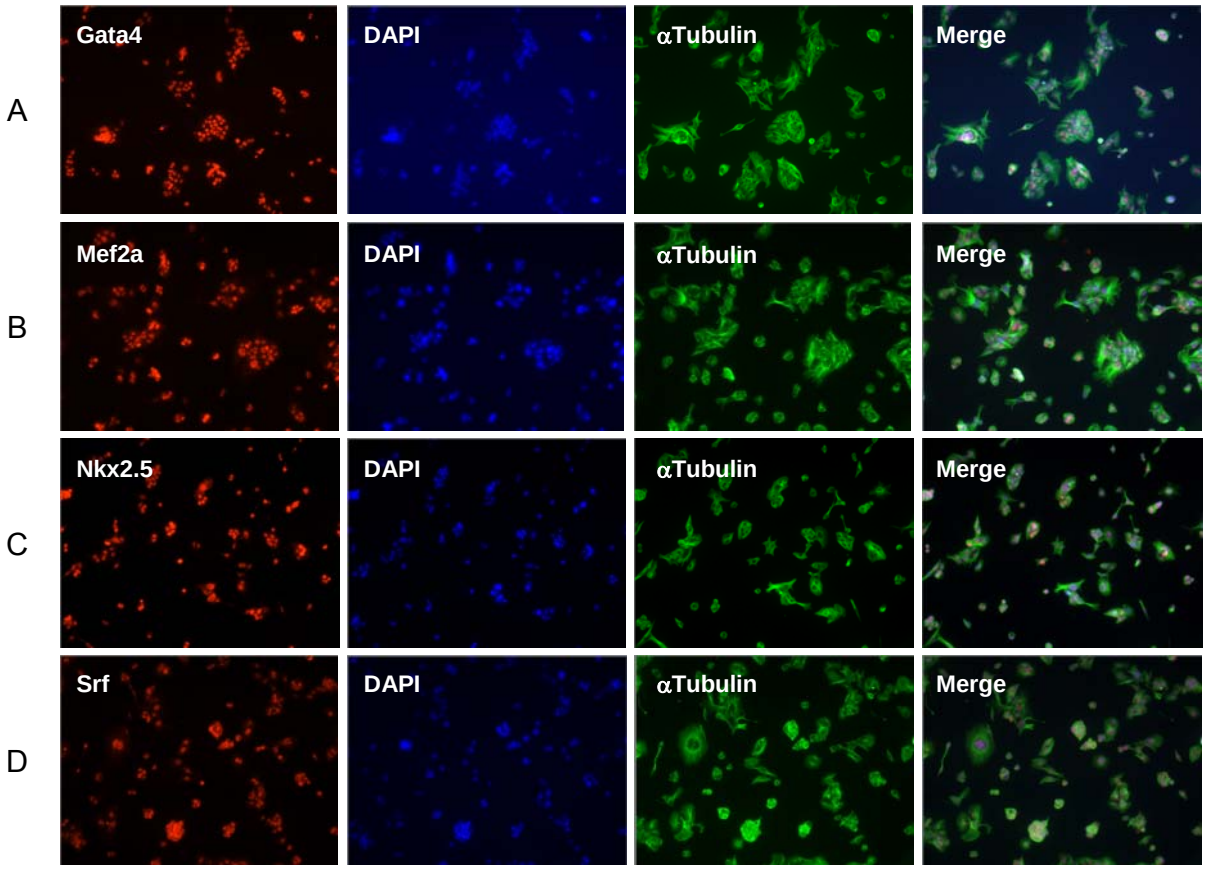


Figure 4-14. Indirect immunofluorescence of HL-1 cells showing localization of TFs (*Alexa Fluor® 594, red*) in the nucleus: A) Gata4, B) Mef2a, C) Nkx2.5 and D) Srf. Nuclei were counterstained using 4',6-diamidino-2-phenylindole (*DAPI; blue*) and an antibody directed against α-tubulin was used as cytoplasmic marker (*Alexa Fluor® 488, green*).

4.3.2 Characterization of Transcription Factor Binding

4.3.2.1 Identification of Transcription Factor Binding Sites (TFBSs)

ChIP-chip analyses of TFBSs led to the identification of several hundred binding sites per TF: Gata4 (447), Mef2a (999), Nkx2.5 (383), and Srf (1,335). The TFBSs were assigned to genes and the distance to the TSSs were investigated (Figure 4-15). The absolute number of TFBSs varied considerably, but their positions relative to TSSs were similar (Figure 4-15). Only ≈ 35% of TFBSs were mapped to the proximal promoter regions (defined as the first 500 bp of the transcribed region and the 4.5 kb upstream of the TSSs) where the vast majority of TFBS have been identified so far. Approximately half the TFBS were found within transcribed regions and 10% of TFBSs were located further than 4.5 kb upstream of any TSS.

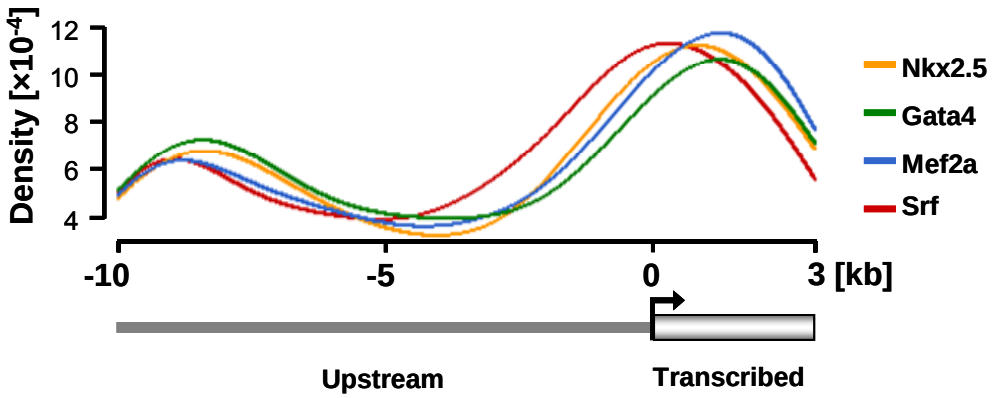


Figure 4-15. Positional distribution of TFBSs relative to the TSSs. The 10 kb upstream and 3 kb downstream regions of each TSS were aligned by the TSS (x-axis). The y-axis shows the densities. The distribution of TFBSs relative to TSSs is similar for the four TFs.

The TFBSs were assigned to genes represented on the array if located within 10 kb upstream or in the transcribed region, as these could potentially be regulated. If TFBSs fell into the regions of several TSSs they were assigned multiple times. Based on this definition 468 Gata4, 1,655 Mef2a, 392 Nkx2.5, and 1,509 Srf target genes were identified. Comparison to published data showed that these include known target genes (Table 4-9). For Gata4 three, for Mef2a ten, for Nkx2.5 two, and for Srf eighteen target genes were confirmed by ChIP-chip analysis. For some of these additional qPCR confirmations were carried out and positive and negative targets could be clearly distinguished (Supplementary Figure 4). Genes previously known to be dysregulated in mutants/knockouts of the respective transcription factor could now be shown to be direct targets. For example, Karamboulas *et al.*[280] described that in P19 cells with disrupted activity of the four Mef2 isoforms (Mef2a to d) the cardiomyocyte development was impaired and expression levels of *Gata4* and *Nkx2.5* were

decreased, suggesting these TFs to be downstream of one of the Mef2 isoforms. The ChIP-chip results now show, that *Gata4* and *Nkx2.5* are direct targets of Mef2a.

Table 4-9. Genes previously described to be regulated by Gata4, Mef2a, Nkx2.5, or Srf confirmed by ChIP-chip analysis. Direct: Binding is described in the given publication, indirect: Target is described to be dysregulated in mutant/knockout of the respective TF.

TF	Target MGI Symbol	Target Gene ID	Genomic Position of TFBS			Lit.	Evidence in Lit.
			Chr.	Start [bp]	End [bp]		
Gata4	Bcl2	ENSMUSG00000057329	1	108539274	108539370	[281]	direct
Gata4	Ctgf	ENSMUSG00000019997	10	24283758	24284260	[282]	indirect
Gata4	Edn1	ENSMUSG00000021367	13	42313374	42313574	[283]	direct
Gata4	Nkx2-5	ENSMUSG00000015579	17	26571092	26571592	[284]	direct
Gata4	Pde1c	ENSMUSG00000004347	6	56289948	56290248	[282]	indirect
Gata4	Tgfb2	ENSMUSG00000039239	1	188364072	188364772	[282]	indirect
			1	188404916	188405016		
Mef2a	Actc1	ENSMUSG00000068614	2	113741448	113742050	[280]	indirect
			2	113742322	113743722		
			2	113744868	113746772		
Mef2a	Cited2	ENSMUSG00000039910	10	17414581	17415173	[285]	direct
Mef2a	Csrp3	ENSMUSG00000030470	7	48708580	48708676	[285]	direct
			7	48708880	48709780		
			7	48714466	48715066		
			7	48715896	48716802		
Mef2a	Cyr61	ENSMUSG00000028195	3	145588398	145589898	[285]	direct
Mef2a	Gata4	ENSMUSG00000021944	14	62196878	62197280	[280]	indirect
			14	62197682	62198778		
			14	62202266	62203356		
			14	62205726	62205826		
Mef2a	Hspb1	ENSMUSG00000004951	5	136162658	136162758	[285]	direct
Mef2a	Mef2c	ENSMUSG00000005583	13	84002684	84003770	[285]	direct
Mef2a	Mid1ip1	ENSMUSG00000008035	X	9872412	9873208	[286]	direct
Mef2a	Nkx2-5	ENSMUSG00000015579	17	26570992	26571998	[280]	indirect
			17	26573654	26573754		
			17	26578208	26578514		
Mef2a	Nppa	ENSMUSG00000041616	4	146843770	146843868	[125,287]	direct
Mef2a	Nr4a1	ENSMUSG00000023034	15	101093222	101093512	[285]	direct
			15	101093718	101093812		
Mef2a	Smyd1	ENSMUSG00000055027	6	71192530	71192730	[285]	direct
Mef2a	Tnnc2	ENSMUSG00000017300	2	164469814	164469910	[285]	direct
Nkx2.5	Ankrd1	ENSMUSG00000024803	19	36181724	36183022	[288]	direct
			19	36183914	36184920		
			19	36194572	36194972		
Nkx2.5	Myocd	ENSMUSG00000020542	11	65021066	65021866	[289]	direct
Nkx2.5	Nr2f2	ENSMUSG00000030551	7	70236595	70236694	[290]	indirect
Nkx2.5	Smpx	ENSMUSG00000041476	X	153041650	153042152	[291]	indirect
			X	153042654	153043652		
			X	153044149	153044650		
Nkx2.5	Ttn	ENSMUSG00000055002	2	76783640	76784140	[292]	indirect
			2	76784339	76784444		
Srf	Acta2	ENSMUSG00000035783	19	34319616	34322805	[293]	direct
			19	34329114	34329216		
			19	34329412	34329710		

			19	34330526	34331434		
			5	143169596	143170588		
Srf	Actb	ENSMUSG00000029580	5	143172086	143174284	294	direct
			5	143176816	143176910		
			2	113742322	113742426		
			2	113742726	113742822		
Srf	Actc1	ENSMUSG00000068614	2	113743026	113743422	295	direct
			2	113744868	113745572		
			2	113746475	113747070		
Srf	Bcl2	ENSMUSG00000057329	1	108539274	108539370	296	direct
			10	110323197	110323494		
Srf	Csrp2	ENSMUSG00000020186	10	110326130	110326332	96	direct
			10	110326736	110326836		
			10	110334876	110334968		
Srf	Dmd	ENSMUSG00000045103	X	79208666	79208966	297,298	direct
Srf	Egr2	ENSMUSG00000037868	10	66933446	66933942	294	direct
Srf	Fos	ENSMUSG00000021250	12	86357142	86357245	299	direct
			12	86362778	86363172		
			14	62197076	62197178		
Srf	Gata4	ENSMUSG00000021944	14	62198174	62198468	300	direct
			14	62202758	62203166		
Srf	Junb	ENSMUSG00000052837	8	87869232	87869527	301	direct
Srf	Myh6	ENSMUSG00000040752	14	53919204	53919498	300	direct
			17	26569398	26569698		
Srf	Nkx2-5	ENSMUSG00000015579	17	26570692	26571795	97	direct
			17	26573654	26573754		
			17	26578312	26578714		
Srf	Nr4a1	ENSMUSG00000023034	15	101093412	101093512	300	direct
			5	58002436	58002532		
Srf	Pcdh7	ENSMUSG00000029108	5	58009066	58009766	300	direct
			5	58010078	58010872		
			5	58012680	58012776		
Srf	Sdc2	ENSMUSG00000022261	15	32866698	32867104	96	direct
			15	32867600	32867700		
Srf	Srf	ENSMUSG00000015605	17	46016998	46018298	96,171	direct
Srf	Tpm1	ENSMUSG00000032366	9	66845652	66845856	302	direct
			9	66849056	66849150		
Srf	Tpm2	ENSMUSG00000028464	4	43543394	43543694	303	direct
			4	43552542	43552744		

4.3.2.2 Defining the *cis* Elements Mediating Transcription Factor Binding

The ChIP-chip method provides high-resolution mapping of binding sites with an average length of 150 bp, increasing the likelihood of finding motifs using *de novo* motif discovery algorithms such as Weeder[304]. The predominant motifs found in the computational search of each of the TF data sets is in agreement with the previously described binding motifs as listed in TRANSFAC[260]. In case of Gata4 the GATA motif (V$GATA_Q6)[305] was returned as core for two new refined binding motifs. The motif derived for Mef2a was highly similar to a motif previously determined by **S**ystematic **E**volution of **L**igands by **EX**ponential Enrichment (SELEX[306], V$MEF2_02[307]). For Nkx2.5 two motifs were identified, highly similar to the motifs[141] previously determined using SELEX (V$NKX25_01 and V$NKX25_02). The consensus binding site for Srf, often called the CArG box, is well established from a large number of biochemical and mutagenesis experiments: CC(A/T)TATA(A/T)GG[308]. However, this motif could not be retrieved from the sequences underlying the Srf binding sites. The motif retrieved in this study had the form $CG(A/T)_nCG$ with n between two and five.

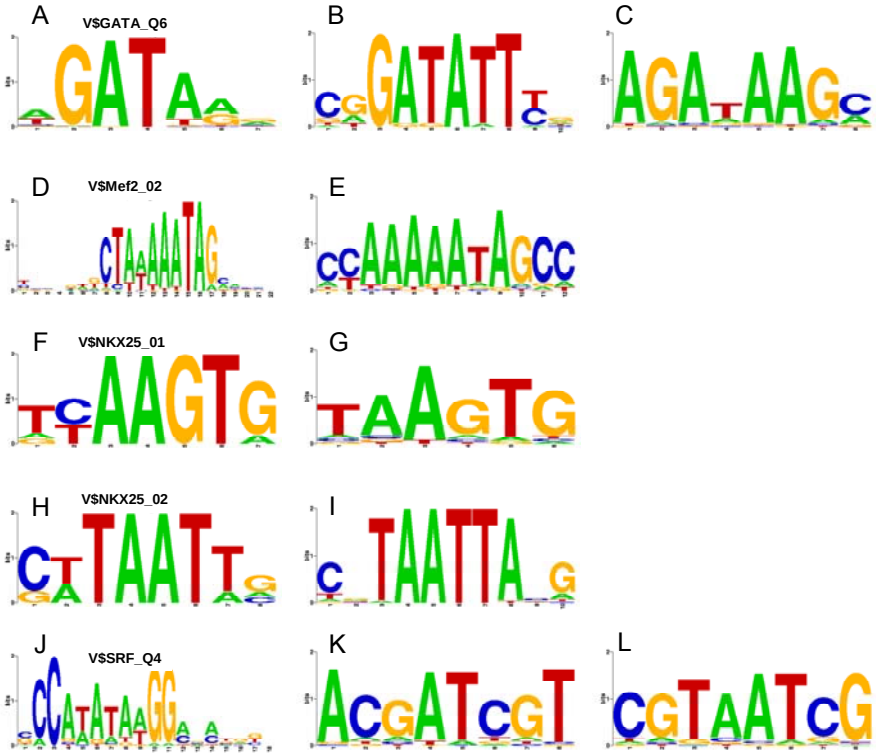


Figure 4-16. Comparison of previously described TF binding motifs for Gata4 (A), Mef2a (D), Nkx2.5 (F and H) and Srf (J), and motifs as computed using Weeder. For Gata4 two refined motifs were found (B and C), in case of Mef2a the motif could be confirmed (E). For Nkx2.5 two matrices were previously described and were retrieved (G and I), while in the case of Srf the known CArG box could not be found while a similar novel motif was identified (K and L).

4.3.2.3 Conservation of TFBSs is Low

Next the sequences underlying the transcription factor binding sites were investigated in more detail. It was analyzed to which extent the TFBSs contain the TRANSFAC[260] motifs and how highly they are conserved (Figure 4-10). First, the sequences of the binding sites were investigated regarding how frequently the respective TF motifs occur. In total the number of sequences matching to the motifs were: Gata4: 1,467, Mef2: 3,372, Nkx2.5: 806, and Srf: 525. For the first three TFs these numbers are higher than the number of TFBSs. It was then analyzed how often a TFBS contains at least one motif. The number of TFBSs where the motif was found at least one time was Gata4 86%, Mef2a 86%, Nkx2.5 85%, and Srf 13%, respectively. Within the binding sites of Gata4, Mef2, and Nkx2.5 the motifs were generally found more than once suggesting frequent multiple binding of the respective TF at a closer proximity than the resolution of the arrays (Gata4 82%, Mef2 69%, Nkx2.5 64%, and Srf 7%).

To identify to which extent the TFBSs are conserved it was investigated how often binding sites occur in regions showing conservation between eighteen species based on the PhastCons elements[247] as supplied by the UCSC Genome Browser[309]. Notably, for Gata4, Mef2a, and Nkx2.5 only 27% of the *bona fide* binding sites were overlapping conserved elements in whole genome vertebrate alignments. In case of Srf only 4% were identified within PhastCons elements.

Since TFBS motifs are short, typically 10-20 bp, they may not necessarily be located in a conserved region as detected by standard algorithms. Therefore, it was analyzed whether the TFBSs are conserved between man and mouse. Less than 15% of the sequences where TF binding motifs were found showed exact conservation (Gata4 9%, Mef2 4%, Nkx2.5 14%, and Srf 12%).

Table 4-10. Number of TF binding motifs and conservation.

	Number of TFBS			
	Gata4	**Mef2a**	**Nkx2.5**	**Srf**
Total number of TFBS	447	999	383	1,335
Total number of TRANSFAC motif matches	1,467	3,372	806	525
TFBS containing at least one TRANSFAC motif	421	858	323	169
TFBS containing TRANSFAC motif multiple times	366	687	245	89
TFBS located in PhastCons conserved regions	122	267	103	51
Man-mouse conserved TRANSFAC motifs	139	148	111	65

4.3.2.4 Co-binding and Co-occurrence of TFs

With the exception of Mef2a and Srf pairwise physical interaction has been described for the investigated TFs[93,139]. Nevertheless, it is unknown how frequently this co-binding occurs to guide the regulation in the heart. To answer this question it was investigated how often two or more ChIP enriched loci are observed within a 500 bp window (Table 4-11). For two different TFs this situation frequently occurred (e.g. Srf and Gata4 162 times, Table 4-11); multiple binding sites for the same TF within 500 bp were comparatively rare (e.g. Srf and Srf 50 times, Gata4 and Gata4 together 22 times, Table 4-11). However, it is likely that many instances of multiple binding of one TF are only detected as one enriched locus due to the limited array resolution. This observation is confirmed by the fact that within one given enriched locus a motif of one TF could be found multiple times (e.g. out of 447 Gata4 binding sites 366 contained the Gata4 motif more than once; compare previous section 4.3.2.3).

Table 4-11. Co-binding of TFs within a 500 bp window.

	Srf	Nkx2.5	Mef2	Gata4
Gata4	162	163	232	22
Mef2	291	226	21	
Nkx2.5	151	11		
Srf	50			

Direct interaction between TFs is not necessary for coregulation of a target gene. Therefore, it was next investigated how frequently the different TFs are assigned to the same transcript, irrespective of the length of the intermediate sequence (Figure 4-17). The results show that genes are frequently bound by more than one TF and all possible combinations occur (Figure 4-17 A). Binding of all four TFs was observed for 129 genes. Gata4 and Nkx2.5 had the lowest number of targets (Gata4 468, Nkx2.5 392) but were observed co-binding to 204 genes and their occurrence is therefore highly correlated (Figure 4-17 B). Although Mef2a and Srf co-occur at a similar number of genes (291) they each have a much higher number of binding sites (Mef2a 968, Srf 1,509).

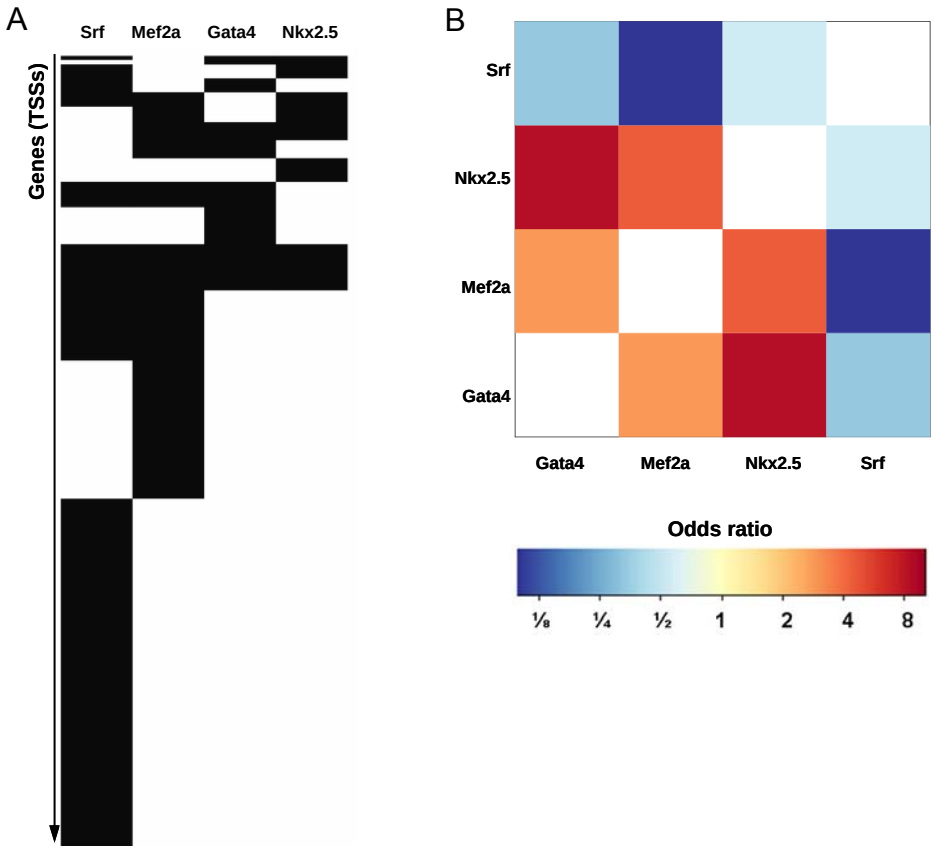


Figure 4-17. Analysis of co-occurrence of Gata4, Mef2a, Nkx2.5, and Srf at a single gene. (A) The combinatorial occurrence of TF binding sites. Each row in the heatmap corresponds to a combination of TFBSs. Dark indicates presence, white absence. The height of the rows is proportional to the number of occurrences of a combination. (B) Odds ratios of pair-wise contingency tables of the occurrence of TFBSs at one gene. Red indicates positively correlated occurrence, blue corresponds to anti-correlation.

After identification of direct targets by ChIP-chip a network was constructed (Figure 4-18). The resultant network is highly complex but can give a first overview of the number of direct targets and the number of co-regulated genes. Gene ontology (GO) terms 'biological function' were assigned to the genes. For clarity the assigned GO terms were restricted to the first level below 'biological function' resulting in 24 different GOs. The direct targets were depicted as nodes connected by edges to the TFs by which they are bound. The GO terms assigned to the targets are color coded within the nodes. Nodes of genes to which to several GOs were assigned have as many sections as GOs coded by the respective color.

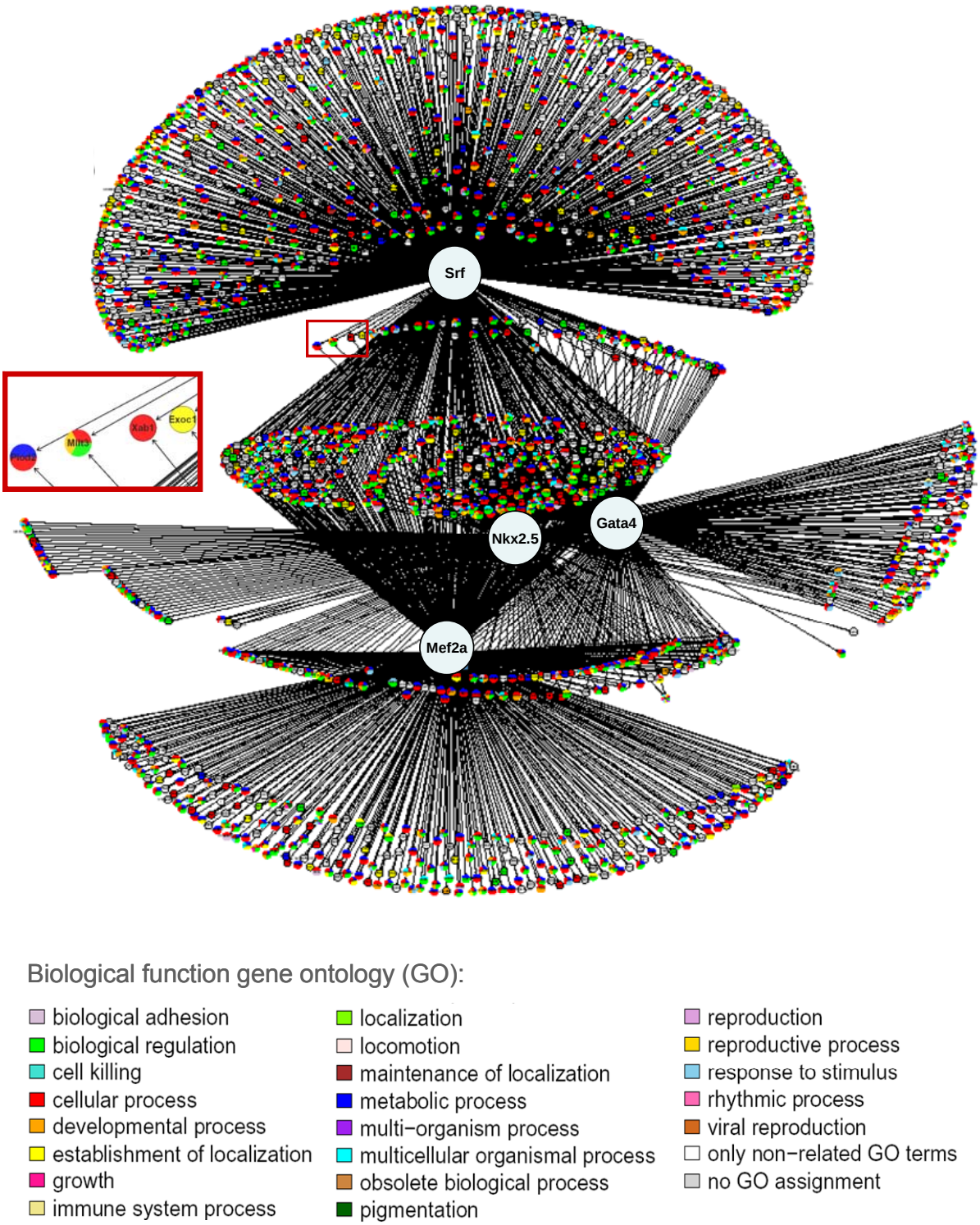


Figure 4-18. Network illustrating the direct targets of Gata4, Mef2a, Nkx2.5, and Srf. Edges connect the TF to its direct targets. Nodes represent TSSs of genes where binding was observed, the MGI symbols are given as identifiers. Nodes are color-coded according to the GOs assigned to the genes; in cases of multiple assignments the nodes are sectioned. The red square gives an exemplary zoom-in for four genes which are direct targets of both Srf and Nkx2.5. The network will be made publicly available online with an appropriate zoom-in function.

4.3.2.5 Functional Annotation of Direct Targets

For a more comprehensive insight into the function of the TF target genes the direct targets were also assigned to biological function GOs on the fifth level allowing for a more precise characterization of their function. Subsequently, the GO terms of genes identified as direct targets were compared to the GO terms of all genes represented on the array. Terms with *p*-values $\leq 1 \times 10^{-4}$ were considered to be significantly overrepresented. Firstly, the GO terms for all direct targets of one TF independent of the binding of further TFs was investigated. Gene ontology annotations confirmed the important roles each of these TFs play in heart development (Supplementary Table 5 to 8). Not only are the GO terms associated with heart development and function but they are also highly related to the phenotypes observed in the mouse models of the respective TFs. For example, among the Nkx2.5 targets identified in this study, the GO term 'heart looping' is significantly overrepresented ($p = 4.1 \times 10^{-4}$). This result is in good agreement with the observation that in *Nkx2.5* hypomorphs looping of the linear heart tube is not initiated[147]. Further details are discussed in section 5.3.2.

Secondly, the analysis was repeated but only genes bound by two or more TFs were considered. These targets were categorized into eleven groups according which combination of TFs bound (Table 4-12). Overrepresented ($P \leq 1 \times 10^{-4}$) GO terms for genes bound by two, three or all four TFs were determined (Supplementary Table 9). These gene groups are of special interest, as particularly essential genes are known to be frequently redundantly regulated. This is generally attributed to the evolutionary advantage such a regulatory system provides, if one of the TFs or binding sites is either lost or mutated. With few exceptions all significantly overrepresented GOs are related to heart development; in all eleven groups terms such as 'heart development', 'circulation', 'cardiac inotropy' or 'muscle contraction' were found. For those 63 targets where binding of all four TFs was observed the most significantly overrepresented GOs were 'heart development' and 'muscle contraction' ($P < 1 \times 10^{-5}$). Among this group of targets well known cardiac genes were found such as the transcription factors *Foxp1*, *Mitf*, *Nfib*, *Hand2*, *Tbx20*, *Rarb*, *Zeb2* and cofactors *Myocd*, *Ankrd1* as well as genes essential for cytoskeleton formation (e.g. *Actc1*, *Lmod2*, *Nebl*, *Myl1*, *Ttn*). However, several genes bound by all four TFs have previously not been described to play a role in cardiac processes, e.g. the transcription factors *Creb3l2*, *Lmo4*, and *Ppp1r12b*, whether a function in cardiac development can be assigned to these genes remains to be investigated.

Table 4-12. The eleven possible combinations for more than one TF binding to a target gene. The GO terms overrepresented in each group are given in Supplementary Table 9.

Combinations of two	Combinations of three	Combination of four
Gata4 & Mef2a	Gata4 & Mef2a & Nkx2.5	Gata4 & Mef2a & Nkx2.5 & Srf
Gata4 & Nkx2.5	Gata4 & Mef2a & Srf	
Gata4 & Srf	Gata4 & Nkx2.5 & Srf	
Mef2a & Nkx2.5	Mef2a & Nkx2.5 & Srf	
Mef2a & Srf		
Nkx2.5 & Srf		

4.3.3 Transcription Factors and Transcriptional Regulation

4.3.3.1 TF Binding is Associated with Higher Expression Levels

Expression array analysis of HL-1 cells was carried out on genome-wide Illumina arrays and transcripts were classified as expressed or non-expressed based on the array intensities. Subsequently, the expression status of the ChIP-chip identified target genes was analyzed. For each of the four TFs approximately 80% of the target genes are expressed.

The median expression levels of all transcripts identified to be targets in the ChIP-chip analysis were compared to the median expression levels of all other transcripts represented on both the Illumina expression and the TF-ChIP arrays but without binding site (unbound). Clearly, the binding of the TF is associated with elevated expression levels. Wilcoxon rank sum test demonstrated that the distribution of the expression levels of target genes was significantly elevated compared to non-targets ($p \leq 0.005$).

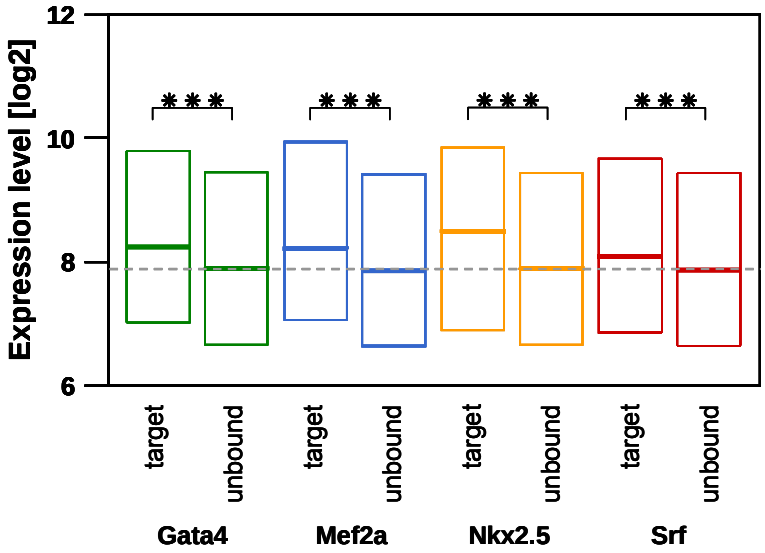
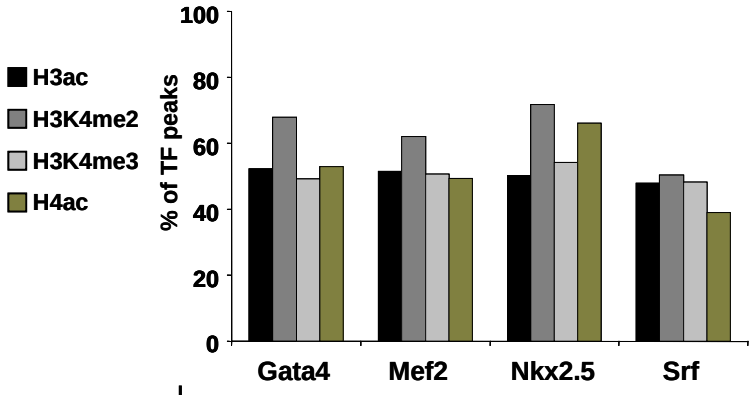


Figure 4-19. Transcripts where binding of Gata4, Mef2a, Nkx2.5, or Srf occurs have elevated expression levels. Transcripts were grouped according to whether binding of a TF was observed on ChIP-chip (target) or not (unbound). For each group the determined expression levels are represented as boxplots. Wilcoxon rank sum test demonstrated that the median expression levels for target genes are significantly elevated. The resultant *p*-values are indicated: $p \leq 0.005$ (***),. $p \leq 0.01$ (**), and $p \leq 0.05$ (*).

4.3.3.2 TFBSs and Histone Modifications Frequently Occur Together

The efficiency of a transcription factor in governing transcription depends on its affinity for its binding site in the promoter of its targets. The accessibility of the promoter is, however, determined by the chromatin configuration. This in its turn is strongly influenced by the modifications on the tails of histones at these positions. The histone modifications H3ac, H4ac, H3K4me2, and H3K4me3 had been previously mapped in HL-1 cells (section 4.2.2). Therefore, it was now investigated to which extent the binding sites of Gata4, Mef2a, Nkx2.5, and Srf occurred at the sites of histone modifications.

As the previous analysis regarding the histone modifications used a smaller array set-up only those TF binding sites were considered that were sufficiently represented. To estimate whether TFBSs are overrepresented in histone modified sites it was first calculated that in a random situation between 23 and 38% of TFBSs would fall together with histone modified sites. The actual number of TFBSs overlapping such sequences was found to be more than twice as high (65-84%), indicating a preferential binding at promoters marked by one or more of the investigated histone modifications (Figure 4-20).



	Gata4	Mef2	Nkx2.5	Srf
Total no. of TF peaks:	287	592	227	734
% Overlapping	79	80	84	65
% Expected at random	30 ± 6	32 ± 3	32 ± 5	30 ± 4

Figure 4-20. Overlap between histone modified sites and TFBSs. TFBSs were only considered if the respective sequences were sufficiently represented on the histone ChIP-array. The expected percentage is based on 100-times random distribution of TFBSs on genomic sequences with marked histone modifications.

The analysis was repeated based on modified domains, however, a preferential binding at one particular histone modification could not be distinguished for any of the TFs as the histone modifications frequently occur together (Table 4-13). The number of TF binding sites occurring at a particular combination was roughly proportional to the absolute number of

domains occurring with this combination (Total domains, Figure 4-21). For example, although the TFs frequently bound at domains coding for H4acH3acK4me2/3 this was also the most frequently observed domain code.

Table 4-13. Overlap between TFBSs and histone modified domains. The last column gives the number of domains with this code found on the TF ChIP-chip array. The last row gives the number of TFBSs for each TF found on the histone ChIP-chip array. The numbers are also visualized in Figure 4-21.

	Gata4	Mef2a	Nkx2.5	Srf	Total number of domains found
H4acH3acK4me2/3	86	147	83	192	508
H4acH3K4me2	30	48	31	24	133
H3acK4me2/3	29	77	17	87	210
H4acH3acK4me2	15	25	5	16	61
H4ac	14	49	17	36	116
H3acK4me2	10	13	1	10	34
H3K4me2/3	10	15	9	14	48
H3K4me2	9	28	10	22	69
H3acK4me3	7	21	2	26	56
H3K4me3	7	23	4	22	56
H3ac	5	18	4	22	49
H4acH3K4me2/3	5	18	9	13	45
H4acH3acK4me3	4	8	3	5	20
H4acH3ac	2	4	1	6	13
H4acH3K4me3	2	3	2	6	13
Sum	235	497	198	501	1431

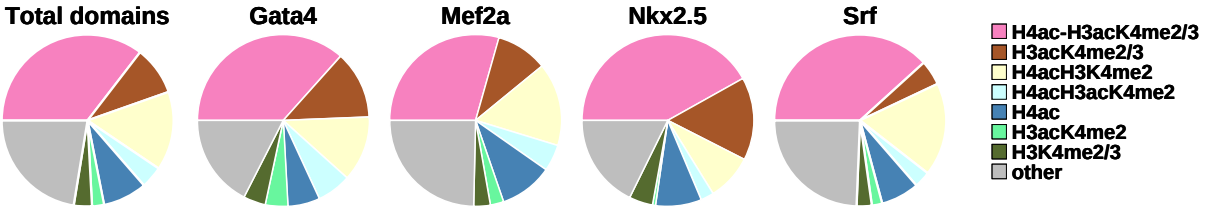


Figure 4-21. Pie chart visualizing the co-occurrence between the transcription factor binding sites and the histone modified domains as listed in Table 4-13. The proportion between the total number of domains found and the number of TFBS co-occurring with the domains is roughly the same for all four TFs.

4.3.3.3 Influence of Histone Modifications on Expression of TF Targets

The high number of co-occurrences between transcription factor binding sites (TFBSs) and modified sites prompted the question whether this influences transcript levels. Each TF the binding sites were categorized into five classes: TF only (only binding of the TF, no histone modification is observed), H3ac, H3K4me2, H3K4me3, and H4ac (histone modification in addition to TF binding). Subsequently, the expression levels of the transcripts assigned to the TFBSs were investigated (Figure 4-22).

The investigated histone modifications are generally thought to be associated with higher transcript levels. Therefore, a Wilcoxon rank sum test was performed to elucidate whether such transcripts associated with TF binding and histone modifications had higher expression levels than such transcripts where only TF binding was observed. A cross

comparison between the modifications was not feasible because of the high co-occurrence of modifications. An analysis based on histone domains (as introduced in section 4.2.2.3) could not be carried out, as the number of transcripts in each class would have been too low.

Interestingly, although each of the TFs (compare Figure 4-19) and each of the histone modifications (compare Figure 4-11A) in themselves were found to be associated with higher transcript levels the effect of the combinations differed for each TF. The levels of transcripts associated with binding of Gata4 and acetylation of histone 3 were significantly elevated. A similar result was obtained for Srf binding sites. In case of Mef2a the median expression levels of all target genes were similar (red line), independent of co-occurrence with histone modifications. The expression levels of transcripts showing binding of Nkx2.5 together with any of the four modifications were elevated compared to binding of Nkx2.5 only.

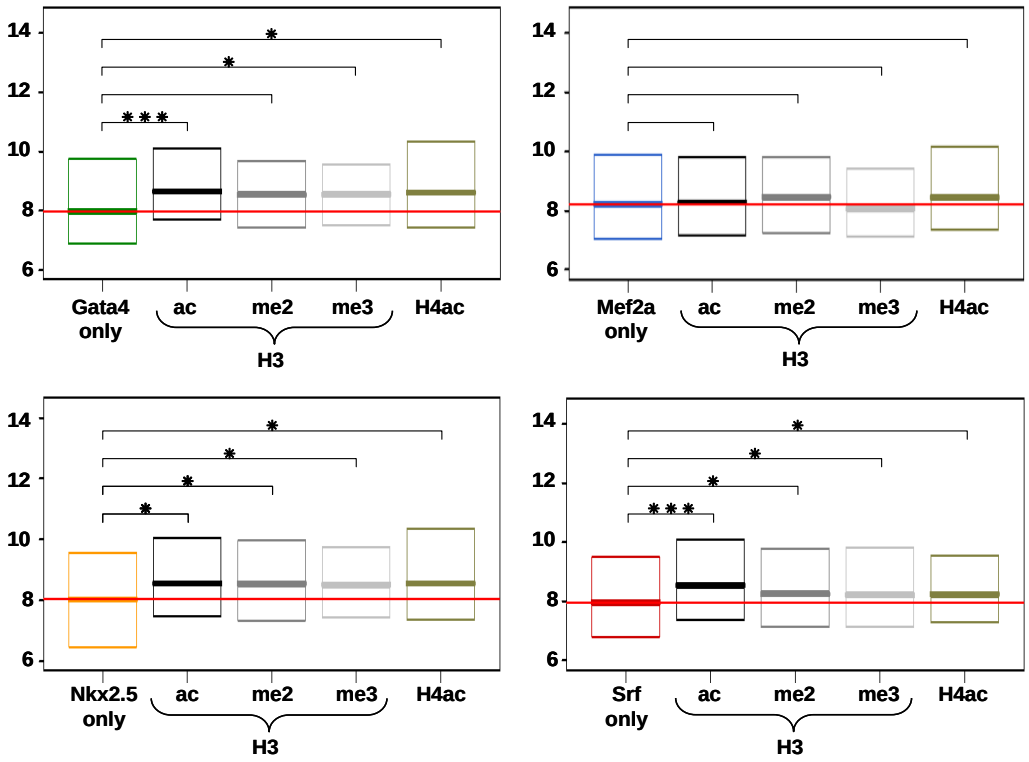


Figure 4-22. The influence of histone modifications at the transcription factor binding sites on expression levels. For each TF the binding sites were categorized into five groups depending on whether co-occurrence with a histone modification could be observed. The expression levels of the transcripts assigned to the binding sites are represented as box plots. Using a Wilcoxon rank sum test it was investigated whether sites where histone modifications were observed are associated with higher expression levels than those without (only).The resultant *p*-values are indicated: $p \leq 0.005$ (***), $p \leq 0.01$ (**) and $p \leq 0.05$ (*).

4.3.4 A Cardiac Regulatory Network

4.3.4.1 Individual Targets can be Activated or Repressed

Although the overall effect of the investigated TFs appears to be activating, this observation cannot be generalized to each individual target gene. To determine functional relevance of the identified binding sites of Gata4, Mef2a, Nkx2.5, and Srf, RNA interference (RNAi) and subsequent expression analysis on Illumina genome-wide expression arrays was carried out. Efficiency of reduction of transcription factor levels was determined both on mRNA level by qPCR and on protein level by immunofluorescence. By monitoring protein levels after transfection the optimal time-point for harvesting cells was determined to be 48 h; after 72 h, recovery of protein levels could be observed (Supplementary Figure 5). For each TF HL-1 cells were treated in duplicate with two different siRNAs causing at least 70% reduction of TF mRNA levels compared to transfections using an unspecific siRNA (siNon, Figure 4-23).

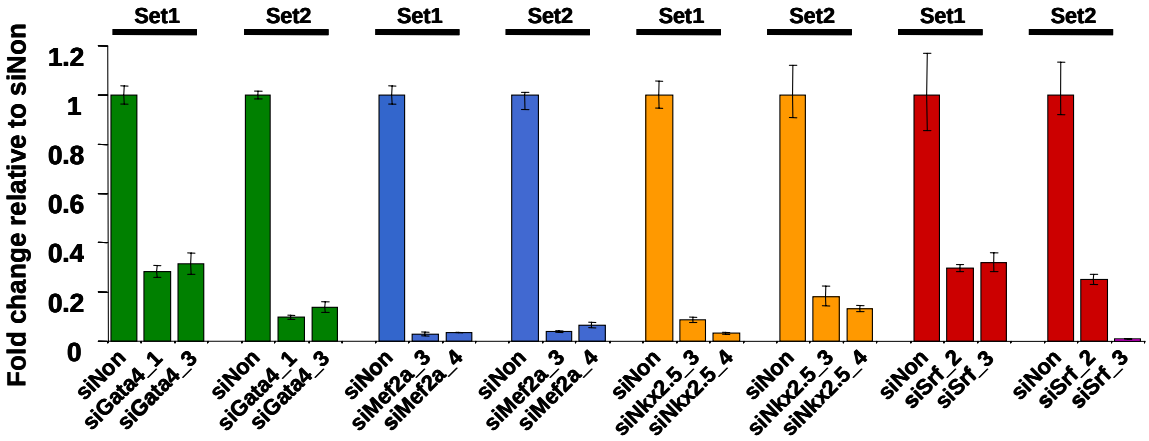


Figure 4-23. Reduction of mRNA levels in HL-1 cells for each TF using RNA interference as determined by qPCR. Experiments were performed in duplicate (set1, set2). mRNA amounts were normalized to Hprt1 and fold changes were calculated relative to an unspecific control siRNA (siNon).

The numbers of transcripts found to be significantly differentially expressed for each siRNA individually and for the two siRNAs targeting the same TF are listed in Table 4-14. For the analysis only such transcripts were considered which were also represented on the TF ChIP arrays (7,895 transcripts). Among the targets which are not expressed in untreated HL-1 cells, none was found to be differentially expressed.

Transcripts appearing to be differentially expressed for only one siRNA are generally considered to be dysregulated due to potential off target effects of the siRNA. However, the regulatory potential of several TFs has been reported to be strongly dosage-dependent (e.g.

Tbx5[310] and Gata4[311]). Therefore, if two siRNAs against the same TF have differing knockdown efficiencies this may also lead to a differing set of dysregulated transcripts in the cells treated with these siRNAs.

Approximately one third of the transcripts measured to be significantly differentially expressed for one siRNA were also identified using the second siRNA for the same TF. Therefore, two groups were defined, one consisting of all genes significantly differentially expressed by either siRNA or both siRNAs (either siRNA). The second group consisted only of those genes significantly differentially expressed by both siRNAs (both siRNAs).

Table 4-14. Number of transcripts significantly differentially expressed ($p \leq 0.05$) in 48 h siRNA treated HL-1 cells as measured on genome-wide Illumina arrays. To compare the number of transcripts identified both in the ChIP and in the siRNA analysis only those 7,895 transcripts were considered represented on both arrays. Compare also Figure 4-24.

	Results for two siRNAs directed against			
	Gata4	***Mef2a***	***Nkx2.5***	***Srf***
upregulated, either siRNA	1,148	136	1,637	527
downregulated, either siRNA	1,593	336	1,723	1,016
upregulated, both siRNAs	175	13	139	51
downregulated, both siRNAs	446	106	643	468

4.3.5 Essential Genes are Redundantly Regulated

The transcripts found to be differentially expressed in siRNA experiments were compared to the targets identified in the ChIP-chip analysis (Figure 4-24). Interestingly, not all binding sites identified in ChIP-chip analysis were assigned to transcripts showing dysregulation in siRNA treated cells, indicating that TF may be bound in a poised state or that additional cofactors maybe lacking.

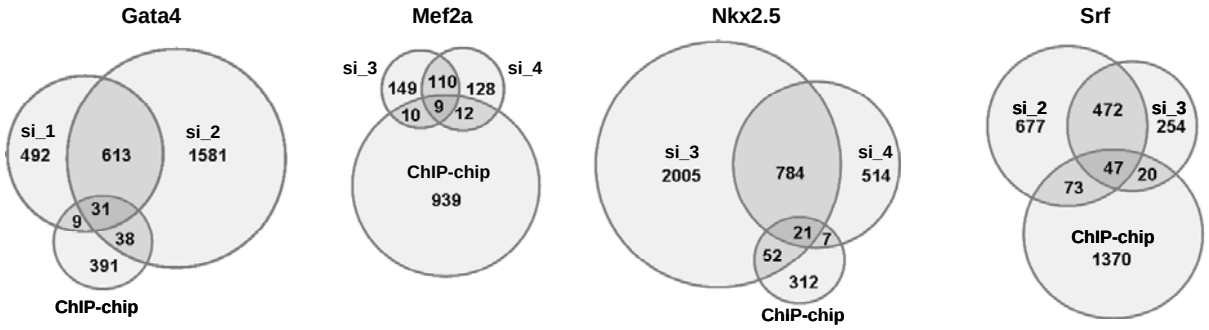


Figure 4-24. Overlap between the significantly differentially expressed transcripts in the two siRNA experiments and the target genes identified in ChIP-chip analysis, respectively. For details on the siRNA results, only, compare Table 4-14.

The reduction of the protein levels of Gata4, Mef2a, Nkx2.5, and Srf should influence direct and indirect target genes. Therefore, it was expected that the number of differentially expressed transcripts (identified in the siRNA experiments) would be higher compared to the number of transcription start sites where direct binding was observed (identified in the ChIP experiments). This was found for Gata4, Nkx2.5, and Srf. In case of Mef2a, however, only a low number of 130 differentially expressed transcripts was identified although the number of directly targeted TSSs exceeded 900.

Using the expression data the effect of transcription factor binding on transcript levels could be classified as activating, repressive, or non-functional (combining the data from Table 4-14 and Figure 4-24). Consequently, the functionality of the TF binding was integrated into the regulatory network. If binding was observed in ChIP but the corresponding transcript was only differentially expressed using one siRNA this was still considered to be sufficiently conclusive of functional binding because two different methods were used. Among these genes several known TF targets were retrieved such as *Actc1*, *Bcl2*, *Ctgf* or *Myocd* (Compare Table 4-9 for references). Interestingly, for transcripts regulated by more than one of the investigated TFs an opposing effect of TF binding was only observed for two cases. *Myocd* (myocardin) is activated by Nkx2.5 and repressed by Srf and *Rbpms* (RNA binding protein gene with multiple splicing) is activated by Nkx2.5 and repressed by Gata4.

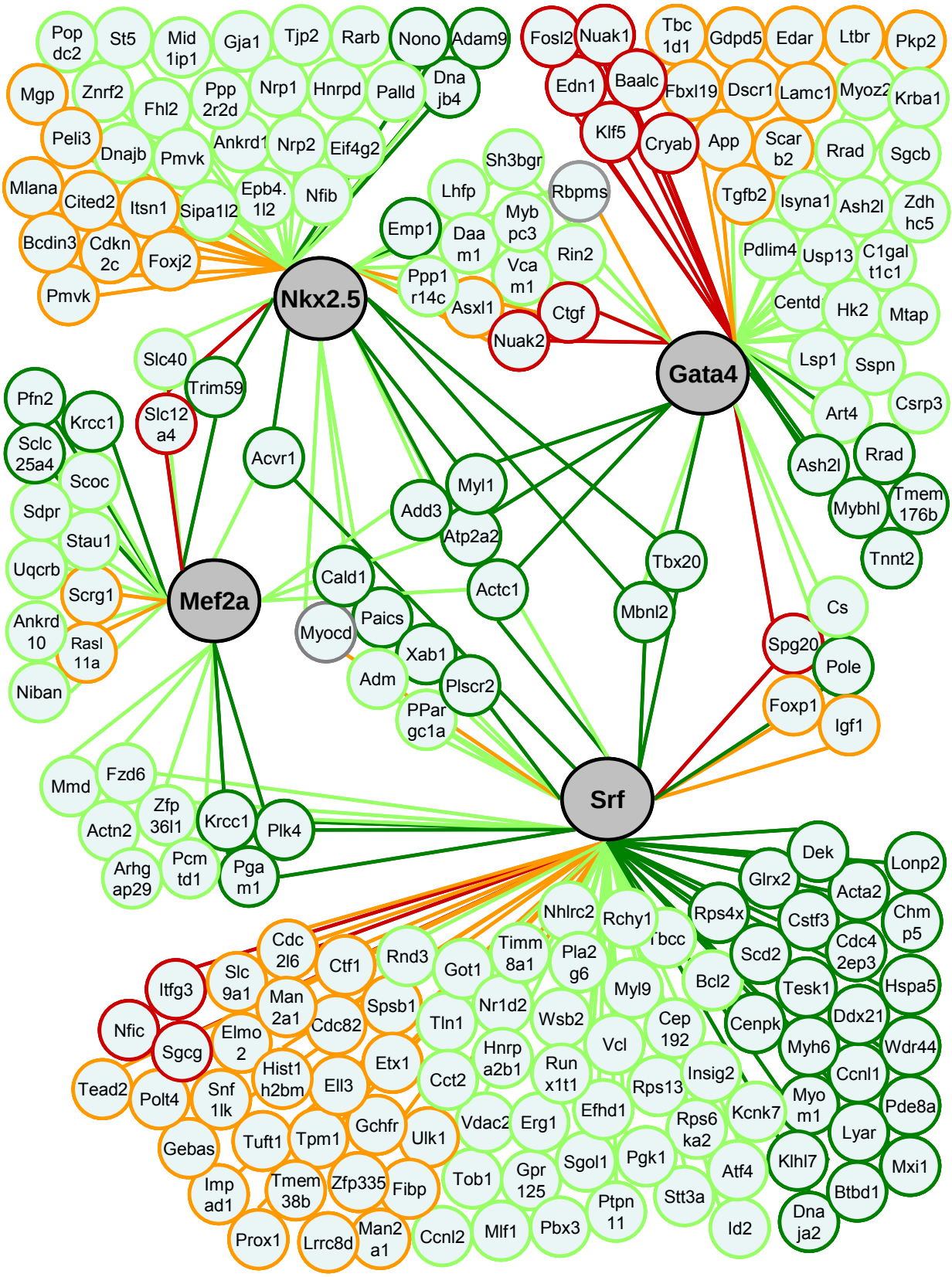


Figure 4-25. TF network showing genes identified to be bound in ChIP-chip analysis and to be differentially expressed after siRNA treatment. Light green edges: expression array data derived from one siRNA indicates activating function of the TF, dark green: both siRNAs. Orange and dark red accordingly for inhibitory TF function. Nodes reflect color of edges. Genes regulated by TFs with opposing functions have gray nodes.

4.3.6 Expression of *Tbx20* is Regulated by Gata4, Mef2a, Nkx2.5, and Srf

Among the directly bound and regulated target genes of Gata4, Mef2a, Nkx2.5, and Srf, several TFs important for muscle development and maintenance were identified, e.g. *Hand2*, *Tbx20* and *Tbx3*. Among these *Tbx20* is of particular interest as mutations in the *Tbx20* gene[312] or changes of *Tbx20* levels are associated with severe congenital malformations both in animal models[313-315] and in patients[312] (unpublished results).

Analyzing the array data for the Gata4, Mef2a, Nkx2.5, and Srf ChIP experiments three sites near the *Tbx20* TSS were found to show enrichment (Figure 4-26A): One in the upstream region (Chr.9:24526600-24527900) and two in the first intron (Chr9: 24521600-24522300 and Chr9: 24523100-24523500). At all three positions enrichment for all four TFs was observed in close proximity and confirmed by qPCR (Supplementary Figure 4). Interestingly, these regions were also associated with the histone modifications H3ac, H4ac, H3K4me2, and H3K4me3 in HL-1 cells as the ChIP-chip array data of these modifications demonstrated (Figure 4-26 B) and as could be confirmed by qPCR (data not shown).

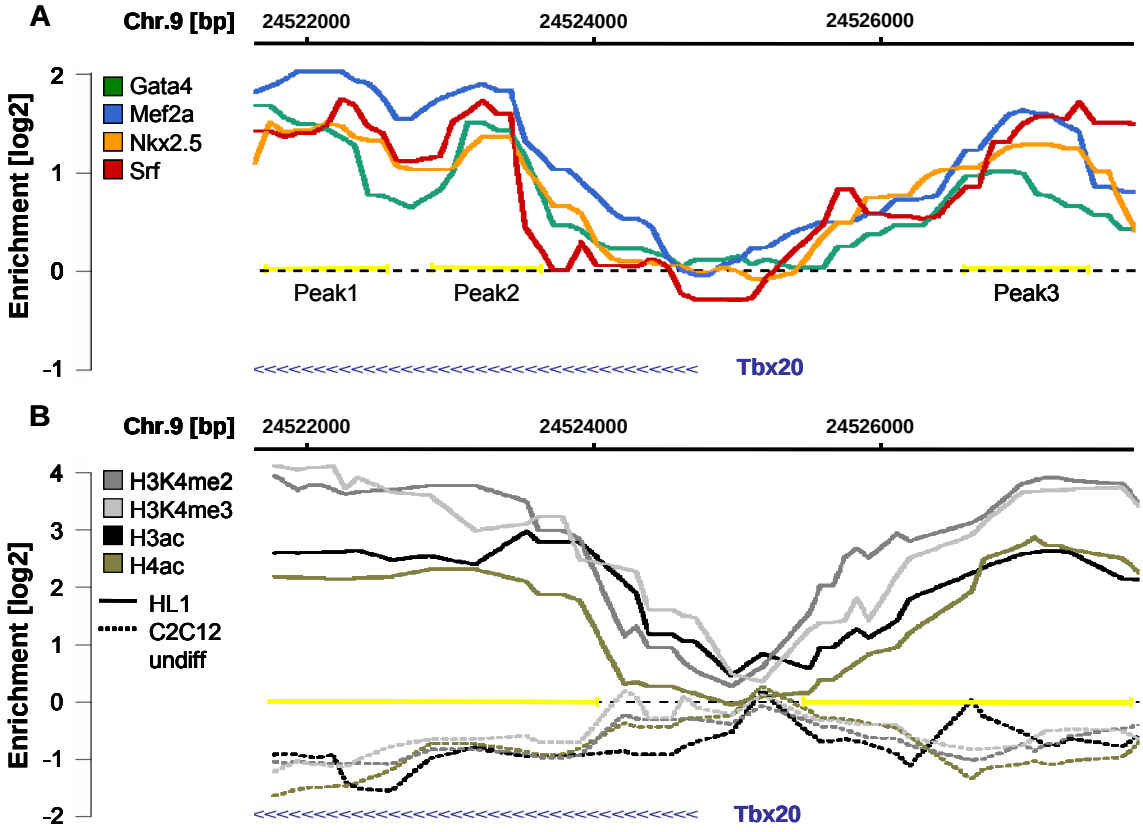


Figure 4-26. ChIP-chip results at the TSS of *Tbx20*. A) For each of the four investigated TFs three binding sites are found two in the first intron of *Tbx20* and one in the upstream region. B) At the sites where TF binding is observed histones carry the modifications H3ac, H4ac, H3K4me2, and H3K4me3 in HL-1 cells. In C2C12 cells (dashed lines) none of the histone modifications is enriched.

These modifications are thought to induce an open chromatin structure making the DNA accessible for TF binding. In C2C12 undifferentiated cells, where *Tbx20* is not expressed these modifications could not be detected. Analysis of the underlying sequences revealed that each site was located within a conserved element (according to PhastCons) and contained the binding motifs of all four TFs.

Measuring *Tbx20* levels in the siRNA knockdown experiments of the respective TFs (Figure 4-27 A) showed reduction of *Tbx20* mRNA levels by 20-50% for both siRNAs targeting *Gata4*, *Mef2a*, and *Srf*, respectively, and for one siRNA targeting *Nkx2.5* (Figure 4-27 B). These results demonstrate, that binding of these TFs is indeed functional and activates *Tbx20* expression.

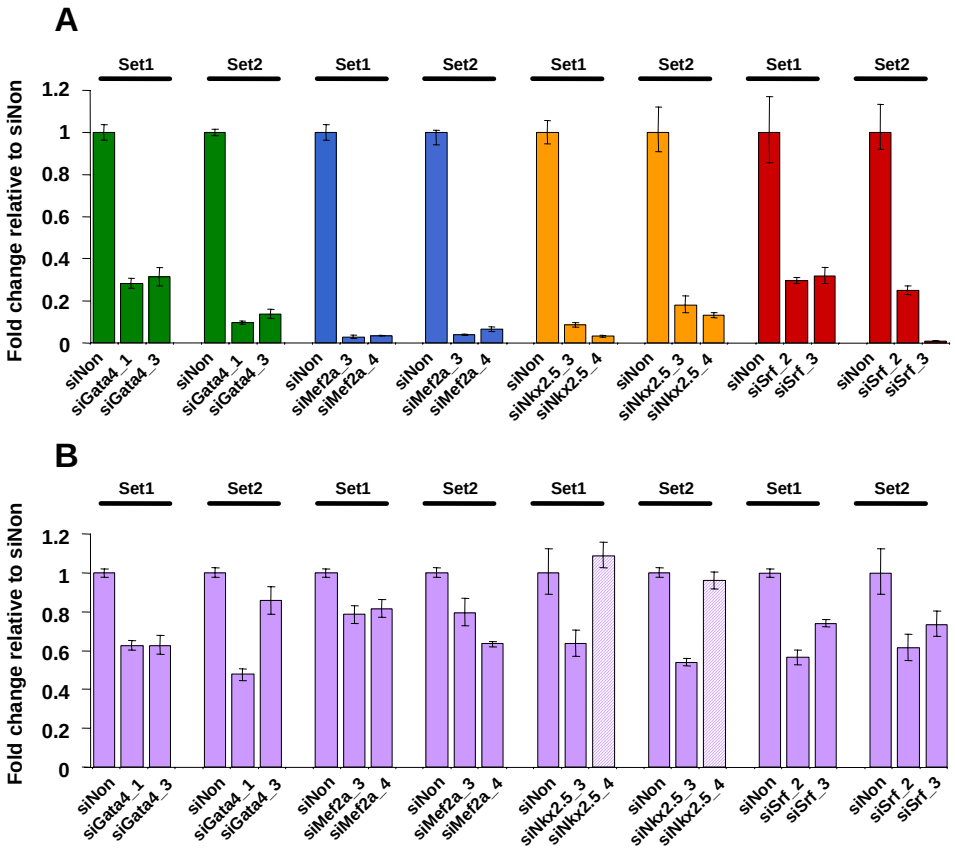


Figure 4-27. *Tbx20* levels are decreased in cells treated with siRNA against one of the four TFs. A) Knockdown of the TF mRNA levels as measured by qPCR. B) Except in cells treated with siNkx2.5_3 the *Tbx20* levels are decreased by ≈ 20 – 50%.

4.3.7 Interim Summary: Regulatory Network of Gata4, Mef2a, Nkx2.5, and Srf

The transcription factors Gata4, Mef2a, Nkx2.5, and Srf are known to be essential for the correct formation of cardiac structures. Investigation by ChIP-chip in cardiomyocytes resulted in the confirmation of known target genes as well as the identification of several hundred novel downstream targets per TF, many of which are redundantly regulated. Gene ontology terms of the target genes confirmed previously reported phenotypes of mouse models. The analysis of the binding site sequences revealed agreement with known binding motifs and showed low conservation of binding sites. Integration with previous results of section 4.2 revealed that the majority of TFBSs co-occur with histone modifications. RNA interference coupled with expression array analysis revealed which binding sites have activating, repressive or no effect on transcription. This resulted in the construction of a regulatory subnetwork where the T-box transcription factor *Tbx20* was identified as a commonly activated target.